


METHODOLOGY ARTICLE

Open Access



A statistical measure for the skewness of X chromosome inactivation based on case-control design

Peng Wang^{1†}, Yu Zhang^{1†}, Bei-Qi Wang¹, Jian-Long Li¹, Yi-Xin Wang², Dongdong Pan³, Xian-Bo Wu⁴, Wing Kam Fung⁵ and Ji-Yuan Zhou^{1*} 

Abstract

Background: Skewed X chromosome inactivation (XCI), which is a non-random process, is frequently observed in both healthy and affected females. Furthermore, skewed XCI has been reported to be related to many X-linked diseases. However, no statistical method is available in the literature to measure the degree of the skewness of XCI for case-control design. Therefore, it is necessary to develop methods for such a task.

Results: In this article, we first proposed a statistical measure for the degree of XCI skewing by using a case-control design, which is a ratio of two logistic regression coefficients after a simple reparameterization. Based on the point estimate of the ratio, we further developed three types of confidence intervals (the likelihood ratio, Fieller's and delta methods) to evaluate its variation. Simulation results demonstrated that the likelihood ratio method and the Fieller's method have more accurate coverage probability and more balanced tail errors than the delta method. We also applied these proposed methods to analyze the Graves' disease data for their practical use and found that rs3827440 probably undergoes a skewed XCI pattern with 68.7% of cells in heterozygous females having the risk allele *T* active, while the other 31.3% of cells keeping the normal allele *C* active.

Conclusions: For practical application, we suggest using the Fieller's method in large samples due to the non-iterative computation procedure and using the LR method otherwise for its robustness despite its slightly heavy computational burden.

Keywords: X chromosome inactivation, Skewness, Case-control design, Confidence interval, Graves' disease

Background

X chromosome inactivation (XCI) is an epigenetic phenomenon. Under XCI, one of two X chromosomes in females is silenced during early embryonic development to achieve dosage compensation between two sexes [1]. As such, the genetic effect of two risk alleles in females is expected to be equivalent to that of one risk allele in males. Most of X-linked genes undergo XCI and only about 15% of genes on X chromosome escape from XCI (XCI-E)

[2]. Both alleles in the genes under XCI-E will be active, which are similar to autosomal genes. Generally, XCI has been treated as random (XCI-R) where both maternal and paternal X chromosomes have equal chance to be inactivated, i.e. for an X-linked gene, nearly 50% of the cells have one allele active while the remaining cells have the other allele active. However, recent studies have revealed that the skewed XCI (XCI-S) is a biological plausibility and even a common feature in both healthy and affected females [3–5]. XCI-S is a non-random process, which has been defined as a significant deviation from XCI-R, for instance, the inactivation of one of the alleles in more than 75% of cells [6–8].

The mechanism of XCI-S remains mysterious and XCI-S in human may be likely caused by secondary selection

*Correspondence: zhoujiyuan5460@hotmail.com

[†]Peng Wang and Yu Zhang contributed equally to this work.

¹State Key Laboratory of Organ Failure Research, Ministry of Education, and Guangdong Provincial Key Laboratory of Tropical Disease Research, Department of Biostatistics, School of Public Health, Southern Medical University, No. 1023, South Shatai Road, Baiyun District, Guangzhou 510515, China

Full list of author information is available at the end of the article



[6, 9, 10]. Specifically, the initial choice of active X chromosome is considered as random. However, during body growth, when an X-linked mutation affects cells proliferation or survival, there will be a larger or smaller proportion of cells with an active mutant allele. Due to the selection pressure, this type of secondary skewing varies in different tissues and is also associated with age. For heterozygous females, positive selection cells with mutant allele will lead to more severe expression of the disease, whereas negative selection cells with mutant allele can provide protection from deleterious effects [11, 12]. For example, in heterozygous females with a mutant FoxP3 allele, the XCI-S against the mutant allele in specific tissues can prevent autoimmune disease, whereas the XCI-S towards the mutant allele in breast epithelial cells can result in breast cancer [13]. On the other hand, some diseases, such as ovarian cancer, Rett syndrome, Klinefelter syndrome, and recurrent miscarriages, are reported to be related to XCI-S [14–17]. Therefore, it is necessary to develop methods for measuring such XCI skewing.

Recently, there has been an increasing interest to incorporate the information on XCI into X-chromosome genetic association studies [18–23]. Clayton's method first takes XCI-R into account and treats males as homozygous females [18]. In this regard, two genotypes of males are coded as 0 or 2, while three genotypes of females are coded as 0, 1 or 2, respectively. This coding strategy also implies that the genetic effect of heterozygous genotype in females lies midway between two homozygous genotypes, which seems reasonable as in heterozygous females about half of cells express the mutant allele while the rest of cells express the normal allele. However, this method does not consider the XCI-E and XCI-S patterns. So, a resampling-based method was proposed by maximizing the likelihood ratio (LR) over all the three biological patterns (XCI-E, XCI-R and XCI-S), where the three genotypes of females

are coded as 0, γ or 2 under XCI-S [21]. Note that γ is an unknown parameter which is used to measure the degree of XCI skewing. For instance, $\gamma = 1$ represents XCI-R; $\gamma = 1.5$ indicates XCI-S where 75% of the cells have the mutant allele active, whereas the other 25% of the cells have the normal allele active. On the other hand, the detection of XCI-S is either by measuring the level of methylation or by integrative analysis of whole exome and RNA sequencing data [24, 25]. Although Xu et al. has recently developed a statistical measure for the skewness of XCI based on family trios [26], there is still no statistical method available in the literature to measure the skewness of XCI for case-control design.

Therefore, in this article, we first showed that γ can be represented as a ratio of two logistic regression coefficients after a simple reparameterization, based on case-control data. We then obtained the point estimate of γ by the maximum likelihood estimates (MLEs) of these two regression coefficients. Further, we derived the confidence interval (CI) of γ by the delta method, the Fieller's method and the LR method. We also applied all the proposed approaches to analyze the Graves' disease data for their practical use.

Results

Statistical properties of confidence interval

Tables 1 and 2 list the estimated coverage probability (CP), left tail error (ML), right tail error (MR) (missing the true value of γ), $ML/(ML+MR)$ and proportion of the discontinuous CIs (DP) of the LR, Fieller's and delta methods under various simulation settings with $N = 500$, $\rho = 0$, and $\lambda_2 = 1.5$ and 2, respectively. From the tables, the LR and Fieller's methods control the CP well when $p = 0.3$. However, when $p = 0.1$, both the LR and Fieller's methods appear to overestimate the CP except for $\gamma = 0$. Note that the CPs of the LR method are closer to the pre-set level

Table 1 Estimated CP (%), ML (%), MR (%), $ML/(ML+MR)$ and DP (%) of the two-sided 95% CI when $N = 500$, $\rho = 0$ and $\lambda_2 = 1.5$ for the LR, Fieller's and delta methods

p	γ	LR				Fieller				Delta			
		CP	(ML, MR)	$\frac{ML}{ML+MR}$	DP	CP	(ML, MR)	$\frac{ML}{ML+MR}$	DP	CP	(ML, MR)	$\frac{ML}{ML+MR}$	DP
0.1	0	94.75	(4.78, 0.47)	0.91	1.96	94.93	(4.89, 0.18)	0.96	2.28	100	(0, 0)	—	0
	0.5	95.78	(1.49, 1.65)	0.47	1.44	96.81	(1.58, 0.62)	0.72	1.49	95.72	(0, 4.28)	0	0
	1	96.13	(0.90, 2.33)	0.28	1.82	98.18	(0.73, 0.60)	0.55	1.82	79.38	(0, 20.62)	0	0
	1.5	96.39	(0.64, 2.84)	0.18	1.69	98.72	(0.37, 0.75)	0.33	1.85	70.83	(0, 29.17)	0	0
	2	96.68	(0, 3.32)	0	2.35	98.99	(0, 1.01)	0	2.49	67.71	(0, 32.29)	0	0
0.3	0	94.72	(3.88, 1.40)	0.73	1.27	94.75	(3.87, 1.38)	0.74	1.31	99.22	(0.78, 0)	1	0
	0.5	95.21	(2.45, 1.96)	0.56	0.73	95.25	(2.44, 1.93)	0.56	0.73	99.75	(0.02, 0.23)	0.08	0
	1	94.77	(1.99, 2.85)	0.41	0.69	94.88	(1.98, 2.78)	0.42	0.65	96.70	(0, 3.30)	0	0
	1.5	95.22	(1.26, 3.33)	0.27	0.88	95.39	(1.25, 3.14)	0.28	0.91	92.29	(0, 7.71)	0	0
	2	94.72	(0.37, 4.91)	0.07	1.29	95.16	(0.34, 4.50)	0.07	1.36	88.11	(0, 11.89)	0	0

