BMC Medical Genetics

BioMed Central

Research article

The impact of population heterogeneity on risk estimation in genetic counseling

Wenlei Liu[1], Nikolina Icitovic[2], Michele L Shaffer[1] and Gary A Chase*[1]

Address: [1]Department of Health Evaluation Sciences, Penn State College of Medicine, A210, Suite 2200, 600 Centerview Drive, Hershey, PA 17033, USA and [2]Department of Statistics, Penn State University, 326 Thomas Building, University Park, PA 16802, USA

Email: Wenlei Liu - wliu@hes.hmc.psu.edu; Nikolina Icitovic - nzi100@psu.edu; Michele L Shaffer - mshaffer@hes.hmc.psu.edu; Gary A Chase* - gchase@hes.hmc.psu.edu

* Corresponding author

Published: 30 June 2004

BMC Medical Genetics 2004, **5**:18 doi:10.1186/1471-2350-5-18

Received: 21 October 2003
Accepted: 30 June 2004

This article is available from: http://www.biomedcentral.com/1471-2350/5/18

Abstract

Background: Genetic counseling has been an important tool for evaluating and communicating disease susceptibility for decades, and it has been applied to predict risks for a wide class of hereditary disorders. Most diseases are complex in nature and are affected by multiple genes and environmental conditions; it is highly likely that DNA tests alone do not define all the genetic factors responsible for a disease, so that persons classified into the same risk group by DNA testing actually could have different disease susceptibilities. Ignorance of population heterogeneity may lead to biased risk estimates, whereas additional information on population heterogeneity may improve the precision of such estimates.

Methods: Although DNA tests are widely used, few studies have investigated the accuracy of the predicted risks. We examined the impact of population heterogeneity on predicted disease risks by simulation of three different heterogeneity scenarios and studied the precision and accuracy of the risks estimated from a logistic regression model that ignored population heterogeneity. Moreover, we also incorporated information about population heterogeneity into our original model and investigated the resulting improvement in the accuracy of risk estimation.

Results: We found that heterogeneity in one or more categories could lead to biased estimates not only in the "contaminated" categories but also in other homogeneous categories. Incorporating information about population heterogeneity into the original model greatly improved the accuracy of risk estimation.

Conclusions: Our findings imply that without thorough knowledge about genetic basis of the disease, risks estimated from DNA tests may be misleading. Caution should be taken when evaluating the predicted risks obtained from genetic counseling. On the other hand, the improved accuracy of risk estimates after incorporating population heterogeneity information into the model did point out a promising direction for genetic counseling, since more and more new techniques are being invented and disease etiology is being better understood.

Background

With the in-depth study of modern genetics, its principles have been discovered and applied widely in clinical settings. Many diseases have been found to "run in families"

exhibiting simple Mendelian inheritance patterns, such as muscular dystrophy and Huntington's disease. Advances in knowledge about the genetic basis of disease enable the expansion of DNA testing both for diagnosis and for prediction of disease susceptibility beyond simply inherited traits. Demand for and expectation of genetic counseling keeps increasing over time. Many methodologies have been developed to estimate the age of onset [1-3] and lifetime risk or recurrence of hereditary disorders [4-7]. These have been successfully applied to determine a person's risk of developing a genetic disease or to determine the risk of having a child with a genetic disease [8,9]. However, there are cases where predictive tests based on family history cannot give satisfactory assessment. For example, BRCA-1 and BRCA-2 are believed to be breast cancer susceptibility genes; Begg [10] has pointed out that lifetime risk estimates for breast cancer derived from samples of multiple-case families are not always applicable to new BRCA-1 or BRCA-2 positive women who request genetic counseling. Because patients undergoing predictive DNA testing usually have no symptoms or clinical presentation, it is particularly important for this type of DNA testing to give precise estimates. Getting wrong answers either way has long-term effects on the individual or family and could lead to irreversible life decisions, e.g. prophylactic mastectomy. Therefore, finding out the cause of biased estimates and its impact on the predicted risks should be an important goal of contemporary genetic counseling.

