Choosing the Right Statistical Analysis to Answer the Research Questions

Made by Billy Witanto, Research Assistant at Eijkman Institute of Molecular Biology

This study is a genetic association study to determine the effect of SNPs towards the aspects of hepatitis B virus (HBV) infection, especially the natural history of chronic hepatitis B (CHB). In genetic association study, we need to conduct a quality control of the genotyping process by Hardy-Weinberg equilibrium (HWE) testing using chi-square goodness of fit. By doing so, the power of our statistical test of linkage and/or association will be trusted, such as when using chi-square test of independence applied to case-control data.

Hosking, L., Lumsden, S., Lewis, K., Yeo, A., McCarthy, L., Bansal, A., Riley, J., Purvis, I., Xu, C. F. 2004. Detection of genotyping errors by Hardy-Weinberg equilibrium testing. *European Journal of Human Genetics*. 12: 395-399.

Salanti, G., Amountza, G., Ntzani, E. E., Ioannidis, J. P. A. 2005. Hardy-Weinberg equilibrium in genetic association studies: an empirical evaluation of reporting, deviations, and power. *European Journal of Human Genetics.* 13: 840-848.

After HWE testing, we need to determine and measure the association of SNPs towards the phase progression of CHB natural history. It can be done using chi-square test of independence, followed by odds ratio (OR) and 95% confidence interval (CI) calculation.

That’s the normal route. But it is important to note the assumptions to conduct the chi-square test. Is it relevant to the research questions that we propose? The assumptions are:

1. Sample observations need to be independent.
2. Constraints on the frequencies should be linear.
3. N-total population size should be large (>50 on minimum basis).
4. No expected frequency should be < 5, as at this level, the distribution loses its continuous character and results in overestimated x^2 value.

Bali, J. & Kant, A. 2017. *Basic of Biostatistics: A Manual for Medical Practitioners*. Jaypee: New Delhi.

One thing that bugs me is chi-square statistic assumes variable to be independent, implying that the occurrence of an event does not affect the outcome of the other. It just tells us about if there is any association or not. But it is not that simple in our genetic association study of single nucleotide polymorphisms (SNPs) towards CHB phase.

Two main pitfalls by using chi-square in this study are:

1. I think we need to place ‘phase’ variable under dependent category, not independent. Because we want to see the outcome of certain genotype of SNPs.
2. There are confounding variables, many of it. By doing chi-square, we ignore that and makes our statistical result lost its power. We should add known covariates/confounding such as age or gender, after bivariate statistical test of course: chi square test of independence.

Szumilas, M. 2010. Explaining Odds Ratios. *J. Can. Acad. Child. Adolesc. Psychiatry*. 19(3): 227-229.

To answer our research question, it is logical to use multivariate regression analysis, either multinomial logistic regression or ordinal logistic regression, depending our dependent variable is sequential or not.

This is from statistical point of view. Let’s see from our research design point of view.

There has been many reviews and meta-analysis toward SNPs, such as Zhang *et al*. (2011) and the newest one, Zhang *et al*. (2019), and many more. All of them said that for further study of SNPs, we need a larger sample size for reproducible result. Small sample size makes inconsistent biased results. And here we are, with only 249 samples with insignificant results. How do we justify this critical fatal repeated error? Do we forget to read review paper of previous studies before conducting a research? It tells us about our poor study design and poor execution of research.

Zhang, T. C., Pan, F. M., Zhang, L. Z., Gao, Y. F., Zhang, Z. H., Gao, J., Ge, R., Mei, Y., Shen, B. B., Duan, Z. H., Li, X. 2011. A meta-analysis of the relation of polymorphism at site -1082 and -592 of the IL-10 gene promoter with susceptibility and clearance to persistent hepatitis B virus infection in the Chinese population. *Infection*. 39: 21-7.

Zhang, Z., Wang, C., Liu, Z., Zou, G., Li, J., Lu, M. 2019. Host Genetic Determinants of Hepatitis B Virus Infection. *Frontiers in Genetics*. 10: 696.

And that’s why I’m working with more advanced statistical analysis, ordinal logistic regression, to give more depths to our poor executed study. Even if the result is not significant, at least we are giving a new perspective in analyzing the CHB natural history progression. It adds something, rather than using an old way of analysis, chi-square.

