 doi:10.20944/preprints202006.0226.v1

Bias, confounding, multiplicity and *researchers' degrees of freedom* combined with segmented publications are not likely to provide accurate estimates of reality: a case study of two candidate-gene association reports in the first episode psychosis patients

Vladimir Trkulja

Department of pharmacology, Zagreb University School of Medicine

Šalata 11, 10000 Zagreb, CROATIA

vtrkulja@mef.hr

phone: +385-98-325-307

 doi:10.20944/preprints202006.0226.v1

Abstract

Estimation of the reality can easily be flawed, hence, in order to result in accurate and useful estimates the process has to be protected from bias and confounding and should follow other methodological milestones inherent to different types of empirical observations. Candidate-gene association studies are a specific form of observations that have been rather extensively applied in psychiatry yielding valuable information on various aspects – when methodologically adequate and used in appropriate settings. However, certain flaws that may occur in such studies might not be bluntly obvious, at least not at first glance, and may pass unnoticed by researchers and reviewers. This case study uses two recent published candidate-gene association reports suggesting involvement of cannabinoid receptor type 1 and of heat shock protein single nucleotide polymorphisms in development of neurocognitive performance and psychopathology in a cohort of adult first episode psychosis patients to point-out the types of flaws inevitably resulting in inaccurate and useless estimates.

 doi:10.20944/preprints202006.0226.v1

Introduction

Genetic/genomic analyses have considerably contributed to understanding of individual susceptibility to certain psychiatric disorders or some of their phenotypic traits, to understanding of pharmacokinetics of psychiatric drugs and individual responses determined by the properties of their target molecules or the networks with which they interfere. Family linkage studies, genome-wide (GWAS) and candidate-gene association studies, each with their inherent reaches and limitations, provide complementary information (Hoehe and Morris-Rosendahl, 2018). The latter are valuable in relatively simple and biologically well-understood systems like drug pharmacokinetics, but have been criticized for their limitations in investigations of disease susceptibility or complex traits (such as cognition or psychopathology) (Border and Keller, 2017). However, their value in such settings has been argued for in cases of investigation of biologically plausible pathways, and assuming that they meet certain methodological standards, including at least the following (Moore, 2017): a) for discovery studies of novel associations, small effects of genetic variants should be assumed, reflecting on the sample size; b) adequate adjustments for confounders (genetic or "common" effects), and c) predefinition of hypotheses and computational methods for their testing, with adequate protection against *false discoveries*. The present case study uses two recent subsequent publications to illustrate common flaws that might not be so obvious at first glance. Two reports [Report 1 (Rojnic Kuzman, et al., 2019a), Report 2 (Bosnjak Kuharic, et al., 2020), previously published protocol (Rojnic Kuzman, et al., 2019b)] pertain to one prospective candidate-gene association study conducted in the same adult first episode psychosis (FEP) patients (N=121) with no history of antipsychotics use, generally free of other psychopathology (cannabis use was not an exclusion criterion), who were assessed on a number of psychopathology instruments and cognitive tests on two occasions.

Case presentation

Report 1 (Rojnic Kuzman, et al., 2019a)

 doi:10.20944/preprints202006.0226.v1

Positive and negative syndrome scale (PANSS), General Assessment of Functioning (GAF) scale and Holmes-Rahe Stress Inventory and 20 cognitive tests were assessed at baseline and 18 months later. All patients were genotyped for a single nucleotide polymorphism (SNP) in the gene encoding the cannabinoid receptor type 1 (*CNR1*) – *CNR1* A>G (rs12720071); 112/121 were genotyped also for *CNR1 C>T* (rs7766029). The primary objective was declared as (quote) "*…association of CNR1 genotypes with changes in neurocognitive test results*". Additional objectives were declared as secondary ("*…association of CNR1 genotypes with change in the perceived levels of stress*") or tertiary ("*…interaction of cannabis use with the association of targeted genotypes and changes in neurocognitive test results*"). Patients were dichotomized as *ever* or *never* cannabis users. The study was reportedly motivated by a GWAS study that suggested involvement of rs7766029 in certain neurocognitive features in patients with schizophrenia, a candidate-gene association study suggesting involvement of rs12720071 in neurocognitive performance in patients with schizophrenia, and several other candidate-gene studies implying associations between several other *CNR1* SNPs and certain brain volumetric indices.