In this paper, we investigated mechanisms for generating biased risk estimates in heterogeneous populations. We assumed that some individuals in certain groups were "contaminated" by having a very low probability, as a result of unmeasured factors, of getting the disease. If individual contamination status is properly taken into account, the estimated risk should be close to its true value. However, if contamination status is overlooked, the estimated risk will be biased. A dichotomous disease was modeled before and after this latent contamination factor was incorporated into a logistic regression model. The accuracy of the estimates was explored by computing relative bias and root mean square error (RMSE) of the estimated risks using simulated data sets.

Methods

Simulations

To find out the effect of population heterogeneity on the estimated disease risks, we carried out a set of simulations. We assumed that the individual disease risks were estimated based on genotype at a biallelic locus (L1) and their exposure to a fixed, dichotomous environmental factor (E1). However, the phenotype resulting from L1 could be overridden by the genotype at another unscreened locus (L2). In addition, interactions between E1 and the joint genotype at L1 and L2 were assumed to be present. For the two alleles, A and B at L1, there were three genotypes, AA, AB and BB. Individual genotypes at L1 were simulated assuming Hardy-Weinberg Equilibrium (HWE) with both alleles having equal frequencies. Individual exposure to E1 was randomly determined using a population exposure rate of 0.20. Disease status was determined by the following logistic regression model:

$$P(affected \mid x_1, x_2, x_3) = \frac{\exp(\alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3)}{1 + \exp(\alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3)}$$

where design variable x_1 indicated whether the individual genotype was AA, design variable x_2 indicated whether the genotype was BB, and independent variable x_3 denoted whether the individual was exposed to the environmental risk factor E1. Allele A was partially dominant to allele B, and AA individuals were most likely to have the disease. In addition, individuals exposed to E1 were more likely to be affected than individuals with the same genotype not exposed to E1. The coefficients of the model were set as follows: α = -2.197, β_1 = 1.35, β_2 = -0.747 and β_3 = 0.811 so that there was no genotype-environment interaction involved in the logistic regression model. The coefficients β_1 and β_2 measure the impact of genotypes AA and BB relative to AB.

To study the impact of population heterogeneity on predictive risks, we looked at three different contamination scenarios. In the first scenario, contamination occurred only in individuals with AB genotype and not exposed to E1. This could happen when there was interaction between some external environment and a joint genotype. For example, individuals with AB genotype at L1 and CC genotype at L2 could get the disease with a very low probability if not exposed to E1. In the second scenario, we assumed that contamination presented in two different categories: AB individuals exposed and not exposed to E1. This could occur if the disease phenotype resulted from AB genotype at L1 was masked by a genotype of the second locus, CC so that all the individuals with genotype AC/BC have very low disease susceptibilities in the absence of genotype-environment interaction. In the presence of genotype-environment interaction, some AC/BC individuals might express normally under one environmental condition but not the other, so that the proportions of contaminated individuals were different in the two categories. In the last scenario, contamination happened in AB individuals not exposed to E1 and in AA individuals exposed to E1, which was possible again when there was interaction between the environment and some (but not all) joint genotypes.

With contamination properly taken into account, the accuracy of predicted risks should be improved. To investigate potential improvement of the predicted risks, we

also estimated the risks by incorporating the contamination factor into the logistic regression model (full model). For contamination in a single category, an independent variable x_4 was used in the full model to denote whether the individual was contaminated or not. Disease status was determined by the full model as follows:

$$P(affected \mid x_1, x_2, x_3, x_4) = \frac{\exp(\alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4)}{1 + \exp(\alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4)}$$

Likewise, for contamination in two categories, an additional independent variable x_5 was used in the full model to denote whether the individual was contaminated or not in the second category.

In each contamination scenario, the proportion of contaminated individuals (contamination factor) varied from 0 to 0.8 at an interval of 0.2. The disease risks of contaminated individuals were set to 0.01. Each data set consisted of 600 unrelated individuals with simulated genotypes, environmental exposure status, contamination status and affection status. 1000 replicated data sets were simulated for each parameter set.