Interpretation of Ordinal Logistic Regression Result of Single Nucleotide Polymorphisms (SNPs) and Chronic Hepatitis B (CHB) Natural History Progression

Before interpreting the result of our regression analysis, it is wise to check whether our data fits the assumption to run the ordinal logistic regression. The assumptions are:

1. Your dependent variable should be measured at the ordinal level.

   Our dependent variable is phase of CHB. The **phase** of CHB is stages of disease and obviously an ordinal data. OK.

2. One or more independent variables that are continuous, ordinal or categorical.

   Our independent variables include:
   Gender–categorical; **Age**–continuous; **HBeAg**–categorical; **Log Viral Load**–continuous; **ALT Group**–categorical; **IL10-592**–categorical; **TNFA-308**–categorical.

3. There is no **multicollinearity**. Multicollinearity occurs when you have two or more independent variables that are highly correlated with each other.

   Multicollinearity can be detected using collinearity statistics in regression option of SPSS. All independent variables have **VIF** between 1 and 10, which mean there is no multicollinearity.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	2,761	,198		13,930	,000		
	Age_Group	,001	,034	,001	,022	,982	,830	1,205
	Gender	-,052	,104	-,019	-,502	,617	,767	1,304
	HBeAg	-2,459	,101	-,982	-24,425	,000	,699	1,431
	Log_VL	,040	,022	,070	1,803	,075	,756	1,322
	ALT_Group	-,431	,098	-,172	-4,412	,000	,744	1,344
	IL10_592	-,158	,070	-,078	-2,261	,026	,944	1,060
	TNF_308	,173	,138	,044	1,252	,214	,935	1,069

a. Dependent Variable: Phase

4. You have proportional odds, which are fundamental assumption of cumulative odds ordinal regression with proportional odds. The assumption means that each independent variable has an identical effect at each cumulative split of the ordinal dependent variable.

This assumption can be checked in the ordinal logistic regression output–**Test of Parallel Lines**. If the p-value is more than 0.05, then the model fits the assumption. Ours is OK.

Test of Parallel Lines[a]

Model	-2 Log Likelihood	Chi-Square	df	Sig.
Null Hypothesis	496,056			
General	482,964	13,092	8	,109

Because of data fulfills the four assumptions, the validity of our method is proven. After this, we need to see the validity of the result before jumping into odds ratio (OR) and its p-value. Reference for the assumptions can be seen at:

https://statistics.laerd.com/spss-tutorials/ordinal-regression-using-spss-statistics.php

Because ordinal logistic regression is a multivariate analysis. First, we need to choose the statistically significant independent variable associated with phase using chi-square test of independence to be included in multivariate analysis. All independent variables except SNPs are statistically associated with phase of CHB natural history (p-value less than 0.05). But, after careful consideration, we excluded HBeAg, Log Viral Load, and ALT Group independent variables, because they are already reflected in dependent variable – phase. Other than that, when including them in regression analysis, because of many zero frequencies of independent-dependent variable combination, the analysis can't be conducted. That left us with age and gender as independent variables to control the SNPs.

The chi-square test of independence of SNPs itself are not statistically significant (p -value more than 0.05). But because this is our independent variable of interest, and the limitation of chi-square test of independence which did not consider confounding can result in loss of statistical power and bias, it is not wise to stop the analysis of SNPs here.

After conducting ordinal logistic regression of SNPs towards phase controlled by age and gender, we have outputs:

Model Fitting Information

Model	-2 Log Likelihood	Chi-Square	df	Sig.
Intercept Only	524,471			
Final	496,056	28,415	4	,000

Link function: Logit.

Before we start looking at the effect of each independent variables in the model, we need to determine whether the model improves the ability to predict the outcome. The statistically significant chi-square statistic indicates that the final model gives a significant improvement in predicting the outcome, compared to just guessed based on marginal probability for the outcome categories.

Goodness-of-Fit

	Chi-Square	df	Sig.
Pearson	449,839	431	,256
Deviance	408,578	431	,775

Link function: Logit.

Pseudo R-Square

Cox and Snell	,108
Nagelkerke	,116
McFadden	,044

Link function: Logit.

These Goodness of Fit statistics are intended to test whether the observed data are consistent with the fitted model. If we do not reject the null hypothesis (p-value more than 0.05), we can conclude that the data and the model predictions are similar, and we have a good model. High pseudo R-square values indicates that the model with age, gender and SNPs is likely to be a good predictor to the outcome of CHB natural history phase.