Table 2 Estimated CP (%), ML (%), MR (%), ML/(ML+MR) and DP (%) of the two-sided 95% CI when $N = 500$, $\rho = 0$ and $\lambda_2 = 2$ for the LR, Fieller's and delta methods

ρ	γ	LR				Fieller				Delta			
		CP	(ML, MR)	$\frac{ML}{ML+MR}$	DP	CP	(ML, MR)	$\frac{ML}{ML+MR}$	DP	CP	(ML, MR)	$\frac{ML}{ML+MR}$	DP
0.1	0	94.59	(4.59, 0.82)	0.85	1.97	94.78	(4.71, 0.51)	0.90	2.72	100	(0, 0)	—	0
	0.5	95.24	(2.12, 1.96)	0.52	1.10	96.20	(2.17, 0.86)	0.72	1.28	94.44	(0, 5.56)	0	0
	1	96.07	(1.39, 2.38)	0.37	1.14	97.78	(1.27, 0.73)	0.64	1.23	84.05	(0, 15.95)	0	0
	1.5	96.56	(0.99, 2.39)	0.29	0.90	98.37	(0.85, 0.72)	0.54	0.98	81.08	(0, 18.92)	0	0
	2	96.49	(0.01, 3.50)	0	0.43	98.37	(0.01, 1.62)	0.01	0.43	79.71	(0, 20.29)	0	0
0.3	0	94.88	(2.80, 2.32)	0.55	0.60	94.93	(2.79, 2.28)	0.55	0.61	98.10	(1.89, 0.01)	0.99	0
	0.5	95.01	(2.40, 2.54)	0.49	0.14	95.01	(2.40, 2.53)	0.49	0.15	99.13	(0.20, 0.67)	0.23	0
	1	95.15	(2.10, 2.71)	0.44	0.08	95.28	(2.09, 2.60)	0.45	0.09	96.37	(0, 3.63)	0	0
	1.5	94.81	(2.03, 3.14)	0.39	0.21	95.03	(2.05, 2.90)	0.41	0.22	92.51	(0, 7.49)	0	0
	2	94.88	(1.80, 3.32)	0.35	0.08	95.18	(1.78, 3.04)	0.37	0.08	91.80	(0, 8.20)	0	0

than the Fieller's method, which indicates the robustness property of the LR method for relatively small samples. Besides, the delta method generally has the worst CP under all the situations. When $p = 0.1$, the delta method overestimates the CP for $\gamma = 0$, while underestimates the CP for $\gamma = 1, 1.5$ and 2 , irrespective of λ_2 being 1.5 or 2 . When p increases from 0.1 to 0.3 , the delta method overestimates the CP for $\gamma = 0, 0.5$ and 1 , while underestimates the CP for $\gamma = 1.5$ and 2 , regardless of $\lambda_2 = 1.5$ or 2 . From the estimated ML, MR and ML/(ML+MR) values, we find that ML and MR of the delta-type CIs are not balanced since nearly all the values of ML/(ML+MR) are far away from 0.5 , while the LR and Fieller's methods have more balanced ML and MR than the delta method, especially when $p = 0.3$. On the other hand, we see that the values of the DP for both the LR and Fieller's methods are not over 3% under our simulation settings. Further, when p increases from 0.1 to 0.3 or λ_2 changes from 1.5 to 2 , DP will generally become smaller.

Tables 3 and 4 give the estimated CP, ML, MR, ML/(ML+MR) and DP of the LR, Fieller's and delta methods when $N = 2000$, $\rho = 0$, and $\lambda_2 = 1.5$ and 2 , respectively. When N increases from 500 to 2000 , the LR method has similar performance with the Fieller's method. It can be seen from the tables that the CPs of all the methods are more accurate. Both the LR and Fieller's methods control the CP well, while the delta method still generally has the poor CP, especially when $p = 0.1$. Note that when $p = 0.3$, all the values of ML/(ML+MR) for the LR method and the Fieller's method are around 0.5 regardless of λ_2 being 1.5 or 2 . But when $p = 0.1$, the values of ML/(ML+MR) for the LR method and the Fieller's method are deviated from 0.5 , especially when γ_0 are 0 and 2 . Further, the Fieller's method has slightly more balanced tail errors than the LR method when $p = 0.1$. In addition, the delta method has the most unbalanced tail errors. We also find that the values of the DP generally decrease to be less than 2% when N increases to be 2000 . In general, the LR

Table 3 Estimated CP (%), ML (%), MR (%), ML/(ML+MR) and DP (%) of the two-sided 95% CI when $N = 2000$, $\rho = 0$ and $\lambda_2 = 1.5$ for the LR, Fieller's and delta methods

ρ	γ	LR				Fieller				Delta			
		CP	(ML, MR)	$\frac{ML}{ML+MR}$	DP	CP	(ML, MR)	$\frac{ML}{ML+MR}$	DP	CP	(ML, MR)	$\frac{ML}{ML+MR}$	DP
0.1	0	94.89	(4.31, 0.80)	0.84	2.04	94.91	(4.36, 0.73)	0.86	2.04	99.98	(0.01, 0.01)	0.50	0
	0.5	94.38	(2.10, 2.84)	0.43	1.03	94.78	(2.13, 2.43)	0.47	0.96	93.29	(0, 6.71)	0	0
	1	95.09	(1.53, 3.19)	0.32	0.73	95.88	(1.52, 2.43)	0.38	0.71	84.16	(0, 15.84)	0	0
	1.5	94.49	(1.24, 4.21)	0.23	0.56	95.59	(1.25, 3.13)	0.29	0.56	81.18	(0, 18.82)	0	0
	2	94.68	(0.01, 5.31)	0	0.22	95.81	(0.01, 4.18)	0	0.22	81.01	(0, 18.99)	0	0
0.3	0	95.23	(2.59, 2.18)	0.54	0.35	95.23	(2.59, 2.18)	0.54	0.33	97.89	(2.07, 0.04)	0.98	0
	0.5	95.00	(2.66, 2.31)	0.54	0.07	95.01	(2.66, 2.30)	0.54	0.07	99.08	(0.12, 0.80)	0.13	0
	1	94.80	(2.60, 2.56)	0.50	0.07	94.86	(2.60, 2.51)	0.51	0.07	95.91	(0, 4.09)	0	0
	1.5	95.05	(2.16, 2.79)	0.44	0.03	95.10	(2.16, 2.74)	0.44	0.03	93.58	(0, 6.42)	0	0
	2	94.58	(2.34, 3.08)	0.43	0.01	94.68	(2.34, 2.98)	0.44	0.01	91.86	(0, 8.14)	0	0

Table 4 Estimated CP (%), ML (%), MR (%), ML/(ML+MR) and DP (%) of the two-sided 95% CI when $N = 2000$, $\rho = 0$ and $\lambda_2 = 2$ for the LR, Fieller's and delta methods