Statistical analysis

Two logistic regression models were used to fit the simulated data sets. The reduced model has two covariates denoting the genotype and one covariate indicating environmental exposure status. The full model has two genotype covariates, one environmental covariate and one or two additional covariates indicating individual's contamination status. Maximum likelihood estimates of the coefficients were computed using computer software SAS version 8.0. To find out how different the estimated disease risks and the true risks were, we calculated the RMSE and relative bias of the predicted risks, averaged over the 1000 replicated data sets. Relative bias was defined as the bias of the estimated risk divided by the true risk to facilitate interpretation of the bias on an appropriate scale.

Results

Tables 1 and 2 list the RMSE and relative bias of the disease risks estimated using the reduced and full model averaged over the 1000 replicated data sets when contamination occurred in AB individuals not exposed to the environmental factor. When there was no contamination, we could see that the relative biases were small (less than 1.8 percent) in all six categories in both models. This implied that our reduced and full models were both efficient and could give precise estimates under no contamination. When contamination occurred, both the relative biases and RMSE of the risks estimated from the reduced model increased in all categories with larger proportion of contaminated individuals. The predicted risks increased in BB individuals exposed to the environment and AA individuals exposed to the environment. They decreased in the other four categories. The logit of the contaminated category corresponded to the intercept of the logistic model. Since the estimated coefficients of logistic regression model were interdependent, changing of one parameter led to changing of all the other parameters. Thus all the predicted risks deviated from their true values as a result of contamination in a single category, though the relative bias increased fastest in the contaminated category. In the contaminated category, the predicted risk differed greatly from its true value (15 percent difference) even with 20 percent contaminated individuals. The deviation reached nearly 60 percent when the proportion of contaminated individuals reached 0.8. When contamination status was incorporated into the model (full model), the relative biases and RMSE remained small in all six categories despite increasing proportion of contaminated individuals. The relative bias was less than 3 percent in all the categories even with 80 percent contaminated individuals.

Tables 3 and 5 present our findings for the impact of two-category contamination on predictive risks. In general, the relative biases and RMSE increased with increasing proportions of contamination. However, there were cases where they decreased with increasing proportions of contamination. For example, in scenario three the absolute value of the relative bias of the estimated risk of AB individuals not exposed to the environmental factor decreased with increasing contamination in AA individuals exposed to the environment, as shown in Table 5. Actually, we could see that the relative bias varied from -0.109 to 0.008 when $r_{AB,NE}$ equaled 0.2. This indicated that with increasing contamination, the predicted risk reduced at first, but then it increased and became larger than its true value. This was possible because the predicted risks were functions of the model coefficients. Contamination caused different coefficients to change in different directions. Therefore, the predicted risks could fluctuate in either direction with increasing contamination. In the third scenario, when contamination proportions in both categories were 0.2, the estimated risk for AB individuals exposed to the environment reduced 13 percent even though there was no contamination in this category. When the two contamination factors were 0.6 and 0.8, the estimated risk for this category decreased nearly 50 percent.

Tables 4 and 6 list the average relative bias and RMSE of the disease risks for the second and third contamination scenarios estimated using the full model. Similar to one group contamination case, when additional knowledge about individual's contamination status was available, the estimated relative biases and RMSE were greatly improved in all categories. They remained small despite increasing

Table 1: Relative bias and RMSE of the disease risks predicted using the reduced model in the first scenario.

Contam Factor [a]	AB NE[b]		AB E[c]		BB NE	
	Relative bias	RMSE	Relative bias	RMSE	Relative bias	RMSE
0	-0.0030	0.0179	0.0098	0.039	-0.0007	0.0168
0.2	-0.1481	0.022	-0.0510	0.0424	-0.0342	0.016
0.4	-0.2928	0.0326	-0.1346	0.0471	-0.0792	0.0162
0.6	-0.4346	0.0452	-0.2032	0.0567	-0.0959	0.0165
0.8	-0.5725	0.0582	-0.2980	0.0710	-0.1374	0.0173
Contam Factor	BB E		AA NE		AA E	
	Relative bias	RMSE	Relative bias		Relative bias	RMSE
0	0.0157	0.0381	0.0065	0.0383	0.0087	0.0647
0.2	0.068	0.0393	-0.0173	0.039	0.0379	0.0704
0.4	0.1038	0.0421	-0.0299	0.0405	0.0730	0.0770
0.6	0.2151	0.0479	-0.0456	0.0412	0.1302	0.094
0.8	0.3071	0.0559	-0.0695	0.0452	0.1819	0.1132

(a) Proportion of individuals in the contaminated category who get the disease with a very low probability (0.01). (b) Not exposed to the environmental factor. (c) Exposed to the environmental factor.