Parameter Estimates

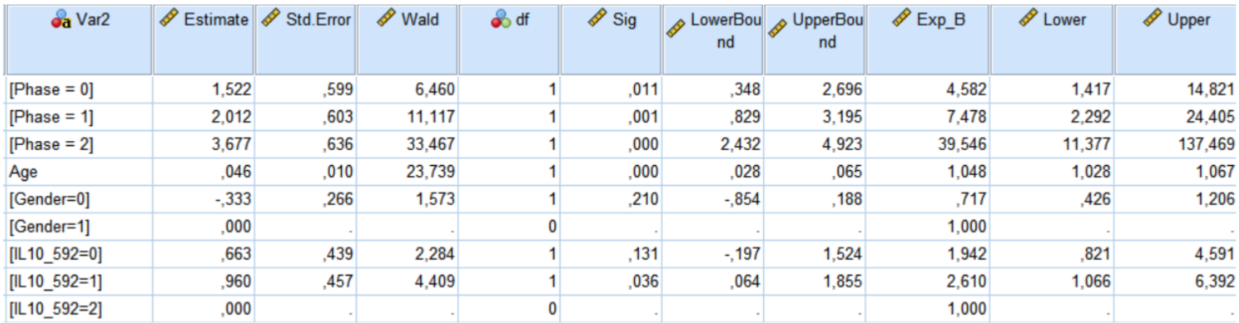
		Estimate	Std. Error	Wald	df	Sig.	95% Confidence Interval	
							Lower Bound	Upper Bound
Threshold	[Phase = 0]	1,522	,599	6,460	1	,011	,348	2,696
	[Phase = 1]	2,012	,603	11,117	1	,001	,829	3,195
	[Phase = 2]	3,677	,636	33,467	1	,000	2,432	4,923
Location	Age	,046	,010	23,739	1	,000	,028	,065
	[Gender=0]	-,333	,266	1,573	1	,210	-,854	,188
	[Gender=1]	0^a	.	.	0	.	.	.
	[IL10_592=0]	,663	,439	2,284	1	,131	-,197	1,524
	[IL10_592=1]	,960	,457	4,409	1	,036	,064	1,855
	[IL10_592=2]	0^a	.	.	0	.	.	.

After getting those good result in validating our data and prediction model, now it is safe to look at the adjusted odds ratio produced by SPSS.

At parameter estimates, there are things to be looked at:

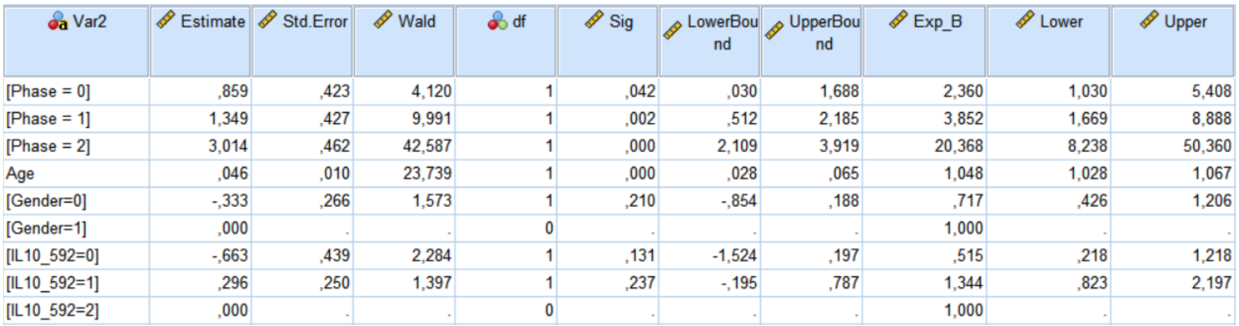
1. The threshold coefficients are representing the intercepts between phase, so the classification and the prediction towards phase are more objective.
2. Estimates of location are the probability of having a level up of one phase to next phase. For age, is in increasing 1 value of age. For Gender or IL10_592, when compared to reference (with 0 estimate).
3. Sig is the p-value, processed from Wald value.