ρ	γ	LR				Fieller				Delta			
		CP	(ML, MR)	$\frac{ML}{ML+MR}$	DP	CP	(ML, MR)	$\frac{ML}{ML+MR}$	DP	CP	(ML, MR)	$\frac{ML}{ML+MR}$	DP
0.1	0	95.45	(2.89, 1.66)	0.64	1.57	95.48	(2.89, 1.63)	0.64	1.66	99.81	(0.14, 0.05)	0.74	0
	0.5	94.92	(2.24, 2.75)	0.45	0.18	95.27	(2.28, 2.34)	0.49	0.23	92.63	(0, 7.37)	0	0
	1	94.39	(2.38, 3.21)	0.43	0.12	95.31	(2.39, 2.28)	0.51	0.11	88.87	(0, 11.13)	0	0
	1.5	94.77	(1.56, 3.67)	0.30	0	95.77	(1.61, 2.62)	0.38	0	87.92	(0, 12.08)	0	0
	2	94.45	(0.42, 5.13)	0.08	0	95.57	(0.42, 4.01)	0.09	0	87.32	(0, 12.68)	0	0
0.3	0	95.02	(2.68, 2.30)	0.54	0.01	95.03	(2.67, 2.30)	0.54	0.01	96.64	(3.08, 0.28)	0.92	0
	0.5	94.97	(2.37, 2.66)	0.47	0	94.97	(2.37, 2.66)	0.47	0	96.94	(1.04, 2.02)	0.34	0
	1	95.13	(2.40, 2.47)	0.49	0	95.17	(2.40, 2.43)	0.50	0	96.05	(0.16, 3.79)	0.04	0
	1.5	94.68	(2.78, 2.54)	0.52	0	94.76	(2.79, 2.45)	0.53	0	94.65	(0, 5.35)	0	0
	2	94.86	(2.53, 2.61)	0.49	0	94.89	(2.56, 2.55)	0.50	0	93.81	(0, 6.19)	0	0

method and the Fieller's method control the CP well with the relatively balanced tail errors on the left and on the right. All the other results of CP, ML, MR, ML/(ML+MR) and DP with $\rho = 0.05$ are given in Tables S1-S4 [see Additional file 1], which are similar to those in Tables 1, 2, 3 and 4 except for $N = 500$ and $p = 0.1$, indicating that Hardy-Weinberg disequilibrium has limited effect on the results. Notice that under the scenario of $N = 500$ and $p = 0.1$, we observe that the CPs of all the methods in Additional file 1: Tables S1 and S2 are better than those in Tables 1 and 2, respectively. One possible explanation of this phenomenon is that the genotype frequency of AA in the control sample increases from 0.01 to 0.0145 when ρ changes from 0 to 0.05.

Sizes and powers

We also simulated the corresponding size and power for testing $\gamma = \gamma_0$ [see Appendix B of Additional file 1]. The size results are given in Additional file 1: Tables S5-S8 and the power results are displayed in Figures S1-S12 [see Additional file 1]. It can be seen that the LR method and the Fieller's method control the size well except for $N = 500$ and $p = 0.1$, while the size of the delta method is either conservative or inflated. On the other hand, the power of the LR method and the Fieller's method are close to each other, but the LR method is generally slightly more powerful than the Fieller's method. However, the power of the delta method can be quite different from those of the LR and Fieller's methods.

Application to Graves' disease data

The GPR174 gene is located on X chromosome, which is associated with autoimmune thyroid disease, including Graves' disease [27]. An X chromosome genome-wide association study (GWAS) was conducted by Chu et al. to study the association between the GPR174 gene and

Graves' disease among Han population [27]. In this study, 14,141 single nucleotide polymorphisms on X chromosome were genotyped. Among them, rs3827440 is a non-synonymous single nucleotide polymorphism within the GPR174 gene, with the minor allele frequency being 0.45 in this population, and thus is a functional variant of interest. Further, statistical analysis of both the GWAS data and the replication data showed that rs3827440 is statistically significantly associated with Graves' disease. At rs3827440, there are two alleles T and C , where T is the susceptible allele which is associated with a higher expression level of the GPR174 gene. Several studies [7, 15] showed that XCI-S is associated with autoimmune thyroid disease. So, we applied the LR, Fieller's and delta methods to explore if rs3827440 undergoes an XCI-S pattern. We only selected the females to estimate the degree of XCI skewing as well as its 95% CI. In the GWAS stage, 2242 females were sampled (1115 cases and 1127 controls). In the case group, the numbers of females with genotypes CC , TC , and TT are 163, 508, and 444, respectively. Those in the control group are 219, 541, and 367, respectively. In the replication stage, 6260 females were sampled with genotype counts 471, 1606, and 1298 in the case group, and 584, 1344, and 957 in the control group for CC , TC , and TT , respectively. The estimated allele frequency of T in females is 0.57 and 0.56 for the GWAS stage and the replication stage, respectively. We applied each of the proposed methods to the data in the GWAS stage and those in the replication stage. After that, we used our proposed methods to deal with the pooled data, by incorporating the stage as a covariate.

Table 5 gives the point estimate $\hat{\gamma}$ and its 95% CIs, based on the LR, Fieller's and delta methods. From the table, we observe that the LR-type CIs and the Fieller's CIs are almost the same. The delta-type CIs are nested within the LR-type and Fieller's CIs for the replication stage and the

Table 5 Statistical inference for γ at rs3827440 in females based on the LR, Fieller's and delta methods

Stage	$\hat{\gamma}$	95% CI		
		LR	Fieller	Delta
GWAS	0.957	[0, 1.657]	[0, 1.658]	[0.241, 1.672]
Replication	1.513	[1.123, 1.930]	[1.122, 1.930]	[1.126, 1.900]
Pooled	1.373	[1.028, 1.719]	[1.028, 1.719]	[1.037, 1.708]

pooled data, which may be caused by the fact that the delta method underestimates the CP. We also find that the LR-type CI and the Fieller's CI are asymmetrical around its point estimate in the GWAS stage but are nearly symmetrical around the point estimate in the replication stage and the pooled analysis, which is probably due to the larger sample size in the replication stage and the pooled dataset. In the GWAS stage, the point estimate $\hat{\gamma}$ is 0.957. All of the three types of CIs contain 1 (XCI-R). In the replication stage, the point estimate $\hat{\gamma}$ is 1.513 and all the CIs do not contain 1. The results in the replication stage suggest the XCI-S pattern at rs3827440 with 75.7% (1.513/2) of cells having the risk allele *T* active and the other 24.3% of cells having the normal allele *C* active. Note that the statistical results for both two stage data suggest different XCI patterns. One possible reason is that the variance of $\hat{\gamma}$ is larger in the GWAS data and there may exist study heterogeneity between those two stages. The results for the pooled data give the point estimate $\hat{\gamma} = 1.373$ by adjusting the stage and all of the three types of CIs do not contain 1. This demonstrates that rs3827440 probably undergoes the XCI-S pattern with 68.7% (1.373/2) of cells keeping the risk allele *T* active, while the other 31.3% of cells keeping the normal allele *C* active. However, this observation needs to be further confirmed by functional analysis of this variant.

Discussion

In this article, we proposed a statistical measure to estimate the degree of the skewness of XCI (i.e. γ). We first showed that γ can be expressed as a ratio of two logistic regression coefficients. Then, we constructed a ratio estimate $\hat{\gamma}$ for γ and also derived three types of CIs (the LR, Fieller's and delta methods) to evaluate its variation. The delta method is a simple and non-iterative procedure but generally has poor statistical properties, which is probably caused by the skewness of $\hat{\gamma}$. On the other hand, the LR method and the Fieller's method are based on a simple reparameterization procedure and thus does not require the normality assumption of $\hat{\gamma}$. The simulation results demonstrate that the LR method and the Fieller's method have better performance than the delta method. On the other hand, note that the LR-type CI will be close to the Fieller's CI when N is large. In this regard, the Fieller's

CI is preferential since it is a non-iterative procedure. However, when N is relatively small, the LR-type CI is recommended for its robustness. In addition, our software SkewXCI is freely available at <http://www.echobelt.org/web/UploadFiles/SkewXCI.html>, which is implemented in R (<http://www.r-project.org/>, version 3.5.1).

