BMC Proceedings

PROCEEDINGS Open Access

Analysis of the progression of systolic blood pressure using imputation of missing phenotype values

Tatsiana Vaitsiakhovich[1*†], Dmitriy Drichel[2†], Marina Angisch[1], Tim Becker[2,1], Christine Herold[2], André Lacour[2*]

From Genetic Analysis Workshop 18
Stevenson, WA, USA. 13-17 October 2012

Abstract

We present a genome-wide association study of a quantitative trait, “progression of systolic blood pressure in time,” in which 142 unrelated individuals of the Genetic Analysis Workshop 18 real genotype data were analyzed. Information on systolic blood pressure and other phenotypic covariates was missing at certain time points for a considerable part of the sample. We observed that the dropout process causing missingness is not independent of the initial systolic blood pressure; that is, the data is not missing completely at random. However, after the adjustment for age, the impact of systolic blood pressure on dropouts was no longer significant. Therefore, we decided to impute missing phenotype values by using information from individuals with complete phenotypic data. Progression of systolic blood pressure ($\Delta SBP/\Delta t$) was defined based on the imputed phenotypes and analyzed in a genome-wide fashion. We also conducted an exhaustive genome-wide search for interaction between single-nucleotide polymorphisms (7.14×10^{10} tests) under an allelic model.

The suggested data imputation and the association analysis strategy proved to be valid in the sense that there was no evidence of genome-wide inflation or increased type I error in general. Furthermore, we detected 2 single-nucleotide polymorphisms (SNPs) that met the criterion for genome-wide significance ($p \le 5 \times 10^{-8}$), which was also confirmed via Monte Carlo simulation. In view of the rather small sample size, however, the results have to be followed-up in larger studies.

Background

We present a genome-wide analysis, that is, an analysis comprising all odd-numbered chromosomes, of the Genetic Analysis Workshop 18 (GAW18) real data. The structure of the GAW18 data is rather complex (longitudinal study, family structure, missing measurements), making it difficult to address all dimensions of the data in a single analysis. Therefore, we restricted our analysis to 142 unrelated individuals available. Moreover, because systolic blood pressure (SBP) was analyzed extensively during the genome-wide association analysis era by large consortia, we decided to investigate the progression of SBP instead.

We address the problem of missing phenotypes and impute missing values for the original sample of 939 individuals. Based on the imputed phenotype data, we conduct a genome-wide analysis of the quantitative trait progression of SBP in time and perform an exhaustive search for SNP-SNP interaction. We assess the validity of the approach in the genome-wide setting and confirm our top-ranking findings via Monte Carlo simulation. Finally, we compare results obtained with the imputed phenotypes to the respective results obtained with the original phenotype data.

Methods

Imputation of missing data in longitudinal studies

In a longitudinal study repeated measurements and records related to a trait under investigation are collected

* Correspondence: vait@imbie.meb.uni-bonn.de; andre.lacour@dzne.de
† Contributed equally
[1]Institute for Medical Biometry, Informatics and Epidemiology (IMBIE), University of Bonn, Sigmund-Freud-Str., D-53105 Bonn, Germany
[2]German Center for Neurodegenerative Diseases (DZNE), Ludwig-Erhard-Allee 2, D-53175 Bonn, Germany
Full list of author information is available at the end of the article

at different time points for a fixed group of individuals. It is often not possible to collect complete data sets because of early dropouts, changed health conditions of subjects, or other reasons. The first step in the analysis of repeated measurements with missing values should be the verification, whether the data is missing completely at random (MCAR), missing at random (MAR), or missing not at random (MNAR) [1]. An observation is said to be MCAR if the missingness is independent of all observed and unobserved assessments. A more relaxed assumption about the missing data mechanism is MAR, where missingness is independent of all unobserved values, although it may be dependent on the observed values. A process that is neither MCAR nor MAR is called MNAR [1-3]. Monotone missing data refers to data from a longitudinal study, where the subjects who left the study before it was finished never returned again.