Although the two SNPs were in a linkage disequilibrium (LD), no haplotype or diplotype analysis was performed and the effect of one SNP was analyzed without the adjustment for the other one. *CNR1* gene is highly polymorphic with several tens of two-allele SNPs and a huge number of copy number variations (CNVs); with LD shown for many SNPs, and with shown associations between different SNPs (genotypes, haplotypes, diplotypes) and CNVs with a variety of neuropsychiatric disorders – including association between the two tested SNPs (or inclusive haplotypes) with cannabinoid use/dependence (Ishiguro, et al., 2013). Apparently, a total of 23 different outcome variables were analyzed: (change vs. baseline in) PANSS, GAF, stress inventory and 20 neurocognitive tests. The multivariate models always adjusted for the baseline value of the dependent variable, age, sex, negative symptoms PANSS and cannabinoid use. Results for 5 neurocognitive tests and the stress inventory in respect to rs7766029 are shown, with 4/6 adjusted coefficients associated with a p-value that would be incompatible with the null at the

 doi:10.20944/preprints202006.0226.v1

nominal alpha 0.05, but "flagging" 1/6 (1/4 tests of verbal fluency) for which the p-index remained <0.15 after false discovery rate (FDR) adjustment (Table 2 in Rojnic Kuzman, et al., 2019a). Of notion: a) it was not explained why the two SNPs in LD were evaluated separately; b) treatments delivered between baseline and end-of-study or those in use at the time of re-testing, and a number of other potential confounders were never mentioned nor included in the analysis; c) it remained unclear to which contrasts reported coefficients referred and why "stress" was listed among neurocognitive tests if considered a secondary objective; d) it remained unclear how many p-values were submitted to the FDR adjustment and why; and why FDR-adjusted p-values <0.15 should be perceived as relevant (i.e., likely not associated with falsely rejected nulls). Further data (Table 3 in Rojnic Kuzman, et. al., 2019a) referred to rs12720071 and reported on stress and two cognitive tests – none "flagging" an independent association with the variant allele carriage vs. wild type (wt) subjects. Two further tables (Table 4 and 5) reported on an (quote) "*interaction between rs7766029 and cannabis use*" in respect to the cognitive test result; and to an (quote) "*interaction between rs7766029 and rs12720071 and cannabis use*" in their effects on the stress score changes. Of notion: a) it was not explained how these interactions were tested. The proper way would be to fit a model that includes rs7766029 (or both SNPs in the case of stress), baseline value, cannabis use and interaction terms (SNP*cannabis use; or two interaction terms, in the case of stress). It appears, however, that each outcome was analyzed separately in subsets based on cannabis use (or subsets by-SNP-by-cannabis use in the case of stress); b) adjustments and the multiplicity issue in these analyses were not addressed. These "*significant interactions*" were stressed as a virtually major finding of the study.

Report 2 (Bosnjak Kuharic, et al., 2020)

The report pertains to the same study but this time the primary objective was defined as evaluation of association between an SNP in the gene encoding heat-shock protein (HSP) – *HSPA1B G>A* (rs1061581) – with (quote) "… *changes in psychopathology and neurocognitive test*

 doi:10.20944/preprints202006.0226.v1

results". The evaluation included PANSS and Calgary Depression Scale for Schizophrenia (CDSS), and the same 20 neurocognitive tests as in Report 1. The study was reportedly motivated by previous candidate-gene studies suggesting that SNPs in genes (*HSPA1A, HSPA1B* and *HSPA1L*) encoding for HSP70 subfamily of the HSP protein family might be associated with a risk of schizophrenia, and the fact that the G allele in rs1061581 had been suggested associated with a lower expression of mRNA for HSP70-2 protein.

Change from baseline in PANSS (total, positive, negative), CDSS, and in neurocognitive tests (24 different outcomes) were each analyzed separately, with the baseline score, age and sex as covariates. The analyses were performed twice – with subjects classified by genotypes, and also wt subjects were contrasted to variant carriers. Results for CDSS, PANSS positive and two cognitive tests are shown (Table 3 in Bosnjak Kuharic, et al., 2020), three of which are "flagged" as "significant" ($p<0.15$ after FDR correction) in the wt vs. variant contrasts. Of notion: a) neither cannabis use nor the reported *CNR1* SNP effects, and the particularly stressed SNP*cannabis use interaction were mentioned in the text, nor were they accounted for in data analysis; b) all other points on non-included adjustments and statistical analysis are the same as for Report 1. The authors concluded (quote): "*We … have for the first time found the association of HSPA1B rs1061581 polymorphism and … changes of psychopathology and neurocognitive test results in patients with FEP*".