Table 2: Relative bias and RMSE of the disease risks predicted using the full model in the first scenario.

Contam Factor [a]	AB NE[b]		AB E[c]		BB NE	
	Relative bias	RMSE	Relative bias	RMSE	Relative bias	RMSE
0	0.0002	0.01786	0.0088	0.03874	0.0053	0.0168
0.2	0.0053	0.01940	0.0094	0.04196	0.0028	0.01626
0.4	0.0078	0.02154	-0.006	0.04085	-0.0092	0.01643
0.6	0.0113	0.02564	0.0023	0.04316	0.0177	0.01692
0.8	0.0230	0.03089	0.0009	0.04522	0.0284	0.01755
Contam Factor	BB E		AA NE		AA E	
	Relative bias	RMSE	Relative bias	RMSE	Relative bias	RMSE
0	0.0170	0.03801	0.0040	0.03805	0.0045	0.06422
0.2	0.0121	0.03699	-0.0042	0.03842	-0.0041	0.06724
0.4	-0.0131	0.03755	0.0001	0.03925	-0.0099	0.06803
0.6	0.0211	0.03736	0.0037	0.03895	0.0000	0.06866
0.8	0.0278	0.03892	0.0036	0.04021	-0.0021	0.06951

(a) Proportion of individuals in the contaminated category who get the disease with a very low probability (0.01). (b) Not exposed to the environmental factor. (c) Exposed to the environmental factor.

proportion of contaminated individuals. In both scenarios, the largest relative risks were about 3 percent even with 80 percent contaminated individuals. These results suggested that the additional contamination covariate was efficient in modeling population heterogeneity. The difference in individual's disease susceptibility was accounted for properly. Knowledge about contamination could improve the accuracy of the predicted risks.

Discussion

Rapid developments in genetics have an increasing impact on medical practice. Genetic counseling has made it possible to predict an individual's risk for complex genetic diseases that do not cleanly follow Mendelian inheritance patterns. However, predictive tests based on family history cannot always give satisfactory assessment due to the complexity of human diseases; "one size fits all" techniques appear to be problematic. Incomplete information

Table 3: Relative bias and RMSE of the disease risks predicted using the reduced model in the second scenario.

		AB NE[c]		AB E[d]		BB NE	
$r_{AB,NE}$[a]	$r_{AB,E}$[b]	Relative risk	RMSE	Relative risk	RMSE	Relative risk	RMSE
0.2	0.2	-0.1811	0.0244	-0.1663	0.0494	0.0066	0.0169
	0.4	-0.2268	0.0274	-0.2882	0.0666	0.0251	0.0166
	0.6	-0.2719	0.0314	-0.4016	0.0858	0.0810	0.0186
	0.8	-0.3063	0.0343	-0.5001	0.1033	0.1284	0.0190
0.6	0.2	-0.4712	0.0487	-0.3140	0.0717	-0.0684	0.0165
	0.4	-0.5147	0.0528	-0.4333	0.0922	-0.0450	0.0161
	0.6	-0.5555	0.0568	-0.5414	0.1116	0.0141	0.0174
	0.8	-0.5962	0.0607	-0.6362	0.1293	0.0652	0.0176
		BB E		AA NE		AA E	
$r_{AB,NE}$	$r_{AB,E}$	Relative risk	RMSE	Relative risk	RMSE	Relative risk	RMSE
0.2	0.2	0.0045	0.0373	0.0060	0.0397	-0.0022	0.0691
	0.4	-0.0810	0.0364	0.0218	0.0408	-0.0551	0.0737
	0.6	-0.1446	0.0373	0.0453	0.0417	-0.1048	0.0865
	0.8	-0.2155	0.0392	0.0740	0.0454	-0.1584	0.1034
0.6	0.2	0.1459	0.0441	-0.0308	0.0410	0.0891	0.0836
	0.4	0.0526	0.0399	-0.0148	0.0406	0.0348	0.0735
	0.6	-0.0175	0.0383	0.0099	0.0399	-0.0175	0.0732
	0.8	-0.0961	0.0373	0.0393	0.0417	-0.0730	0.0797

(a) Proportion of AB individuals not exposed to the environment who get the disease with a very low probability (0.01). (b) Proportion of AB individuals exposed to the environment who get the disease with a very low probability (0.01). (c) Not exposed to the environmental factor. (d) Exposed to the environmental factor.