Standard methods to treat missing data, for example, complete case analysis and simple imputation methods (last-observed-carried-forward, mean, etc), are not appropriate for a valid inference without assumption that the data is MCAR, or when this assumption is violated. A proper analysis of missing data requires a knowledge or estimate of the mechanism that generated the missing data [4]. A test for MCAR versus MAR can be found in Little [5]. Unfortunately, one cannot determine whether missingness is MNAR or MAR based solely on the given data [4].

In analyzing the GAW18 data we were dealing with quantitative, non-monotone data from the longitudinal study. For available phenotypic variables we introduced the following notations: AGE_i for age, SBP_i for SBP, MED_i for intake of antihypertensive medication, and SMK_i for smoking status at the initial measurement i = 0 and up to 3 follow-ups (i = 1,2,3).

We analyzed the quantitative trait progression of SBP in time, $\Delta SBP/\Delta t$, which was defined as the difference between 2 values of SBP recorded at the last and the first examinations attended by an individual, divided by the time Δt in years elapsed between these 2 examinations.

We suspected that the data was not MCAR and tried to determine whether MAR or MNAR was more proper to assume. We considered a linear regression model, $SBP_0 = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$, where a variable x_i equals 1 if the measure of SBP_i is given, and 0 if the measure of SBP_i is missing, $\beta_i \in \mathbb{R}$, i = 0,1,2,3. Using this model to test the hypothesis $\beta_1 = \beta_2 = \beta_3 = 0$ resulted in the p-value 5.21×10^{-7}. This led us to conclude that the data is not MCAR, given that the presence or absence of the follow-up measures is influenced by SBP_0. However, when we included the covariate AGE_0, the p-value became much higher (p = 0.07), which suggested that age might be a significant predictor of missingness and that the data is likely MAR.

To impute the missing values and create complete phenotype data for each subject, we applied a kind of simple imputation method [2,3], which represents our modification of this well-known technique. For our purposes, we imputed the missing values in AGE_i, SBP_i, MED_i, SMK_i variables, i = 0,1,2,3, without considering hypertension status or diastolic blood pressure. To start the imputation we extracted 187 completers, that is, individuals who had participated in all 4 examinations and for whom all measurements had been recorded. We used the information collected for completers to determine the settings of the association analysis, as well as to complete the missing phenotype values for non-completers.

We suggest the following algorithm to impute the missing values in the GAW18 real phenotype data for the purposes of the analysis of $\Delta SBP/\Delta t$:

1. Determine the average calendar years of the 4 examinations (here: 1993, 1998, 2003, 2009).

2. Impute the missing values of AGE_i , i = 0,1,2,3, using the results of Step 1 and the birth year of every individual.

3. Calculate the correlation coefficients α_1 , α_2 , α_3 of ΔSBP on AGE_0, MED_0, SMK_0, respectively, and β of ΔSBP on SBP_0 for completers.

4. Given a completer C and a non-completer N, define a function where wide hat denotes the normalized value of the corresponding variable for instance, and where $I_{(\cdot)}$ equals 1 if the corresponding measurement for a non-completer is given, or equals 0 otherwise.

$$d(C,N) := \sum_{i=0}^{3} [\left|\frac{\alpha_1}{\beta}\right| I_{AGE_i^N}(\widehat{AGE}_i^N - \widehat{AGE}_i^C)^2 + I_{SBP_i^N}(\widehat{SBP}_i^N - \widehat{SBP}_i^C)^2 + \left|\frac{\alpha_2}{\beta}\right| I_{MED_i^N}(\widehat{MED}_i^N - \widehat{MED}_i^C)^2 + \left|\frac{\alpha_3}{\beta}\right| I_{SMK_i^N}(\widehat{SMK}_i^N - \widehat{SMK}_i^C)^2] \quad (1)$$

$$\widehat{AGE}_i^N = \frac{AGE_i^N - \mu_i^N}{\sigma_i^N}, \mu_i^N = \sum_N \frac{AGE_i^N}{N}, \sigma_i^N = \sqrt{\sum_N (AGE_i^N)^2 - (\mu_i^N)^2} \quad (2)$$

5. For every fixed non-completer determine the "closest" completer by finding the minimal value of the function d on the list of all completers.