Our proposed methods have several limitations. First, our methods assume that the genetic effect of the mutant allele among all the cells is additive on the disease. On the other hand, notice that Model (1) under XCI is different from genetic models (dominant, additive, and recessive) on autosomes or on X chromosome under XCI-E. Specifically, genetic model defines the relationship between two alleles at a locus and usually varies from locus to locus [28–30]. However, when XCI occurs, only one allele is active at each locus in each cell and most of the loci share the same XCI pattern. As such, the magnitude of $\gamma/2$ is a measure of the proportion of cells with the mutant allele active among all the cells in heterozygous females. For instance, adrenoleukodystrophy has been previously viewed as an X-linked recessive disorder where the female carriers are commonly thought to be normal or only mildly affected [31]. However, a recent study showed that the heterozygous females with adrenoleukodystrophy have a wide spectrum of clinical manifestations, ranging from mild to severe phenotypes, which is probably due to the various degree of XCI-S towards the mutant allele. Second, we simply cut the estimated CI within the interval $[0, 2]$ and this may lead to potential loss of information. However, if we incorporate this interval constraint into statistical inference, then the LR, Fieller's and delta methods no longer follow a simple chi-square distribution or a standard normal distribution due to the boundary problem [32]. An alternative method is the Bayesian inference, where such constraint can be regarded as prior information. For instance, when no other information is available, we can choose an uniform prior distribution within the interval $[0, 2]$ for γ . Once the posterior distribution is derived, its percentiles or variance can be used to construct the corresponding CI. Third, note that the validity of our proposed measure is based on the assumption that there exists association between disease and allele *A*. Therefore, in GWAS, we can first screen the associated single nucleotide polymorphisms as candidate loci before making any inference about γ . If such association is not statistically significant, our proposed methods may not be reliable. In this situation, according to Fieller's theorem, the Fieller's CI and the LR-type CI can be discontinuous as shown in Tables 1, 2, 3, and 4, which is difficult to interpret.

Generally, the LR method and the Fieller's method have accurate CP and control the ML and MR well, and hence are recommended in practical application. In future work, we will incorporate the information on the

interval constraint into analysis so as to further improve the efficiency of the proposed methods. Moreover, we will generalize our methods to quantitative traits.

Conclusions

When the sample size is greater than 2000, the Fieller’s method has similar performance to the LR method and thus is preferable due to the non-iterative computation procedure. However, the LR method is recommended otherwise because it has better statistical properties, especially in small samples.

Methods

Point estimation for γ

Consider an X-linked diallelic locus with normal allele a and mutant allele A . We only select the females because XCI is unrelated to males. For females, suppose that aa , Aa and AA are three genotypes and let $X = \{0, \gamma, 2\}$ be the corresponding genotypic value, respectively, with $\gamma \in [0, 2]$. For a case-control design, let $Y = 1$ (0) denote that the female is affected (unaffected). Then, the association between Y and X can be expressed using a logistic regression model

$$\text{Logit}(\Pr(Y = 1|X, \mathbf{z})) = \beta_0 + \beta X + \mathbf{b}^T \mathbf{z}, \tag{1}$$

where β_0 is the intercept, β is the regression coefficient for X , \mathbf{z} is a vector of covariates that need to be adjusted (e.g. age), and \mathbf{b}^T is a vector of regression coefficients for \mathbf{z} .

To estimate γ , we decompose the genotypic value X as $X = \gamma X_1 + (2 - \gamma)X_2$, where $X_1 = I_{\{G=Aa \text{ or } AA\}}$, $X_2 = I_{\{G=AA\}}$, G denotes the genotype of the female and $I_{\{\cdot\}}$ is the indicator function. It can be seen that X_1 indicates if the genotype contains the mutant allele A and X_2 represents if the genotype is the homozygote AA . As such, Model (1) becomes

$$\begin{aligned} \text{Logit}(\Pr(Y = 1|X_1, X_2, \mathbf{z})) \\ = \beta_0 + \beta \gamma X_1 + \beta(2 - \gamma)X_2 + \mathbf{b}^T \mathbf{z}. \end{aligned}$$

Let $\beta_1 = \beta \gamma$ and $\beta_2 = \beta(2 - \gamma)$. Then, the above model can be rewritten as

$$\begin{aligned} \text{Logit}(\Pr(Y = 1|X_1, X_2, \mathbf{z})) \\ = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \mathbf{b}^T \mathbf{z}. \end{aligned} \tag{2}$$

Further, due to this reparameterization, γ can be represented as

$$\gamma = \frac{2\beta_1}{\beta_1 + \beta_2}, \tag{3}$$

when $\beta = (\beta_1 + \beta_2)/2 \neq 0$. γ can only be well defined in presence of the association between the disease and the allele A . Note that $\gamma \in [0, 2]$ means that β_1 and β_2 have the same sign. That is, the genetic effect of heterozygous genotype in females lies between those of two

homozygotes, which is generally satisfied in real applications. From Eq. (3), we have $\gamma = 0$ (2) if and only if $\beta_1 = 0$ and $\beta_2 \neq 0$ ($\beta_1 \neq 0$ and $\beta_2 = 0$), representing XCI-S fully towards the normal (mutant) allele a (A), while $\gamma = 1$ if and only if $\beta_1 = \beta_2 \neq 0$, which means XCI-R. So, if we get the MLEs $\hat{\beta}_1$ and $\hat{\beta}_2$ of β_1 and β_2 , then $\hat{\gamma} = 2\hat{\beta}_1/(\hat{\beta}_1 + \hat{\beta}_2)$ is the MLE of γ by the invariance property of MLE.

Note that $\hat{\beta}_1$ and $\hat{\beta}_2$ can be easily calculated through the standard logistic regression procedure. Specifically, suppose that we collect N unrelated females from a homogeneous population. Then, the log-likelihood function of the sample can be written as

$$\begin{aligned} l_1(\beta_0, \beta_1, \beta_2, \mathbf{b}^T) \\ = \sum_{i=1}^N \left[y_i \left(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \mathbf{b}^T \mathbf{z}_i \right) \right. \\ \left. - \log \left(1 + \exp \left(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \mathbf{b}^T \mathbf{z}_i \right) \right) \right], \end{aligned}$$

where y_i , x_{i1} , x_{i2} and \mathbf{z}_i respectively are the values of Y , X_1 , X_2 and \mathbf{z} of female i . Then, $\hat{\beta}_1$ and $\hat{\beta}_2$ are obtained by maximizing the above log-likelihood function, i.e.

$$l_1(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\mathbf{b}}^T) = \underset{\beta_0, \beta_1, \beta_2, \mathbf{b}^T}{\text{argmax}} l_1(\beta_0, \beta_1, \beta_2, \mathbf{b}^T),$$

where $\hat{\beta}_0$ and $\hat{\mathbf{b}}^T$ are the MLEs of β_0 and \mathbf{b}^T , respectively.