Table 4: Relative bias and RMSE of the estimated disease risks predicted using the full model in the second scenario.

		AB NE[c]		AB E[d]		BB NE	
$r_{AB,NE}$[a]	$r_{AB,E}$[b]	Relative risk	RMSE	Relative risk	RMSE	Relative risk	RMSE
0.2	0.2	0.0127	0.0198	0.0135	0.0434	0.0099	0.0169
	0.4	0.0019	0.0191	-0.0015	0.0472	-0.0054	0.0162
	0.6	-0.0060	0.0198	-0.0094	0.0517	0.0135	0.0174
	0.8	0.0062	0.0204	0.0054	0.0613	0.0206	0.0169
0.6	0.2	0.0177	0.0257	0.0096	0.0454	0.0118	0.0169
	0.4	0.0109	0.0256	-0.0049	0.0503	-0.0014	0.0164
	0.6	0.0036	0.0266	-0.0194	0.0545	0.0187	0.0174
	0.8	0.0155	0.0274	-0.0095	0.0656	0.0281	0.0169
		BB E		AA NE		AA E	
$r_{AB,NE}$	$r_{AB,E}$	Relative risk	RMSE	Relative risk	RMSE	Relative risk	RMSE
0.2	0.2	0.0136	0.0376	0.0017	0.0395	-0.0021	0.0690
	0.4	-0.0063	0.0376	-0.0010	0.0399	-0.0104	0.0699
	0.6	0.0022	0.0383	0.0022	0.0394	-0.0104	0.0720
	0.8	0.0115	0.0395	0.0081	0.0402	-0.0087	0.0755
0.6	0.2	0.0182	0.0383	0.0024	0.0400	-0.0014	0.0706
	0.4	-0.0068	0.0377	0.0003	0.0399	-0.0118	0.0714
	0.6	-0.0025	0.0385	0.0044	0.0396	-0.0150	0.0731
	0.8	-0.0011	0.0390	0.0114	0.0402	-0.0173	0.0744

(a) Proportion of AB individuals not exposed to the environment who get the disease with a very low probability (0.01). (b) Proportion of AB individuals exposed to the environment who get the disease with a very low probability (0.01). (c) Not exposed to the environmental factor. (d) Exposed to the environmental factor.

Table 5: Relative bias and RMSE of the disease risks predicted using the reduced model in the third scenario.

		AB NE[c]		AB E[d]		BB NE	
$r_{AB,NE}$[a]	$r_{AA,E}$[b]	Relative risk	RMSE	Relative risk	RMSE	Relative risk	RMSE
0.2	0.2	-0.1090	0.0203	-0.1286	0.0465	0.0177	0.0177
	0.4	-0.0663	0.0175	-0.2135	0.0574	0.0550	0.0171
	0.6	-0.0312	0.0176	-0.3090	0.0718	0.1079	0.0186
	0.8	0.0080	0.0171	-0.3856	0.0846	0.1884	0.0217
0.6	0.2	-0.4019	0.0422	-0.2663	0.0648	-0.0520	0.0172
	0.4	-0.3671	0.0387	-0.3400	0.0773	-0.0071	0.0163
	0.6	-0.3330	0.0359	-0.4214	0.0911	0.0539	0.0175
	0.8	-0.2949	0.0326	-0.4846	0.1023	0.1396	0.0202
		BB E		AA NE		AA E	
$r_{AB,NE}$	$r_{AA,E}$	Relative risk	RMSE	Relative risk	RMSE	Relative risk	RMSE
0.2	0.2	-0.0211	0.0383	-0.0432	0.0414	-0.0585	0.0729
	0.4	-0.1316	0.0368	-0.0894	0.0466	-0.1820	0.1120
	0.6	-0.2365	0.0400	-0.1189	0.0541	-0.2964	0.1585
	0.8	-0.3079	0.0439	-0.1498	0.0605	-0.3981	0.2054
0.6	0.2	0.1094	0.0436	-0.0775	0.0460	0.0268	0.0705
	0.4	-0.0138	0.0377	-0.1213	0.0529	-0.1033	0.0874
	0.6	-0.1335	0.0373	-0.1473	0.0602	-0.2263	0.1296
	0.8	-0.2158	0.0400	-0.1746	0.0663	-0.3363	0.1786