6. Impute the missing values of SBP_i, MED_i , SMK_i , i = 0,1,2,3, for a given non-completer by those from the "closest" completer.

Imputation was performed for 752 non-completers. Table 1 provides an example of imputation.

Quality control (GAW18 real genotype data)

First, we filtered 142 individuals from the GAW18 data set who were indicated as being nonrelated. To check for unobserved relatedness among these individuals, we computed a pair-wise identity-by-state (IBS) matrix from the genotype data. For a pair of individuals, their IBS-value was computed as the portion of alleles identical by

Table 1 Imputation of the missing phenotype data

ID	*N/C*	*d*(*C,N*)	AGE_0	AGE_1	AGE_2	AGE_3	SBP_0	SBP_1	SBP_2	SBP_3	MED_0	MED_1	MED_2	MED_3	SMK_0	SMK_1	SMK_2	SMK_3
T2DG2501049	N		**39**	45	50	**56**	**110**	125	139	**149**	**0**	0	0	**1**	**0**	0	0	**0**
T2DG0300150	C	0.072	35	41	45	51	110	121	135	149	0	0	0	1	0	0	0	0
T2DG1700856	C	0.108	33	38	43	50	110	123	130	132	0	0	0	1	0	0	0	0
T2DG0600470	C	0.268	43	46	51	58	133	127	131	185	0	0	0	0	1	1	0	1
...				...				...				...				...		
T2DG0300165	C	18.105	50	57	61	66	151	175	176	139	0	0	0	1	0	0	0	0
T2DG0600450	C	21.673	48	53	61	65	136	190	128	130	0	1	1	1	1	0	0	0

The results of imputation are shown for a middle-aged non-completer (*N*) with ID number T2DG2501049 (see GAW18 real phenotype data) with 2 missing values in each phenotypic variable. The real SBP measurements indicate the possible development of hypertension by this individual. In the table, several values of the distance-function *d(C,N)* are shown in ascending order. The completer (*C*) T2DG0300150 is closest to the given non-completer. The imputed values are shown in bold.

state divided by the number of alleles genotyped in both individuals. The mean IBS-value over all pairs of individuals was 0.74; its SD (standard deviation) was 0.0063. Seven pairs had IBS-values higher than 0.84, which was more than 14 SD above the mean, indicating evidence of relatedness. For each of the 7 pairs, the individual with the lower genotyping rate was removed from the analysis. We also excluded 8 other individuals who had genotype missing rates exceeding 40%, leaving 127 individuals for analysis.

Finally, we removed 42,519 SNPs because of either low genotyping rate (<99%), deviations from Hardy-Weinberg equilibrium ($p<1 \times 10^{-6}$), or low minor allele frequency (MAF<1%). That left 429,516 quality controlled SNPs for analysis, and the overall genotyping rate of the remaining data was 99.9%.

Statistical analysis

For the association analysis of $\Delta SBP/\Delta t$ reasonable covariates would be AGE_0*, SBP_0*, MED_0*, SMK_0*, ΔMED, ΔSMK, and *SEX*. Here, for every individual, the variables AGE_0*, SBP_0*, MED_0*, SMK_0*are defined to be equal to the first non-missing AGE_i , SBP_i, MED_i , SMK_i values, respectively; i = 0,1,2,3; ΔMED and ΔSMK are defined analogously to ΔSBP. Note that when using imputed phenotype data, AGE_0*= AGE_0 , SBP_0*=SBP_0, etc. From the plausible covariates listed above, we retained for the association analysis only those covariates that were significantly associated with $\Delta SBP/\Delta t$ in the sample of completers: AGE_0*, SBP_0*, MED_0*, and *SEX*.