Confidence interval of γ based on delta method

Once the point estimate of γ is derived, we need to calculate the standard error or CI to evaluate its precision. Since $\hat{\gamma}$ is also a ratio estimate, a natural idea is to use the first order Taylor series expansion of $\hat{\gamma}$ and then obtain its asymptotic variance. Specifically, by the consistency of MLE, $\hat{\beta}_1$ and $\hat{\beta}_2$ are close to β_1 and β_2 , respectively, when N is large. Note that $\beta = (\beta_1 + \beta_2)/2$ and thus γ can be rewritten as β_1/β . Making a first order Taylor expansion of $\hat{\gamma}$ around the point (β_1, β) and evaluating this at $(\hat{\beta}_1, \hat{\beta})$, we have

$$\hat{\gamma} \approx \frac{\beta_1}{\beta} + (\hat{\beta}_1 - \beta_1) \frac{1}{\beta} - (\hat{\beta} - \beta) \frac{\beta_1}{\beta^2},$$

where $\hat{\beta} = (\hat{\beta}_1 + \hat{\beta}_2)/2$. Taking variance from both sides, the above equation becomes

$$\begin{aligned} \text{Var}(\hat{\gamma}) \approx \frac{1}{\beta^2} \text{Var}(\hat{\beta}_1) \\ + \frac{\beta_1^2}{\beta^4} \text{Var}(\hat{\beta}) - \frac{2\beta_1}{\beta^3} \text{Cov}(\hat{\beta}_1, \hat{\beta}). \end{aligned} \tag{4}$$

Notice that

$$\text{Var}(\hat{\beta}) = \frac{1}{4} \left(\text{Var}(\hat{\beta}_1) + \text{Var}(\hat{\beta}_2) + 2 \text{Cov}(\hat{\beta}_1, \hat{\beta}_2) \right)$$

and

$$\text{Cov}(\hat{\beta}_1, \hat{\beta}) = \frac{1}{2} \left(\text{Var}(\hat{\beta}_1) + \text{Cov}(\hat{\beta}_1, \hat{\beta}_2) \right),$$

where $\text{Var}(\hat{\beta}_1)$, $\text{Var}(\hat{\beta}_2)$ and $\text{Cov}(\hat{\beta}_1, \hat{\beta}_2)$ are the elements of the variance-covariance matrix V of $\hat{\beta}_1$ and $\hat{\beta}_2$. Generally, V has no simple form when covariates are included in the model, but can be derived from the empirical Fisher's information matrix \hat{I} for $(\beta_0, \beta_1, \beta_2, \mathbf{b}^T)^T$ [33]. For Model (2),

$$\hat{I} = \mathbf{U}^T \hat{W} \mathbf{U},$$

where $\mathbf{U} = (\mathbf{1}, \mathbf{X}_1, \mathbf{X}_2, \mathbf{z})$ is the design matrix, $\mathbf{X}_1 = (x_{11}, x_{21}, \dots, x_{N1})^T$, $\mathbf{X}_2 = (x_{12}, x_{22}, \dots, x_{N2})^T$, $\mathbf{z} = (z_1, z_2, \dots, z_N)^T$, $\hat{W} = \text{diag}(\hat{w}_1, \hat{w}_2, \dots, \hat{w}_N)$ is a diagonal matrix with diagonal elements

$$\hat{w}_i = \hat{f}_i (1 - \hat{f}_i) \quad (i = 1, 2, \dots, N),$$

and

$$\hat{f}_i = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \hat{\mathbf{b}}^T \mathbf{z}_i)}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \hat{\mathbf{b}}^T \mathbf{z}_i)}$$

represents the estimated penetrances for female i . Once \hat{I} is estimated, the partial information matrix \hat{I}_1 for β_1 and β_2 given β_0 and \mathbf{b}^T can be computed and thus $V = \hat{I}_1^{-1}$. If there is no covariate in the model, then V has the following form [see Appendix A of Additional file 1]

$$\begin{pmatrix} \frac{1}{n_{aa}\hat{w}_{aa}} + \frac{1}{n_{Aa}\hat{w}_{Aa}} & -\frac{1}{n_{Aa}\hat{w}_{Aa}} \\ -\frac{1}{n_{Aa}\hat{w}_{Aa}} & \frac{1}{n_{Aa}\hat{w}_{Aa}} + \frac{1}{n_{AA}\hat{w}_{AA}} \end{pmatrix},$$

where n_{aa} , n_{Aa} and n_{AA} are the numbers of the females with aa , Aa and AA , respectively, and $N = n_{aa} + n_{Aa} + n_{AA}$; \hat{w}_{aa} , \hat{w}_{Aa} and \hat{w}_{AA} are the weighted elements for aa , Aa and AA , respectively, with

$$\hat{w}_G = \hat{f}_G (1 - \hat{f}_G) \quad (G = aa, Aa, \text{ or } AA),$$

and

$$\hat{f}_{aa} = \frac{\exp(\hat{\beta}_0)}{1 + \exp(\hat{\beta}_0)},$$

$$\hat{f}_{Aa} = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1)}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1)}$$

and

$$\hat{f}_{AA} = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_2)}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_2)}$$

representing the estimated penetrances for aa , Aa and AA , respectively.

Replacing β and β_1 by $\hat{\beta}$ and $\hat{\beta}_1$ in Eq. (4), we estimate the delta-type standard error [34]

$$\begin{aligned} \hat{\text{Var}}(\hat{\gamma}) &\approx \frac{1}{\hat{\beta}^2} \text{Var}(\hat{\beta}_1) \\ &+ \frac{\hat{\beta}_1^2}{\hat{\beta}^4} \text{Var}(\hat{\beta}) - \frac{2\hat{\beta}_1}{\hat{\beta}^3} \text{Cov}(\hat{\beta}_1, \hat{\beta}). \end{aligned}$$

As such, the delta-type CI (γ_L^d, γ_U^d) at level $(1 - \alpha)$ can be expressed as

$$\left(\hat{\gamma} - Z_{1-\frac{\alpha}{2}} \sqrt{\hat{\text{Var}}(\hat{\gamma})}, \hat{\gamma} + Z_{1-\frac{\alpha}{2}} \sqrt{\hat{\text{Var}}(\hat{\gamma})} \right),$$

where $Z_{1-\alpha/2}$ denotes the $(1 - \alpha/2)$ -quantile of the standard normal distribution. Note that the estimated CI may be out of the range of $[0, 2]$ when the variation is large, which should be cut off. To test the null hypothesis $H_0 : \gamma = \gamma_0$ against the alternative hypothesis $H_1 : \gamma \neq \gamma_0$, we have

$$\frac{\hat{\gamma} - \gamma_0}{\sqrt{\text{Var}(\hat{\gamma})}} \sim N(0, 1)$$

under H_0 , where γ_0 is an arbitrary constant between $[0, 2]$, such as 1 (XCI-R).

The delta method is a non-iterative procedure and thus is easy to be implemented. However, the CI of a ratio estimate is generally skewed, while the delta-type CI is symmetrical [35, 36]. Therefore, it is necessary to propose the Fieller's and likelihood ratio methods to overcome this shortcoming in the following sections.

Confidence interval of γ based on Fieller's method

The Fieller's method is another widely used non-iterative approach for constructing CI for ratio estimate [37]. This type of CI can be asymmetrical around $\hat{\gamma}$. To propose the Fieller's CI, we first need to build a Wald test for testing $\gamma = \gamma_0$. Specifically, under $\gamma = \gamma_0$, we have $\beta_1 - \gamma_0\beta = 0$. Therefore, the Wald test for testing $\gamma = \gamma_0$ can be written as

$$\frac{\hat{\beta}_1 - \gamma_0\hat{\beta}}{\sqrt{\text{Var}(\hat{\beta}_1) + \gamma_0^2 \text{Var}(\hat{\beta}) - 2\gamma_0 \text{Cov}(\hat{\beta}_1, \hat{\beta})}},$$

which follows a standard normal distribution. Then, the confidence limits γ_L^f and γ_U^f ($\gamma_L^f < \gamma_U^f$) for Fieller's CI at level $(1 - \alpha)$ can be found by solving the following equation

$$\frac{\hat{\beta}_1 - \gamma_0\hat{\beta}}{\sqrt{\text{Var}(\hat{\beta}_1) + \gamma_0^2 \text{Var}(\hat{\beta}) - 2\gamma_0 \text{Cov}(\hat{\beta}_1, \hat{\beta})}} = Z_{1-\frac{\alpha}{2}}.$$

Rearranging the above equation yields a quadratic equation with respect to γ_0

$$D\gamma_0^2 + E\gamma_0 + F = 0,$$

where

$$D = \hat{\beta}^2 - Z_{1-\frac{\alpha}{2}}^2 \text{Var}(\hat{\beta}),$$

$$E = 2 \left(Z_{1-\frac{\alpha}{2}}^2 \text{Cov}(\hat{\beta}_1, \hat{\beta}) - \hat{\beta}_1 \hat{\beta} \right)$$

and

$$F = \hat{\beta}_1^2 - Z_{1-\frac{\alpha}{2}}^2 \text{Var}(\hat{\beta}_1).$$

Suppose $\Delta = E^2 - 4DF > 0$, then this equation must have two unequal roots with γ_L^f and γ_U^f being $(-E \pm \sqrt{\Delta})/2D$. According to Fieller's theorem, we know that $D > 0$ implies $\Delta > 0$. In this situation, the Fieller's CI is continuous and can be denoted by (γ_L^f, γ_U^f) .