(a) Proportion of AB individuals not exposed to the environment who get the disease with a very low probability (0.01). (b) Proportion of AA individuals exposed to the environment who get the disease with a very low probability (0.01). (c) Not exposed to the environmental factor. (d) Exposed to the environmental factor.

Table 6: Relative bias and RMSE of the disease risks predicted using the full model in the third scenario.

		AB NE[c]		AB E[d]		BB NE	
$r_{AB,NE}$[a]	$r_{AA,E}$[b]	Relative risk	RMSE	Relative risk	RMSE	Relative risk	RMSE
0.2	0.2	0.0030	0.0194	0.0005	0.0419	0.0086	0.0174
	0.4	0.0096	0.0221	-0.0049	0.0431	0.0072	0.0174
	0.6	0.0117	0.0261	-0.0133	0.0451	0.0250	0.0177
	0.8	0.0197	0.0313	-0.0123	0.0460	0.0238	0.0165
0.6	0.2	-0.0052	0.0193	-0.0120	0.0444	0.0127	0.0166
	0.4	0.0141	0.0231	-0.0087	0.0464	0.0171	0.0171
	0.6	0.0179	0.0262	-0.0080	0.0460	0.0052	0.0172
	0.8	0.0285	0.0329	-0.0075	0.0457	0.0319	0.0172
		BB E		AA NE		AA E	
$r_{AB,NE}$	$r_{AB,E}$	Relative risk	RMSE	Relative risk	RMSE	Relative risk	RMSE
0.2	0.2	0.0051	0.0377	-0.0013	0.0388	-0.0067	0.0698
	0.4	-0.0076	0.0364	-0.0013	0.0401	-0.0121	0.0728
	0.6	0.0060	0.0369	-0.0011	0.0397	-0.0144	0.0757
	0.8	0.0171	0.0382	0.0012	0.0409	-0.0096	0.0789
0.6	0.2	0.0093	0.0385	-0.0041	0.0401	-0.0133	0.0831
	0.4	-0.0009	0.0377	-0.0006	0.0402	-0.0166	0.0860
	0.6	-0.0081	0.0390	0.0028	0.0403	-0.0113	0.0896
	0.8	0.0262	0.0399	-0.0007	0.0393	-0.0098	0.0916

(a) Proportion of AB individuals not exposed to the environment who get the disease with a very low probability (0.01). (b) Proportion of AB individuals exposed to the environment who get the disease with a very low probability (0.01). (c) Not exposed to the environmental factor. (d) Exposed to the environmental factor.

about disease etiology might lead to individuals classified in the same group by DNA tests to have different susceptibilities to a disease.

The problem of model misspecification is not new. Several studies investigated the asymptotic relative efficiency (ARE) of misspecified model in testing association between exposure and response [11-13]. ARE can be defined loosely as the ratio of sample size needed by the correct test to attain the same power as the mismodeled test. Among these studies, Begg and Lagakos explored the consequences of model misspecification in logistic regression and showed both theoretically and numerically that models with missing or incorrect covariates required larger sample sizes to achieve the same power of testing association between the exposure and response than the correct models [13]. Although efforts have been made to study the effect of model misspecification, little attention has been paid to investigate the problem in the context of genetic counseling, where predicting disease risk is the primary goal. In this paper, we studied the impact of population heterogeneity on predicted disease risks. A logistic regression model was fitted assuming the individual contamination status was unknown (contamination status is a missing covariate). We quantified the bias of the predicted risks based on the level of population contamination through a simulation study. Our results showed that contamination in one or more categories could cause the estimated risks in all categories to deviate from their true values. The departure could be in either direction and the biases were unpredictable. We focused our simulations in three specified situations, though the results could be easily generalized to other scenarios. This implies that without thorough knowledge about genetic basis of the disease, risks estimated from DNA tests may be misleading. Since human bodies are so complicated and disease systems are so sophisticated, it is hard to detect contamination status for many genetic disorders. Therefore, caution should be taken when evaluating the predicted risks obtained from genetic counseling.