For single-marker analysis, we applied the standard linear regression test with 1 degree of freedom (DF) for quantitative traits and used the implementation in INTERSNP [6].

For 2-marker analysis, we used a test for allelic interaction. We compared the sum of standard errors (SSE) of the model $M^{A,I}= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{1,2} x_1 x_2$ against the model $M^{A} = \beta_0 + \beta_1 x_1 + \beta_2 x_2$, where β_0 is the intercept parameter x_i is the number of copies of minor alleles of SNP i, i = 1,2; for a given individual, $x_1 x_2$ is the interaction term; and βs are the coefficients to be estimated. Let $SSE^{A,I}$ and SSE^{A} be SSEs of the respective models, and let n be the number of individuals in the analysis. Then, the test statistic $t = (SSE^{A} - SSE^{A,I})/SSE^{A}$ is F-distributed with (n − 4, 1) DF and yields a test for allelic interaction. Analysis was performed by a parallelized version of INTERSNP, which uses the OpenMP API specification for parallel programming (http://openmp.org/wp). In the interaction analysis, we applied a stronger MAF criterion of 5% and analyzed 7.14×10^{10} SNP pairs.

When the sample size is small, low MAF can lead to a small number of "effective" analysis observations and the application of the F-distribution might lead to inflated p-values. To overcome this issue, we validated the top-ranking p-values of the single-marker analysis via Monte-Carlo simulation. In each permutation replicate, the quantitative trait values were randomly distributed over the individuals of the sample. In total, 10^8 permutation replicates were conducted and for a given SNP, its point-wise Monte-Carlo p -value was computed as $k/10^8$, where k is the number of permutation replicates with a test statistic t larger than that of the real data. For computation time reasons, it was not possible to validate the results of the interaction analysis via simulation, since at least 10^{13} replicates would have been needed.

Results

The analysis of $\Delta SBP/\Delta t$ yielded a genome-wide inflation factor of λ = 0.979, indicating that, overall, our analysis can be judged as valid in the sense that we do not have evidence for inflated p-values caused either by population stratification or by our phenotype imputation method. Table 2 lists the best results of the single-marker analysis. Two SNPs, rs3093642 ($p = 1.35 \times 10^{-9}$) and rs13157168 ($p = 2.45 \times 10^{-8}$), met the criterion for genome-wide significance. These SNPs have rather low MAF, and the results might be artifacts. However, Monte-Carlo simulation with 10^8 permutation replicates confirmed the findings; none of the replicates had a

Table 2 Single-marker analysis of Δ*SBP*/Δ*t* on unrelated individuals

SNP	Position	Alleles	MAF	HWE	β ± σ	p_i	p_i^{MC}	p_r	p_r^{MC}
rs3093642	1-25687742	T/C	0.02	1.23e-5*	2.68 ± 0.41	1.35e-9	0†	0.57	0.49
rs13157168	5-86627138	A/G	0.01	0.893	3.83 ± 0.64	2.45e-8	0†	0.48	0.41
rs956918	9-436983	C/A	0.05	0.002	1.54 ± 0.29	4.14e-7	4.00e-7	0.04	0.05
rs6960510	7-147977001	G/A	0.04	0.641	1.91 ± 0.37	1.32e-6	1.60e-6	0.27	0.26
rs7120076	11-111404684	G/A	0.02	1.23e-5*	2.17 ± 0.43	1.38e-6	1.00e-7	0.46	0.39
rs11803060	1-76549233	C/T	0.04	0.644	1.82 ± 0.37	3.14e-6	3.10e-6	0.40	0.39
rs11665668	19-52169702	A/G	0.04	0.029	1.69 ± 0.35	3.47e-6	3.10e-6	0.09	0.08
rs1340503	9-78571504	C/T	0.20	0.471	0.90 ± 0.19	3.78e-6	4.10e-6	0.09	0.09
rs6860971	5-88552414	T/C	0.04	0.103	1.52 ± 0.32	7.18e-6	8.90e-6	0.36	0.34
rs1202389	7-148913645	T/G	0.06	0.449	1.43 ± 0.31	7.46e-6	9.80e-6	0.49	0.48
rs10438410	15-92055287	T/C	0.03	0.749	2.07 ± 0.44	7.61e-6	9.30e-6	0.44	0.40
rs1524846	15-24025870	G/T	0.02	0.785	2.25 ± 0.48	8.18e-6	8.80e-6	0.01	0.02
rs12269879	11-108413351	A/G	0.02	0.857	2.67 ± 0.58	8.91e-6	9.70e-6	0.25	0.22