After that, we calculate the odds ratio (OR) (Exp_B) with 95% confidence interval (CI) (Lower and Upper), and we got these results:



Var2	Estimate	Std.Error	Wald	df	Sig	LowerBound	UpperBound	Exp_B	Lower	Upper
[Phase = 0]	1,522	,599	6,460	1	,011	,348	2,696	4,582	1,417	14,821
[Phase = 1]	2,012	,603	11,117	1	,001	,829	3,195	7,478	2,292	24,405
[Phase = 2]	3,677	,636	33,467	1	,000	2,432	4,923	39,546	11,377	137,469
Age	,046	,010	23,739	1	,000	,028	,065	1,048	1,028	1,067
[Gender=0]	-,333	,266	1,573	1	,210	-,854	,188	,717	,426	1,206
[Gender=1]	,000	.	.	0	.	.	.	1,000	.	.
[IL10_592=0]	,663	,439	2,284	1	,131	-,197	1,524	1,942	,821	4,591
[IL10_592=1]	,960	,457	4,409	1	,036	,064	1,855	2,610	1,066	6,392
[IL10_592=2]	,000	.	.	0	.	.	.	1,000	.	.

This one is the result using mutant genotype (CC) as reference. For the one with normal genotype (AA) as reference:



Var2	Estimate	Std.Error	Wald	df	Sig	LowerBound	UpperBound	Exp_B	Lower	Upper
[Phase = 0]	,859	,423	4,120	1	,042	,030	1,688	2,360	1,030	5,408
[Phase = 1]	1,349	,427	9,991	1	,002	,512	2,185	3,852	1,669	8,888
[Phase = 2]	3,014	,462	42,587	1	,000	2,109	3,919	20,368	8,238	50,360
Age	,046	,010	23,739	1	,000	,028	,065	1,048	1,028	1,067
[Gender=0]	-,333	,266	1,573	1	,210	-,854	,188	,717	,426	1,206
[Gender=1]	,000	.	.	0	.	.	.	1,000	.	.
[IL10_592=0]	-,663	,439	2,284	1	,131	-1,524	,197	,515	,218	1,218
[IL10_592=1]	,296	,250	1,397	1	,237	-,195	,787	1,344	,823	2,197
[IL10_592=2]	,000	.	.	0	.	.	.	1,000	.	.

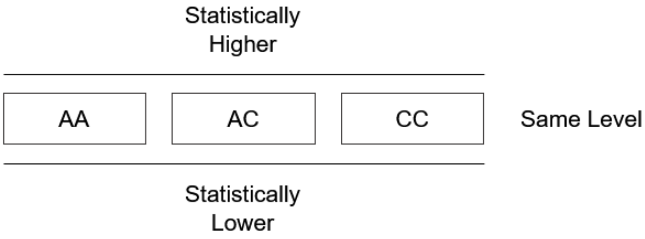
From these result, there are two conclusion:

1. Heterozygous (AC) and mutant (CC) genotype influence on the outcome of CHB natural history phase is not statistically different than normal (AA) genotype.
2. But there is a statistically significant difference between AC and CC. Heterozygous genotype has 2.61 times more to develop advanced phase compared to mutant genotype.

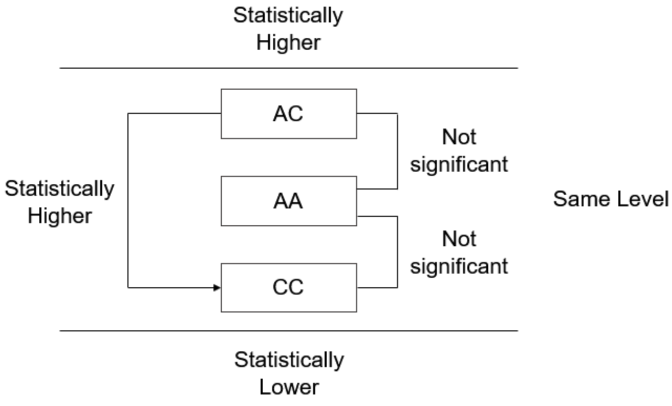
How can we justify using CC as another reference? Why not only AA?

It is because if we only using AA as reference, we can't really tell which one is higher/lower than the other. But, because there is a difference between AC than CC, we can make a rank of it: AC with the most probability to develop advanced phase (ENH especially), then normal with moderate probability, and mutant with lower probability.