Our simulation using the full model did show that major improvements could be made if individual disease status was available and incorporated into the prediction model. This pointed out a promising direction for genetic counseling, since more and more new techniques are being invented and genetic disorders are being better understood.

Conclusions

Our simulation results showed that heterogeneity in one or more categories could lead to biased estimates of disease risk not only in the "contaminated" categories but also in other homogeneous categories. The predicted risks could fluctuate in either direction and the biases were unpredictable. These findings imply that without thorough knowledge about genetic basis of the disease, risks estimated from DNA tests may be misleading. Caution should be taken when evaluating the predicted risks obtained from genetic counseling.

Competing interests

None declared.

Authors' contributions

WL performed the statistical analysis and drafted the manuscript. NI and MLS participated in designing and performing the statistical analysis. GAC conceived of the study, and participated in its design and coordination. All authors read and approved the final manuscript.

References

1. Brackenridge CJ, Teltscher B: **Estimation of the age at onset of Huntington's disease from factors associated with the affected parent.** *J Med Genet* 1975, **12:**64-69.
2. Heimbuch RC, Matthysse S, Kidd KK: **Estimating age-of-onset distributions for disorders with variable onset.** *Am J Hum Genet* 1980, **32:**564-574.
3. Ripatti S, Gatz M, Pedersen NL, Palmgren J: **Three-state frailty model for age at onset of dementia and death in Swedish twins.** *Genet Epidemiol* 2003, **24:**139-149.
4. Paulson GW: **Predictive tests in Huntington's disease.** *Res Publ Assoc Res Nerv Ment Dis* 1976, **55:**317-329.
5. Chakravarti A, Buetow KH: **A strategy for using multiple linked markers for genetic counseling.** *Am J Hum Genet* 1985, **37:**984-997.
6. Chase GA, Markson LE, Brookmeyer R, Folstein SE: **Covariate-dependent genetic counseling in Huntington's disease.** *J Neurogenet* 1986, **3:**215-223.
7. Gold LS, Gaylor DW, Slone TH: **Comparison of cancer risk estimates based on a variety of risk assessment methodologies.** *Regul Toxicol Pharmacol* 2003, **37:**45-53.
8. Roth MP, Petersen GM, McElree C, Vadheim CM, Panish JF, Rotter JI: **Familial empiric risk estimates of inflammatory bowel disease in Ashkenazi Jews.** *Gastroenterology* 1989, **96:**1016-1020.
9. Yoshida K, Tamai M, Kubota T, Kawame H, Amano N, Ikeda S, Fukushima Y: **Analysis of 14 individuals who requested predictive genetic testing for hereditary neuromuscular diseases.** *Rinsho Shinkeigaku* 2002, **42:**113-117.
10. Begg C: **On the use of familial aggregation in population-based case probands for calculating penetrance.** *J Natl Cancer Inst* 2002, **94:**1221-1226.
11. Lagakos SW: **Effects of mismodelling and mismeasuring explanatory variables on tests of their association with a response variable.** *Stat Med* 1988, **7:**257-74.
12. Tosteson TD, Tsiatis AA: **The asymptotic relative efficiency of score tests in a generalized linear model with surrogate covariates.** *Biometrika* 1988, **75:**507-514.
13. Begg MD, Lagakos S: **On the consequences of model misspecification in logistic regression.** *Environ Health Perspect* 1990, **87:**69-75.

Pre-publication history

The pre-publication history for this paper can be accessed here:

http://www.biomedcentral.com/1471-2350/5/18/prepub