*Indicates *p*-values for Hardy-Weinberg equilibrium close to the exclusion criterion.

†None of 10^8 permutation replicates had a more extreme test statistic than the real data.

The best results of the single-marker analysis are shown with SNP ID; physical base-pair position; minor/major alleles, MAF; Hardy-Weinberg equilibrium (HWE) *p*-value; regression coefficient β with standard deviation σ; *p*-value p_i of the single-marker analysis performed on the imputed phenotypes; *p*-value of the Monte-Carlo simulation p_i^{MC} corresponding to p_i; *p*-value p_r of the single-marker analysis performed on SNPs listed in the table and on real phenotypes; and *p*-value of the Monte-Carlo simulation p_r^{MC} corresponding to p_r.

more extreme test statistic than the real data. Nevertheless, the top SNP rs3093642 should be treated with caution because its *p*-value for deviation from Hardy-Weinberg equilibrium was rather close to the predefined exclusion criterion. The second SNP rs13157168 is an intronic variant in the *RASA1* gene. Association of *RASA1* with variable capillary and arteriovenous malformations was reported by Boon et al [7]. Note that the *p*-values p_r of the single-marker analysis performed on SNPs on the *real phenotypes* are markedly higher than the corresponding *p*-values p_i for the imputed phenotypes. SNPs with an at least suggestive level of evidence for association, unfortunately, were not found among the top ranks.

Table 3 presents the top results of the interaction analysis. 2 SNPs from the table, rs4686358 and rs7987982, were previously reported in connection with potentially related phenotypes. The variant rs4686358 is located close to rs35964523 (≈130 kilobases [kb]), a SNP that was implicated in total cholesterol levels ($p = 7.9 \times 10^{-6}$) [8]. SNP rs7987982 is an intronic variant of the *COL4A1* gene. Variants in *COL4A1* are significantly associated