Case discussion

The two reports (Rojnic Kuzman, et al., 2019a, Bosnjak Kuharic, et al., 2020) are segmented publications from one study, which raises both scientific and ethical issues(Supak Smolcic, 2013), but the focus here is on three other points: *bias/confounding,* a phenomenon named *researchers degrees of freedom* (Gelman and Loken, 2013) and *multiplicity*, as these are the key determinants rendering the reported estimates highly unlikely to be accurate, thus making them meaningless (or misleading). While we have no option but to observe in samples and, hence, inevitably have to deal with error due to a chance (*sampling error*), inaccurate estimates are due

 doi:10.20944/preprints202006.0226.v1

to bias/confounding. To increase the probability of accurate estimates, we need to stick to the methods developed to confront it as the only measure against useless observations (Ioannidis, 2005). Should one try to recreate what happened in this study, the following appears as a likely scenario: patients were assessed on 5 psychopathology instruments, 20 cognitive tests, and on 3 further instruments for which no results were reported (Rojnic Kuzman, et al., 2019b). They were genotyped for the reported SNPs, but likely also for at least 4 other SNPs in *MTHFR* and *NDUFV2* genes (Rojnic Kuzman, et al., 2019b) (never mentioned/reported) and then, a number of tests of all possible SNPs and all possible outcomes (including likely various interactions) followed. The two papers presented "*first discoveries*" of effects of individual candidate genes on complex traits in a setting with vaguely suggested mechanistic plausibility in FEP patients. The sample was actually too small for the purpose with a vague reference to sample size calculation (Rojnic Kuzman, et al., 2019a, Bosnjak Kuharic, et al., 2020): it seems that it was not planned in respect to any specific of the many outcome measures and did not account for the (expected) distribution of patients across the targeted SNPs (i.e., subsets by genotypes across different genes), and particularly not for any relevant interaction, but rather in respect to detection of a relatively large effect (declared as Cohen's *g* of 0.53) in any pairwise comparison.

Bias/confounding. It does not make much sense to test the two *CNR1* SNPs without mutual adjustment (or not to test haplotypes or diplotypes, since in LD) (and not accounting for other confounders). It makes even less sense to test the effect of *HSPA1B* without accounting for (already suggested effects of) *CNR1* SNPs, cannabinoid use and their supposed interaction. Beyond that, selecting the two *CNR1* SNPs in a setting in which cannabis use is also considered a factor (but without a clear grading of intensity of use or dependence) while disregarding others that are known to be linked to cannabis dependence/use or effects, cognition and psychopathology (Ishiguro, et al., 2013) also does not seem to make much sense. One could consider potential other biases, since it remains unclear whether assessors were aware of patients' genotypes or whether "treatment as usual" was indeed delivered uniformly.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 June 2020 doi:10.20944/preprints202006.0226.v1

Researchers' degrees of freedom without "fishing". The requirement of predefinition of the statistical procedures could be generally outlined as follows (Gelman and Loken, 2013): i) predefine the outcomes and effects of primary interest (including potential interactions); ii) predefine statistical models including choice of control variables and possible transformations. Actually, the entire procedure is pre-chosen from a variety of possible procedures. In an opposite situation, an unlimited number of tests are performed and the "best results" are then reported. This is sometimes referred to as "p-hacking" or "researchers' degrees of freedom" (Gelman and Loken, 2013). Such a qualification implies that researchers purposely conducted many different analyses on a single data set in a dedicated search for "something publishable" (regardless of how far from the truth it may be), which however is not really a common practice – scientists do not generally do tthat (Gelman and Loken, 2013). Yet, it is possible to have an analysis that is highly dependent on the data without any conscious intention of "p-hacking", which Gelman and Loken (2013) named *researchers degrees of freedom without fishing*. It could be outlined as a situation in which a certain model is fit to the data, but another model fit to the same data would have resulted in a different estimate. The applied model could be defended relatively reasonably as being due to what *was seen in the data* and this could then extend to a *variety of associations* under the same or a similar model (Gelman and Loken, 2013). In essence, with no ill intentions, researchers use their scientific common sense to formulate reasonable hypotheses given the data they have. A typical mistake, however, is thinking that a flagged "statistical significance" is then evidence that supports the hypotheses (Gelman and Loken, 2013).