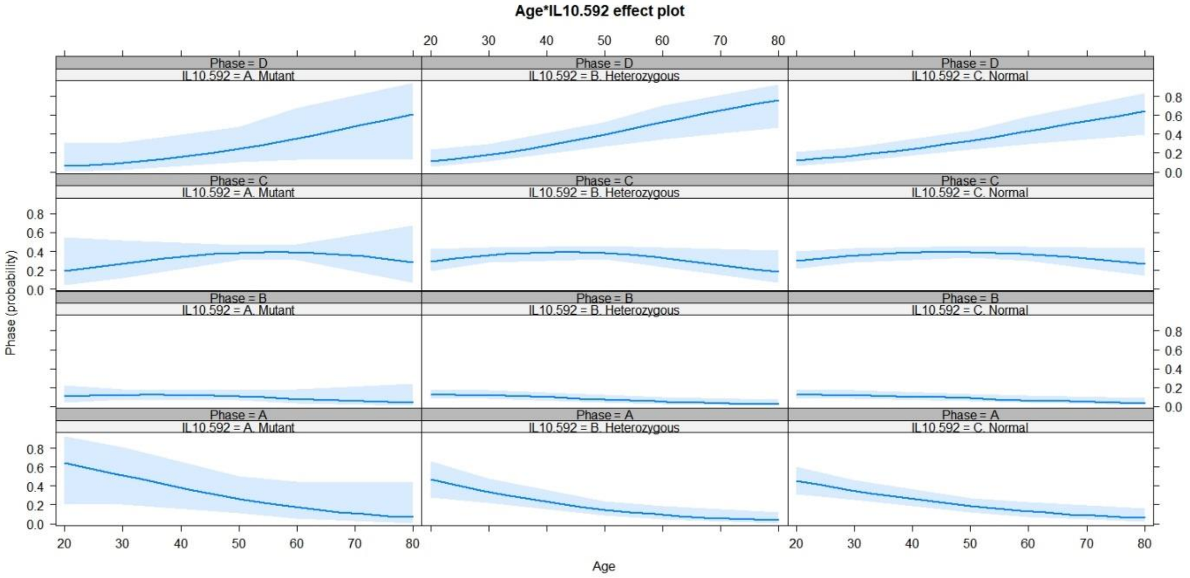
With these graphs, it will be easier to describe.
When there is no significant difference between genotype, we can put them in the same level.



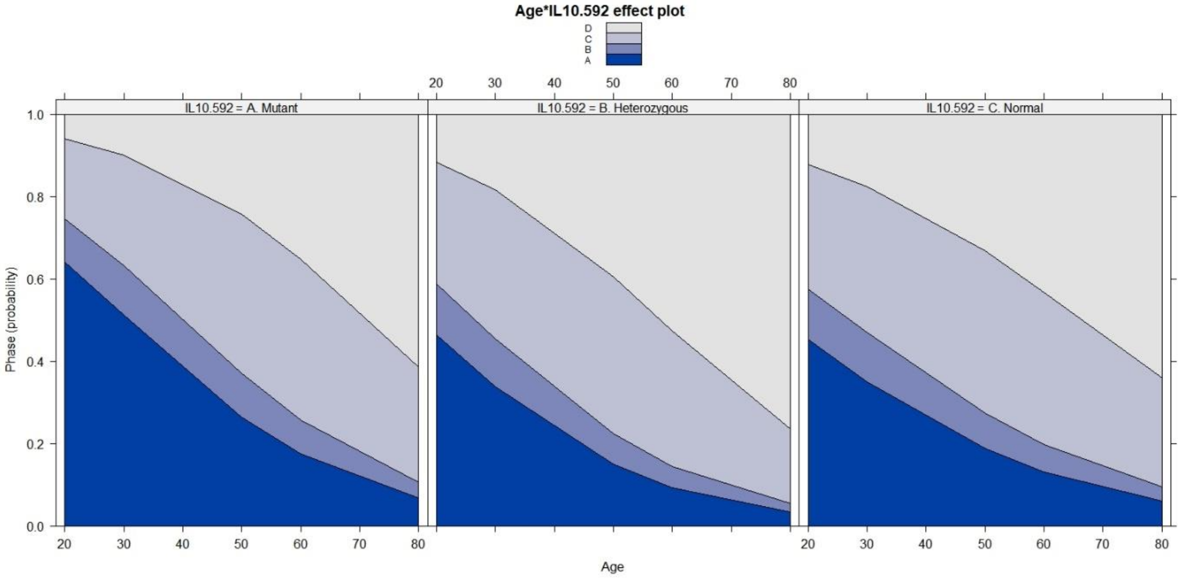
With only this graph, we can conclude that SNPs can’t be used as predictor towards the progression of CHB natural history. But the order, which one is higher or lower is not clear enough. So, after using mutant genotype (CC) as reference, the order is clearer to interpret, and the analysis become more comprehensive.