Table 3 Two-marker interaction analysis of Δ*SBP*/Δt on unrelated individuals

SNP$_1$						SNP$_2$								
rs	Position	Alleles	MAF	HWE	p_i	rs	Position	Alleles	MAF	HWE	p_i	β ± σ	p_i^{INT}	p_r^{INT}
rs4686358	3-1028349	T/C	0.15	0.419	0.52	rs17236800	7-37944375	G/A	0.09	0.053	0.19	3.27 ± 0.39	1.20e-13	0.157
rs1321632	3-197092657	T/C	0.06	0.456	0.01	rs7987982	13-110931311	C/T	0.09	0.263	0.22	1.73 ± 0.21	1.24e-13	0.395
rs12632248	3-185139913	A/G	0.10	0.518	0.13	rs10770039	11-9587187	T/C	0.28	0.017	0.56	2.97 ± 0.36	2.78e-13	0.503
rs1482590	3-65219095	T/C	0.20	0.361	0.09	rs10488631	7-128594183	C/T	0.23	0.754	0.22	1.72 ± 0.21	6.82e-13	0.598
rs11716921	3-17175592	G/T	0.17	0.104	0.11	rs6870951	5-133064960	T/C	0.06	0.511	0.43	4.06 ± 0.51	1.45e-12	0.086
rs2220368	5-176140452	A/G	0.21	0.695	0.92	rs7810996	7-73624615	T/C	0.31	0.511	0.78	2.85 ± 0.37	3.48e-12	0.081
rs249209	5-79867209	C/A	0.25	0.069	0.34	rs427488	17-12428237	C/T	0.11	0.163	0.18	2.51 ± 0.32	3.62e-12	0.416
rs7719329	5-108391660	T/C	0.09	0.053	0.13	rs4868699	5-176134402	C/T	0.22	0.222	0.02	2.04 ± 0.27	4.45e-12	0.013
rs524138	5-31465137	A/G	0.07	0.417	0.06	rs4885678	13-80440491	T/G	0.22	0.981	0.50	3.41 ± 0.45	5.90e-12	0.048
rs17242130	5-81154906	C/A	0.06	0.517	0.32	rs2836412	21-39825012	C/T	0.25	0.617	0.08	3.18 ± 0.42	6.38e-12	0.665
rs1677694	5-79936297	T/C	0.17	7.19e-4*	0.23	rs12673145	7-81423463	G/T	0.11	0.679	0.16	2.91 ± 0.38	7.07e-12	0.027
rs3757572	7-45146656	C/A	0.27	0.552	0.55	rs7275197	21-39317417	G/A	0.14	0.235	0.25	3.70 ± 0.49	8.58e-12	0.137
rs1802074	7-37947103	A/G	0.13	0.579	0.50	rs3887013	15-93614127	T/C	0.09	0.240	0.57	3.92 ± 0.52	9.41e-12	0.043

* Indicates *p*-values for Hardy-Weinberg equilibrium close to the exclusion criterion.

The best results of the 2-marker interaction analysis are shown. Each line contains a pair of SNPs with their IDs; physical base positions; minor/major alleles; Hardy-Weinberg equilibriums (HWEs); *p*-values p_i of the single-marker analysis performed on the imputed phenotypes; regression coefficient β with standard deviation σ of the interaction analysis; *p*-value p_i^{INT} of the interaction analysis performed on the imputed phenotypes; and *p*-value p_r^{INT} of the interaction analysis performed on SNPs listed in the table and on the real phenotypes.

with brain small vessel disease, in which stiffness of blood vessels is affected [9].

Discussion

In longitudinal studies, repeated measurements related to a trait under investigation are performed subsequently within a particular time period. The attractiveness of longitudinal studies is the ability to capture the dynamics of traits. Investigating the progression of disease by using this additional dimension of time might reveal insights on underlying disease mechanisms. Given the nature and the running time of a longitudinal study, it is expected that some measurements might not be available, whereas others might be missed. Imputation of missing values helps to enrich data and to use maximum information, although it has own weaknesses and it is not self-evident that it leads to improved power.

In our setting, the imputation of missing values in the GAW18 phenotype data set helped to streamline the definition of the trait under investigation ($\Delta SBP/\Delta t$). In particular, the imputation reduced the impact of outlier SBP values, and made it possible to define the values of the trait in a more uniform time period. This, in turn, helped to reduce the potentially negative impact of short time periods on the values of the trait.

The method of choice to check the validity of imputation is to carry out a sensitivity analysis to investigate the dependence of the results on the particular imputation method. In our case, the sensitivity analysis could have been carried out by the multiple imputation technique. Unfortunately, this technique is time-consuming and requires software implementation; consequently, it was not applied in our work. This is a potential limitation of our approach. An argument in favor of our imputation method is that the mean of SBP values within an examination is preserved by the imputation.