Note that $D > 0$ is equivalent to $|\hat{\beta}/\sqrt{\text{Var}(\hat{\beta})}| > Z_{1-\frac{\alpha}{2}}$. That is, there exists statistically significant association between the disease and the allele A at the significance level α . However, if there is no such association (i.e. $D < 0$), the Fieller's CI will be unbounded. For instance, if $\Delta < 0$, the Fieller's CI will be $(-\infty, \infty)$. If $\Delta > 0$, the Fieller's CI will be $(-\infty, \gamma_L^f) \cup (\gamma_U^f, \infty)$, which is the discontinuous CI. In real applications, it generally makes little sense to infer about γ if there is no association between the disease and the allele A according to its definition. In addition, the Fieller's CI should also be restricted to the interval $[0, 2]$ when needed.

The Fieller's method usually demonstrates better coverage probability than the delta method. Notice that the Fieller's CI is based on the inversion of the Wald test. Since the LR test is expected to have more robust properties in small samples, so it is desirable to propose the LR method in the next section.

Confidence interval of γ based on likelihood ratio method

To obtain the LR-based CI, we first construct a likelihood ratio test for testing $\gamma = \gamma_0$. As mentioned above, we have derived the MLEs $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$ and $\hat{\mathbf{b}}^T$ of $\beta_0, \beta_1, \beta_2$ and \mathbf{b}^T under H_1 . To calculate the likelihood ratio test statistic λ , we further evaluate the likelihood function under $H_0 : \gamma = \gamma_0$. If H_0 holds, the genotypic value X equals 0, γ_0 and 2 for aa, Aa and AA , respectively. In this regard, Model (1) is reduced to be a standard logistic model and the log-likelihood function under H_0 can be written as

$$l_0(\beta_0, \beta, \mathbf{b}^T) = \sum_{i=1}^N \left[y_i (\beta_0 + \beta x_i + \mathbf{b}^T \mathbf{z}_i) - \log \left(1 + \exp(\beta_0 + \beta x_i + \mathbf{b}^T \mathbf{z}_i) \right) \right],$$

where x_i is the genotypic value of X of female i . Let $\tilde{\beta}_0, \tilde{\beta}$ and $\tilde{\mathbf{b}}^T$ be the MLEs of β_0, β and \mathbf{b}^T under H_0 ,

respectively. Then,

$$l_0(\tilde{\beta}_0, \tilde{\beta}, \tilde{\mathbf{b}}^T) = \underset{\beta_0, \beta, \mathbf{b}^T}{\text{argmax}} l_0(\beta_0, \beta, \mathbf{b}^T),$$

and λ can be computed as

$$\lambda = 2 \left(l_1(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\mathbf{b}}^T) - l_0(\tilde{\beta}_0, \tilde{\beta}, \tilde{\mathbf{b}}^T) \right).$$

λ asymptotically follows a chi-square distribution with the degree of freedom being one (i.e. χ_1^2).

Now, we introduce how to obtain the LR-type CI. For each given γ_0 , we can calculate the corresponding value of the log-likelihood function l_0 and $\tilde{\boldsymbol{\theta}} = (\tilde{\beta}_0, \tilde{\beta}, \tilde{\mathbf{b}}^T)^T$ under H_0 . So, l_0 and $\tilde{\boldsymbol{\theta}}$ are single variable functions with respect to γ_0 and can be denoted by $l_0 = l_0(\gamma_0)$ and $\tilde{\boldsymbol{\theta}} = \tilde{\boldsymbol{\theta}}(\gamma_0)$, respectively. At the significance level α , the confidence limits γ_L^l and γ_U^l ($\gamma_L^l < \gamma_U^l$) of the LR-type CI is determined by the following equation with respect to γ_0

$$l_0(\gamma_0) - l_1(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\mathbf{b}}^T) + \frac{q_{1-\alpha}}{2} = 0, \tag{5}$$

where $q_{1-\alpha}$ denotes the $(1 - \alpha)$ -quantile of the χ_1^2 distribution. Obviously, Eq. (5) has no closed form solutions and numerical method can be adopted. Note that $\boldsymbol{\theta} = (\beta_0, \beta, \mathbf{b}^T)^T$ are nuisance parameters in Eq. (5) which depend on γ_0 . To solve this equation, it generally requires several iterations with different values of γ_0 , and for each γ_0 , the iterative maximization over the remaining parameters is also needed to determine $\tilde{\boldsymbol{\theta}}$. This procedure is relatively time-consuming. Therefore, to reduce the computational burden, borrowing the idea of Venzon and Moolgavkar [38], we can find the roots of Eq. (5) by solving the following system of non-linear equations

$$l_0(\gamma_0, \boldsymbol{\theta}) - l_1(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\mathbf{b}}^T) + \frac{q_{1-\alpha}}{2} = 0$$

$$\frac{\partial l_0}{\partial \boldsymbol{\theta}}(\gamma_0, \boldsymbol{\theta}) = \mathbf{0},$$

which is easily implemented in the commonly used software (e.g. nleqslv package in R). Note that the above system differs only in the first equation from the system (with the first equation being replaced by $\frac{\partial l_0}{\partial \gamma_0}(\gamma_0, \boldsymbol{\theta}) = 0$) that defines the MLEs $\hat{\gamma}$ and $\hat{\boldsymbol{\theta}} = (\hat{\beta}_0, \hat{\beta}, \hat{\mathbf{b}}^T)^T$. Therefore, finding a root of such system almost has the same difficulty as that of finding the MLEs of Model (2) [38]. As such, this algorithm is generally more efficient.

On the other hand, based on the fact that the Wald test and the LR test are asymptotically equivalent in large samples, we know that the confidence limits of the Fieller's CI and the LR-type CI should be close to each other. Therefore, we used the confidence limits of the Fieller's CI as the initial values for γ_0 . For example, when searching the

lower limit, we chose the initial values for γ_0 and θ as γ_L^f and $\tilde{\theta}(\gamma_L^f)$, respectively, where $\tilde{\theta}(\gamma_L^f)$ can be computed from the standard logistic regression procedure. Similarly, we used the same strategy to search the upper limit. The algorithm based on this choice of the initial values works well in most situations. However, in some scenarios, the Fieller's CI and the LR-type CI may be very different. Thus, using the confidence limits of the Fieller's CI as the initial values may cause that the algorithm does not converge. In this regard, we should directly solve the single variable function of Eq. (5). For example, we can use the bisection method to find the roots of Eq. (5) within the interval $[0, 2]$ (e.g. rootSolve package in R).

Like the Fieller's CI, the LR-type CI can be unbounded when there is no association between the disease and the allele A . Specifically, when Equation (5) has no root, then the LR-type CI will be $(-\infty, \infty)$. Otherwise, there will be two roots γ_L^l and γ_U^l . If $\gamma_L^l < \hat{\gamma} < \gamma_U^l$, then the LR-type CI is continuous and can be represented as (γ_L^l, γ_U^l) . If $\hat{\gamma} \notin (\gamma_L^l, \gamma_U^l)$, then the LR-type CI will be $(-\infty, \gamma_L^l) \cup (\gamma_U^l, \infty)$, which is the discontinuous CI. Similar to the delta and Fieller's methods, the LR-type CI is also truncated by $[0, 2]$ if necessary.