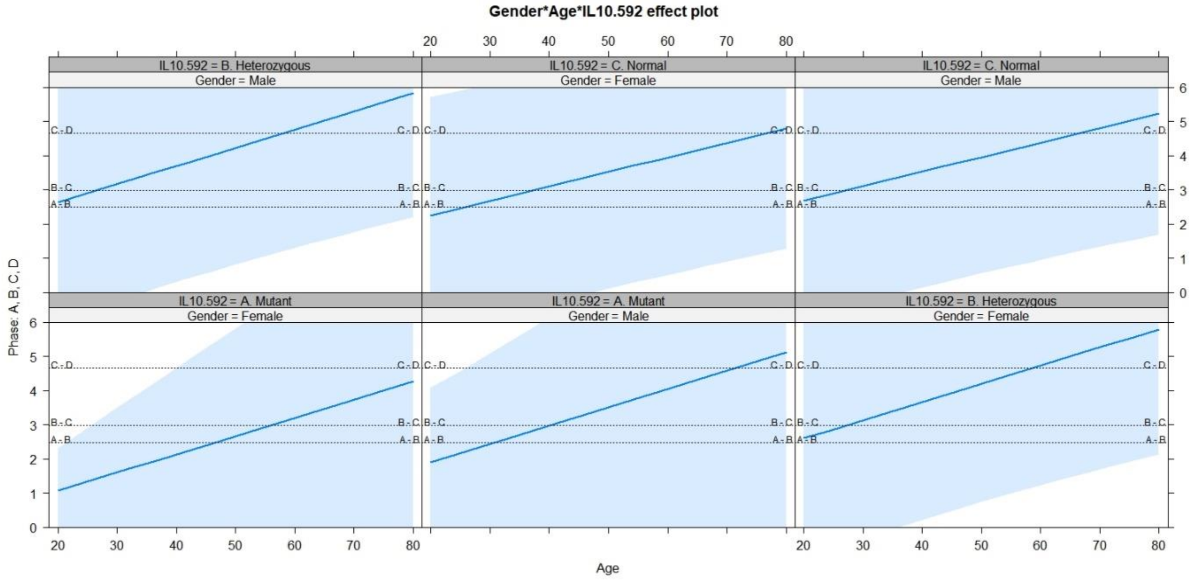
Other than these logically-made graphs, there are also graphs from the model we made. It tells us more about the relationship of independent variables towards the dependent variable.



This graph is quite easy to describe. No need for me to explain. This one is modelled only in male.



For this graph, we can interpret the result by telling the amount of surface area each phase have. In example: an increase in age will result in decrease of IT phase and increase of ENH phase. Heterozygous genotype start and end with bigger surface area compared to mutant and normal.



This graph is almost the same with the first one, but the intercept between each phase is clearer. In example: A-B means the border line that tells the transition of IT to IC phase. Other than that: in female is less likely to develop ENH compared to male with mutant (CC) genotype. And heterozygous is likely to develop ENH in both female and male, and compared to another genotypes.

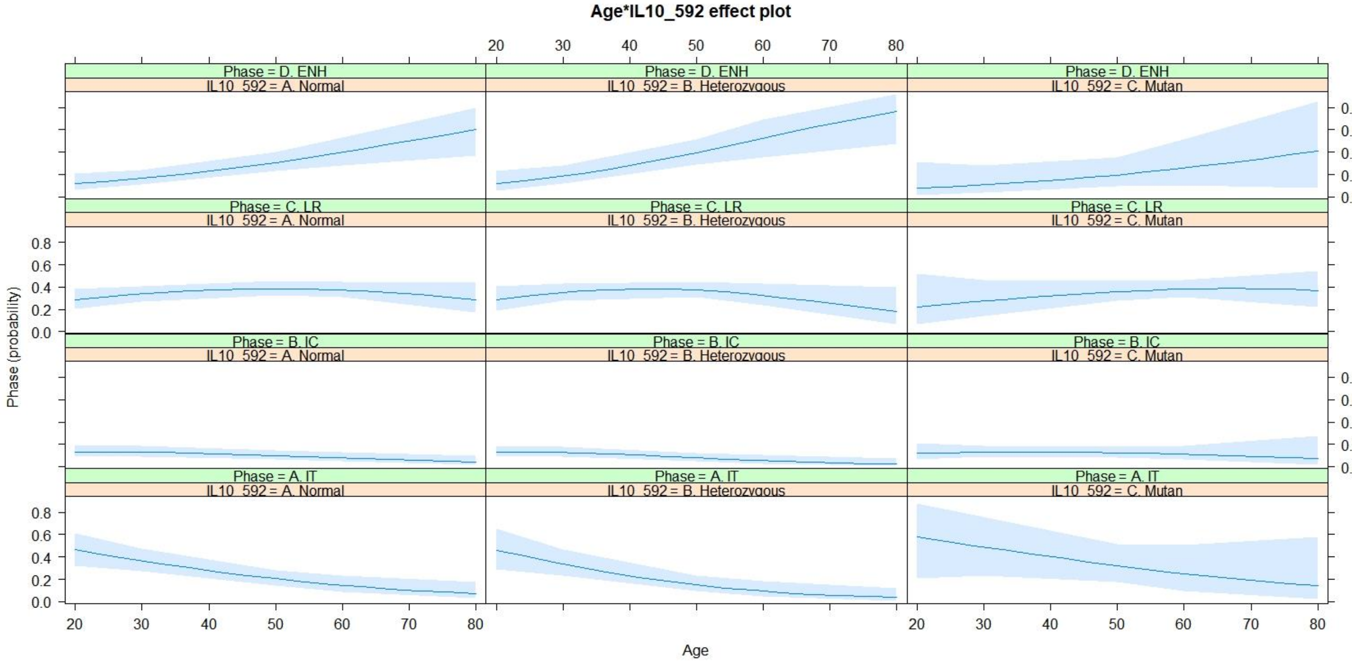
References for this data output interpretation:

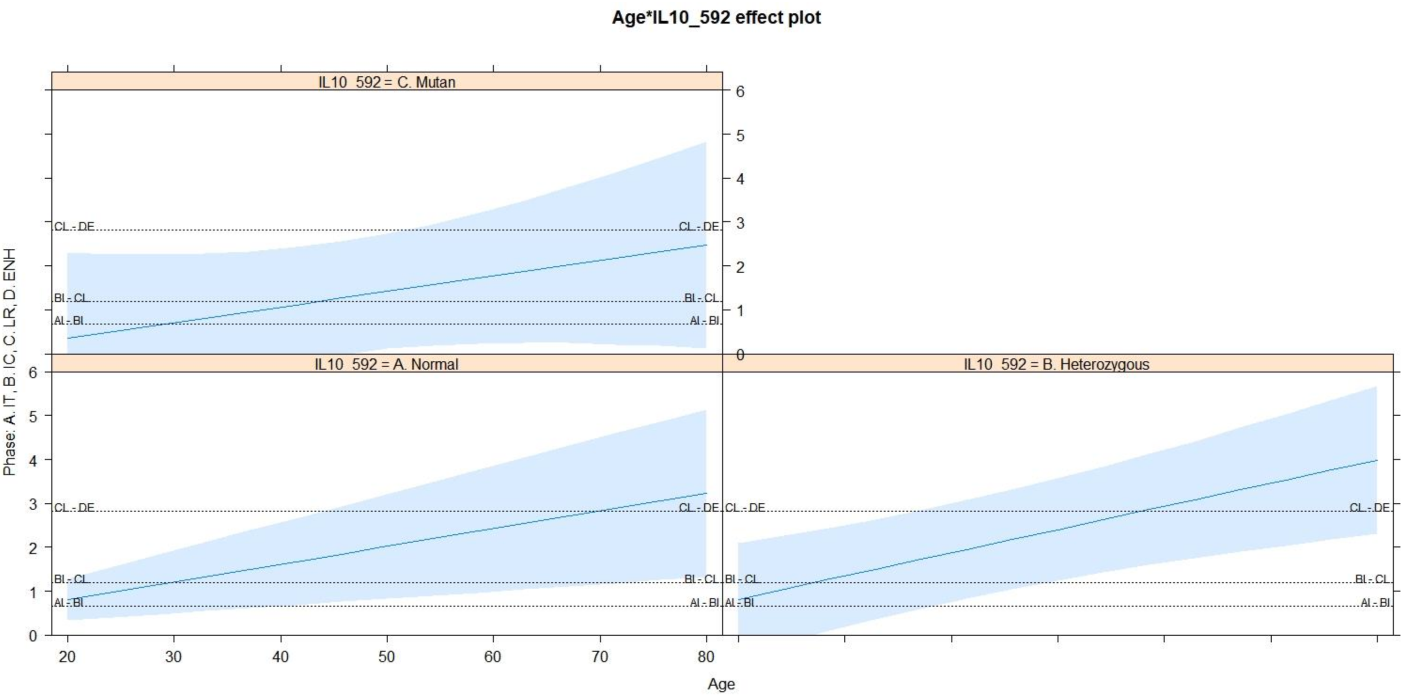
http://www.restore.ac.uk/srme/www/fac/soc/wie/research-new/srme/modules/mod5/4/index.html
https://stats.idre.ucla.edu/spss/output/ordered-logistic-regression/
https://data.library.virginia.edu/fitting-and-interpreting-a-proportional-odds-model/
https://data.library.virginia.edu/visualizing-the-effects-of-proportional-odds-logistic-regression/