Conclusions

Our analysis strategy helped to identify some suggestive association findings. At this point, however, the results do not prove that the analysis using imputed phenotype data is superior to that without imputation. To prove this statement, it would be necessary to know whether some of the SNPs we have identified represent "true" associations. Those have to be investigated in studies with independent data. In summary, our work can be seen as a suggestion for longitudinal data analysis, which was quite reasonable in the sense that it did not show genome-wide inflation or increased type I error in general.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

TV designed and performed the imputation of the missing phenotypes. TV and DD carried out the analysis of unrelated individuals. MA worked on the pre-processing of the data. CH provided support in using INTERSNP software. TB conceived of the study and suggested main lines of the analysis. AL coordinated the study, participated in its design and improvement. All authors read and approved the final manuscript.

Acknowledgements

T. Vaitsiakhovich is supported by the "Deutsche Forschungsgemeinschaft," project BE3828/3-2. The GAW18 whole genome sequence data were provided by the T2D-GENES Consortium, which is supported by NIH grants U01 DK085524, U01 DK085584, U01 DK085501, U01 DK085526, and U01 DK085545. The other genetic and phenotypic data for GAW18 were provided by the San Antonio Family Heart Study and San Antonio Family Diabetes/Gallbladder Study, which are supported by NIH grants P01 HL045222, R01 DK047482, and R01 DK053889. The Genetic Analysis Workshop is supported by NIH grant R01 GM031575.
This article has been published as part of *BMC Proceedings* Volume 8 Supplement 1, 2014: Genetic Analysis Workshop 18. The full contents of the supplement are available online at http://www.biomedcentral.com/bmcproc/supplements/8/S1. Publication charges for this supplement were funded by the Texas Biomedical Research Institute.

Authors' details

[1]Institute for Medical Biometry, Informatics and Epidemiology (IMBIE), University of Bonn, Sigmund-Freud-Str., D-53105 Bonn, Germany. [2]German Center for Neurodegenerative Diseases (DZNE), Ludwig-Erhard-Allee 2, D-53175 Bonn, Germany.

Published: 17 June 2014

References

1. Rubin DB: **Inference and missing data.** *Biometrika* 1976, **63**:581-592.
2. Fielding S, Fayers PM, McDonald A, McPherson G, Campbell MK: **RECORD Study Group:Simple imputation methods were inadequate for missing not at random (MNAR) quality of life data.** *Health Qual Life Outcomes* 2008, **6**:57.
3. Kenward MG, Lesaffre E, Molenberghs G: **An application of maximum likelihood and generalized estimating equations to the analysis of ordinal data from a longitudinal study with cases missing at random.** *Biometrics* 1994, **50**:945-953.
4. Ibrahim JG, Chu H, Chen MH: **Missing data in clinical studies: issues and methods.** *J ClinOncol* 2012, **30**:3297-3303.
5. Little RJA: **A test of missing completely at random for multivariate data with missing values.** *J Am Stat Assoc* 1988, **83**:1198-1202.
6. Herold C, Steffens M, Brockschmidt FF, Baur MP, Becker T: **INTERSNP: genome-wide interaction analysis guided by a priori information.** *Bioinformatics* 2009, **25**:3275-3281.
7. Boon LM, Mulliken JB, Vikkula M: **RASA1: variable phenotype with capillary and arteriovenous malformations.** *Curr Opin Genet Dev* 2005, **15**:265-269.
8. Barber MJ, Mangravite LM, Hyde CL, Chasman DI, Smith JD, McCarty CA, Li X, Wilke RA, Rieder MJ, Williams PT, *et al*: **Genome-wide association of lipid-lowering response to statins in combined study populations.** *PLoS One* 2010, **5**:e9763.
9. Tarasov KV, Sanna S, Scuteri A, Strait JB, Orrù M, Parsa A, Lin PI, Maschio A, Lai S, Piras MG, *et al*: **COL4A1 is associated with arterial stiffness by genome-wide association scan.** *Circ Cardiovasc Genet* 2009, **2**:151-158.

doi:10.1186/1753-6561-8-S1-S83
Cite this article as: Vaitsiakhovich *et al.*: **Analysis of the progression of systolic blood pressure using imputation of missing phenotype values.** *BMC Proceedings* 2014 **8**(Suppl 1):S83.