The LR-based CI and the Fieller's CI can be asymmetrical which is an appealing choice, compared to the delta method. This is because the distribution of a ratio estimate is generally non-normal with a heavy tail, especially when N is small. Additionally, it will be quite straightforward to incorporate covariates using the LR method.

Simulation settings

For simplicity, we assumed that there is no covariate included in the model in our simulation study. We incorporated the covariate into the real data analysis later. For a case-control design, we presumed that the genotype distribution in the case group and that in the control group of females follow trinomial distributions with probabilities (h_0, h_1, h_2) and (g_0, g_1, g_2) , respectively, where h_0 (g_0), h_1 (g_1) and h_2 (g_2) are the frequencies of aa , Aa and AA in the case (control) group, respectively. Let $h_0/g_0 = m$, $h_1/g_1 = \lambda_1(h_0/g_0) = \lambda_1 m$ and $h_2/g_2 = \lambda_2(h_0/g_0) = \lambda_2 m$, where λ_1 and λ_2 are the odds ratios for Aa and AA compared to aa in females, respectively. By $h_0 + h_1 + h_2 = 1$, we have $m = 1/(g_0 + \lambda_1 g_1 + \lambda_2 g_2)$, $h_0 = g_0/(g_0 + \lambda_1 g_1 + \lambda_2 g_2)$, $h_1 = \lambda_1 g_1/(g_0 + \lambda_1 g_1 + \lambda_2 g_2)$ and $h_2 = \lambda_2 g_2/(g_0 + \lambda_1 g_1 + \lambda_2 g_2)$. Thus, (h_0, h_1, h_2) is calculated by (g_0, g_1, g_2) , λ_1 and λ_2 . The value of (g_0, g_1, g_2) is determined by the allele frequency of A (denoted by $p = 1 - q$) and the inbreeding coefficient ρ , i.e. $g_0 = q^2 + \rho pq$, $g_1 = 2(1 - \rho)pq$ and $g_2 = p^2 + \rho pq$. Specifically, $\rho = 0$ implies that Hardy-Weinberg equilibrium holds in the control group and $\rho \neq 0$ indicates Hardy-Weinberg disequilibrium.

Furthermore, from Model (2), $\beta_0 = \log(m)$, $\beta_1 = \log(\lambda_1)$ and $\beta_1 + \beta_2 = \log(\lambda_2)$. Then, $\gamma = 2 \log(\lambda_1)/\log(\lambda_2)$, i.e. $\lambda_1 = \lambda_2^{\gamma/2}$. As such, we defined the simulation settings as follows: p is fixed at 0.1 and 0.3, and ρ is set to be 0 and 0.05. The true value of γ is fixed at 0, 0.5, 1, 1.5 and 2. λ_2 is assigned to be 1.5 and 2. We selected N as 500 (2000) where both the case and control groups have 250 (1000) females. The confidence level is fixed at $(1 - \alpha) = 95\%$ and the number of replications k is 10,000.

We compared the performance of three types of CIs based on CP, ML, MR, ML/(ML+MR) and DP. CP is defined to be the proportion that the CI contains the true value of γ among k replications, regardless of whether the CI is continuous or not. ML and MR are calculated by $ML = \#[(\gamma_0 < \gamma_L) \cap (\gamma_L \leq \hat{\gamma} \leq \gamma_U)]/k$, and $MR = \#[(\gamma_0 > \gamma_U) \cap (\gamma_L \leq \hat{\gamma} \leq \gamma_U)]/k$, respectively, where $\#$ denotes the counting measure, and γ_L and γ_U are the confidence limits of the estimated CI. Note that $\gamma_L \leq \hat{\gamma} \leq \gamma_U$ means that the CI is continuous. As such, we only consider the continuous CIs when estimating ML and MR, since it is impossible to distinguish between the left side and the right side if the CI is discontinuous. Further, DP is computed as $1 - \#(\gamma_L \leq \hat{\gamma} \leq \gamma_U)/k$. We believed that a good CI should control the CP well as well as have the balanced ML and MR. ML and MR are used together to measure the location of CI. If a balance between ML and MR is achieved, then ML/(ML+MR) is close to 0.5. On the other hand, note that the delta-type CI is always bounded. Therefore, the DP for the delta method will stay at 0. However, for the Fieller's CI and the LR-type CI, small DP is desirable.

Additional file

Additional file 1: Appendices and Supplementary tables and figures.

Appendix A Derivation of the closed form of V without covariates;

Appendix B Size and power results for testing $\gamma = \gamma_0$; **Tables S1-S2**

Estimated CP (%), ML (%), MR (%), ML/(ML+MR) and DP (%) of the two-sided 95% CI when $N = 500$, $\rho = 0.05$, and $\lambda_2 = 1.5$ or 2 for the LR, Fieller's and delta methods, respectively; **Tables S3-S4** Estimated CP (%), ML (%), MR (%), ML/(ML+MR) and DP (%) of the two-sided 95% CI when

$N = 2000$, $\rho = 0.05$, and $\lambda_2 = 1.5$ or 2 for the LR, Fieller's and delta methods, respectively; **Tables S5-S8** Estimated size (%) for testing

$H_0 : \gamma = \gamma_0$ with $N = 500$ or 2000, and $\rho = 0$ or 0.05 for the LR, Fieller's and delta methods, respectively; **Figures S1-S6** Power comparison of the LR, Fieller's and delta methods against γ . The simulation is based on 10,000 replicates with $N = 500$, $\rho = 0$ or 0.05, and $\gamma_0 = 0, 1$ or 2, respectively;

Figures S7-S12 Power comparison of the LR, Fieller's and delta methods against γ . The simulation is based on 10,000 replicates with $N = 2000$, $\rho = 0$ or 0.05, and $\gamma_0 = 0, 1$ or 2, respectively. (PDF 117 kb)

Abbreviations

CI: Confidence interval; CP: Coverage probability; DP: Proportion of discontinuous CIs; GWAS: Genome-wide association study; LR: Likelihood ratio; ML: Left tail error; MLE: Maximum likelihood estimate; MR: Right tail error; XCI: X chromosome inactivation; XCI-E: Escape from X chromosome inactivation; XCI-R: Random X chromosome inactivation; XCI-S: Skewed X chromosome inactivation.

Acknowledgements

The authors would like to thank the editor and three anonymous reviewers for their valuable comments which highly improved the presentation of the article and the authors also appreciate Dr. Yuqiang Xue's kind assistance.

Funding

This work was supported by the National Natural Science Foundation of China [81773544, 81373098, 81573207], the National and Guangdong University Students' Innovation and Enterprise Training Project of China [201612121017], and the General Program of Applied Basic Research Programs of Yunnan Province [2017FB002]. All the funding supporters had no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Availability of data and materials

The dataset supporting the conclusions of this article can be found in the article of Chu et al. [27].

Authors' contributions

PW helped design the study, drafted the article, and conducted the simulation study. YZ helped conduct the simulation study and the data analysis, and design the study's analytic strategy. BQW revised the article critically and helped interpret the results of the data analysis. JLL helped design the study, plot the figures and conduct the literature review. YXW helped interpret the results of the data analysis and prepare the introduction section and the discussion section. DP helped draft the article, helped the analysis and the interpretation of the data, and revised the article. XBW and WKF helped design the study, reviewed the whole paper and critically revised the article. JYZ helped design the study, supervised the field activities, and directed its implementation, including quality assurance and control. All authors read and approved this version of the manuscript.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details

¹State Key Laboratory of Organ Failure Research, Ministry of Education, and Guangdong Provincial Key Laboratory of Tropical Disease Research, Department of Biostatistics, School of Public Health, Southern Medical University, No. 1023, South Shatai Road, Baiyun District, Guangzhou 510515, China. ²Department of Occupational and Environmental Health, School of Public Health, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China. ³Yunnan Key Laboratory of Statistical Modeling and Data Analysis, Yunnan University, Kunming, China. ⁴Department of Epidemiology, School of Public Health, Southern Medical University, Guangzhou, China. ⁵Department of Statistics and Actuarial Science, University of Hong Kong, Hong Kong, China.