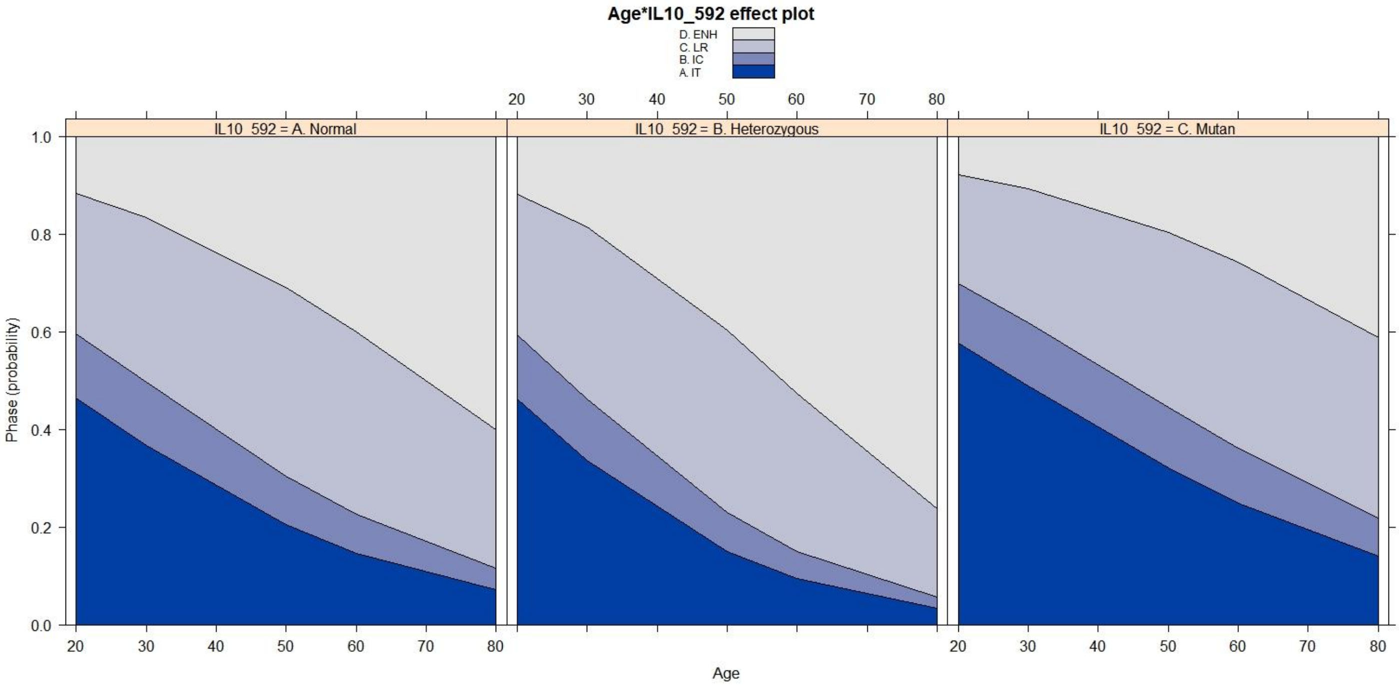
Notes:

There are some phase classification error in 5 data, resulting in shifted distribution of the phase. Please ignore the result above, just focus in how to explain the data. To see the result, please use R with file in *Revision Final 23/12/2019* folder.

Other than that, gender is excluded from covariates, because not significant chi square test of independence p-value.







These are the right pictures.

The significant result: the odds of genotype AC to develop advanced phase was 2.385 (95% CI: 1.02 – 5.66, Wald x^2 = 3.991, p-value = 0.046) times higher compared to genotype CC as reference.
No statistically significant difference of odds ratio between AC or CC compared to AA as reference (p-value > 0.05).