Received: 31 July 2018 Accepted: 18 December 2018

Published online: 07 January 2019

References

- Lyon MF. Gene action in the X-chromosome of the mouse (*Mus musculus* L.) *Nature*. 1961;190:372–3.
- Berlitch JB, Ma W, Yang F, Shendure J, Noble WS, Distechi CM, et al. Escape from X inactivation varies in mouse tissues. *PLoS Genet*. 2015;11:e1005079.
- Plenge RM, Stevenson RA, Lubs HA, Schwartz CE, Willard HF. Skewed X-chromosome inactivation is a common feature of X-linked mental retardation disorders. *Am J Hum Genet*. 2002;71:168–73.
- Amos-Landgraf JM, Cottle A, Plenge RM, Friez M, Schwartz CE, Longshore J, et al. X chromosome inactivation patterns of 1005 phenotypically unaffected females. *Am J Hum Genet*. 2006;79:493–9.
- Busque L, Paquette Y, Provost S, Roy DC, Levine RL, Mollica L, et al. Skewing of X-inactivation ratios in blood cells of aging women is confirmed by independent methodologies. *Blood*. 2009;113:3472–4.
- Minks J, Robinson WP, Brown CJ. A skewed view of X chromosome inactivation. *J Clin Invest*. 2008;118:20–3.
- Chabchoub G, Uz E, Maalej A, Mustafa CA, Rebai A, Mnif M, et al. Analysis of skewed X-chromosome inactivation in females with rheumatoid arthritis and autoimmune thyroid diseases. *Arthritis Res Ther*. 2009;11:R106.
- Renault NKE, Pritchett SM, Howell RE, Greer WL, Sapienza C, Ørstavik KH, et al. Human X-chromosome inactivation pattern distributions fit a model of genetically influenced choice better than models of completely random choice. *Eur J Hum Genet*. 2013;21:1396–402.
- Brown CJ. Skewed X-chromosome inactivation: cause or consequence? *J Natl Cancer Inst*. 2010;91:303–4.
- Belmont JW. Genetic control of X inactivation and processes leading to X-inactivation skewing. *Am J Hum Genet*. 1996;58:1101–8.
- Medema RH, Boudewijn MT. The X factor: skewing X inactivation towards cancer. *Cell*. 2007;129:1275–86.
- Deng X, Berlitch JB, Nguyen DK, Distechi CM. X chromosome regulation: diverse patterns in development, tissues and disease. *Nat Rev Genet*. 2014;15:367–78.
- Zuo T, Wang L, Morrison C, Chang X, Zhang H, Li W, et al. FOXP3 is an X-linked breast cancer suppressor gene and an important repressor of the HER-2/ErbB2 oncogene. *Cell*. 2007;129:1275–86.
- Li G, Jin T, Liang H, Tu Y, Zhang W, Gong L, et al. Skewed X-chromosome inactivation in patients with esophageal carcinoma. *Diagn Pathol*. 2013;8:56–62.
- Simmonds MJ, Kawoura PK, Brand OJ, Newby PR, Jackson LE, Hargreaves CE, et al. Skewed X chromosome inactivation and female preponderance in autoimmune thyroid disease: an association study and meta-analysis. *J Clin Endocrinol Metab*. 2013;99:127–31.
- Iitsuka Y, Bock A, Nguyen D, Samango-Sprouse CA, Simpson JL, Bischoff FZ. Evidence of skewed X-chromosome inactivation in 47, XXY and 48, XYYY Klinefelter patients. *Am J Med Genet Part A*. 2001;98:25–31.
- Sangha KK, Stephenson MD, Brown CJ, Robinson WP. Extremely skewed X-chromosome inactivation is increased in women with recurrent spontaneous abortion. *Am J Hum Genet*. 1999;65:913–17.
- Clayton D. Testing for association on the X chromosome. *Biostatistics*. 2008;9:593–600.
- Clayton D. Sex chromosomes and genetic association studies. *Genome Med*. 2009;1:110–6.
- Hickey PF, Bahlo M. X chromosome association testing in genome wide association studies. *Genet Epidemiol*. 2011;35:664–70.
- Wang J, Yu R, Shete S. X-chromosome genetic association test accounting for X-inactivation, skewed X-inactivation, and escape from X-inactivation. *Genet Epidemiol*. 2014;38:483–93.
- Chen Z, Ng HKT, Li J, Liu Q, Huang H. Detecting associated single-nucleotide polymorphisms on the X chromosome in case control genome-wide association studies. *Stat Methods Med Res*. 2017;26:567–82.
- Ma L, Hoffman G, Keinan A. X-inactivation informs variance-based testing for X-linked association of a quantitative trait. *BMC Genomics*. 2015;16:241–9.
- Busque L, Mio R, Mattioli J, Brais E, Blais N, Lalonde Y, et al. Nonrandom X-inactivation patterns in normal females: lyonization ratios vary with age. *Blood*. 1996;88:59–65.
- Szelinger S, Malenica I, Corneveaux JJ, Siniard AL, Kurdoglu AA, Ramsey KM, et al. Characterization of X chromosome inactivation using integrated analysis of whole-exome and mRNA sequencing. *Plos One*. 2014;9:e113036.
- Xu SQ, Zhang Y, Wang P, Liu W, Wu XB, Zhou JY. A statistical measure for the skewness of X chromosome inactivation based on family trios. *BMC Genet*. 2018;19:109.
- Chu X, Shen M, Xie F, Miao XJ, Shou WH, Liu L, et al. An X chromosome-wide association analysis identifies variants in GPR174 as a risk factor for Graves' disease. *J Med Genet*. 2013;50:479–85.

28. Dobyns WB, Filauro A, Tomson BN, Chan AS, Ho AW, Ting NT, et al. Inheritance of most X-linked traits is not dominant or recessive, just X-linked. *Am J Med Genet Part A*. 2004;129:136–43.
29. Biggar RJ, Bergen AW, Poulsen GN. Impact of X chromosome genes in explaining the excess risk of cancer in males. *Am J Epidemiol*. 2009;170:65–71.
30. Ober C, Loisel DA, Gilad Y. Sex-specific genetic architecture of human disease. *Nat Rev Genet*. 2008;9:911–22.
31. Lourenço CM, Simão GN, Santos AC, Marques JW. Sex-specific genetic architecture of human disease. *Arq Neuropsiquiatr*. 2012;70:487–91.
32. Molenberghs G, Verbeke G. Likelihood ratio, score, and Wald tests in a constrained parameter space. *Am Stat*. 2007;61:22–7.
33. Wood SN. *Generalized Additive Models: An Introduction With R*. 1st ed. London: Chapman & Hall Ltd; 2006.
34. Oehlert GW. A note on the delta method. *Am Stat*. 1992;46:27–9.
35. Tin M. Comparison of some ratio estimators. *J Am Stat Assoc*. 1965;60:294–307.
36. Choquet D, L'écuyer P, Léger C. Bootstrap confidence intervals for ratios of expectations. *Acm T Model Comput S*. 1999;9:326–48.
37. Fieller EC. Some problems in interval estimation. *J R Stat Soc B*. 1954;16:175–85.
38. Venzon DJ, Moolgavkar SH. A method for computing profile-likelihood-based confidence intervals. *J R Stat Soc C-Appl*. 1988;37:87–94.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