Multiplicity. Within the frequentist philosophy, a large number of null-hypothesis tests in a single empirical observation even when ideally protected from all possible biases results in an increased probability that at least one is falsely rejected (familywise error rate, FWER, increases), e.g., for 23 tests FWER=$1-0.95^{23}$=69.3%. This topic extends by far out of the scope of this case study; so let's just state that general views on how, when and why FWER should be considered and handled are still debated, but recognized in both frequentist and Bayesian

 doi:10.20944/preprints202006.0226.v1

frameworks (Hochberg and Westfall, 2000, Greenland and Hofman, 2019). In some settings, however, certain rules have been widely accepted: for example, in common-variant GWAS studies, P-value threshold is usually set at 5 x10^{-8} (Fadista, et al., 2016). False Discovery Rate (FDR) is not a method to control FWER (a probability) – it is a method of adjusting P-values in order to reduce the proportion of falsely rejected nulls among the rejected ones (Hochberg and Westfall, 2000). A statement that FDR was set at 15% means that among those P-values flagged as "significant" after FDR adjustment, there were 15% falsely rejected nulls (e.g., 1/6). To protect against *false discoveries* FDR is usually set 5%, or at 1%, e.g., in shotgun proteomics. The choice of 15% is difficult to understand. The control of FWER may also include grading of outcomes in terms of their importance, where those depicted as "primary", from which inference is intended, are scrutinized in respect to FWER, while those "lower-ranked", based on which no inference is intended, are treated more flexibly. Declarative definition of one or the other outcome of the same study as "primary" or as "lower-ranked" as one sees fit for the purpose of serial publications is a completely meaningless thing, as it does not protect against anything. The same goes for inference based on non-scrutinized outcomes ("explorations"), or for application of the same level of scrutiny to "primary" and "lower-ranked" outcomes. The "grading" of importance of hypotheses is then just a declarative form without a content.

In conclusion, the present case study illustrates how with no ill intentions researchers might take a path not likely to result in accurate estimates and how this could go unnoticed by the reviewers and, possibly, by at least some of the readers.

References

Border, R., Keller, M.C:, 2017. Commentary: Fundametnal problems with candidate gene-by-environment interaction studies – reflections on Moore et Thoemmes (2016). J. Child. Psychol. Psychiatry, 58, 328-330.

Bosnjak Kuharic, D., Bozina, N., Ganoci, L., Makaric, P., Kekin, I., Prpic, N., et al., 2020. Association of HSPA1B genotypes with psychopathology and neurocognition in patients with the first

episode of psychosis: a longitudinal 18-month follow-up study. Pharmacogenomics J., doi: 10.1038/s41397-020-0150-9

Fadista, J., Manning, A.K., Florez, J.C., Groop, L. 2016. The (in)famous GWAS P-value threshold revisited and updated for low-frequency variants. Eur. J. Hum. Gen. 24, 12020-1205.

Gelman, A., Loken, E., 2013. The garden of forking paths: why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time. http://www.stat.columbia.edu/~gelman/research/unpublished/p_hacking.pdf, accessed June 15, 2020.

Hochberg, Y., Westfall, P.H., 2000. On some multiplicity problems and multiple comparison procedures in biostatistics, in: Sen, P.K., Rao, C.R., (Eds.), Handbook of Statistics, vol. 18. Elsevier Science BV, pp. 75-113.

Hoehe, M.R., Morris-Rosedahl, D.M., 2018. The role of genetics and genomics in clinical psychiatry. Dialogues Clin. Neurosci. 20, 169-177.

Greenland, S., Hofman, A., 2019. Multiple comparisons controversies are about context and costs, not frequentism or Bayesianism. Eur. J. Epidemiol. 34, 801-808.

Ioannidis, J.P.A., 2005. Why most published research findings are false. PLoS Med. 2(8): e124. https://doi.org/10.1371/journal.pmed.0020124, accessed June 15, 2020.

Ishiguro, H., Leonard, C.M., Sgro, S., Onaivi, E.S., 2013. Cannabinoid gene variations in neuropsychiatric disorders, in: Murillo-Rodriguez, E., Onaivi, E.S., Darmani, N.A., Wagner, E. (Eds.), Endocannabinoids: molecular, pharmacological, behavioral and clinical features. Bentham Science Publishers, Bussum, The Netherlands, pp. 3-24.

Moore, S.R., 2017. Commentary: what is the case for candidate gene approaches in the era of high-throughput genomics? A response to Border and Keller (2017). J. Child. Psychol. Psychiatry, 58, 331-334.

Rojnic Kuzman, M., Bosnjak Kuharic, D., Ganoci, L., Makaric, P., Kekin, I., Rossini Gajsak, L., et al., 2019a. Association of CNR1 genotypes with changes in neurocognitive performance after

 doi:10.20944/preprints202006.0226.v1

eighteen-month treatment in patients with first-episode psychosis. Eur. Psychiatry, 61, 88-96.

Rojnic Kuzman, M., Makaric, P., Bosnjak Kuharic, D., Kekin, I., Rossini Gajsak, L., Boban, M., et al., 2019b. Integration of complementary biomarkers in patients with first episode psychosis: research protocol of a prospective follow up study. Psychiatria Danubina, 31, 162-171.

Supak Smolcic, V., 2013. Salami publications: definitions and examples. Biochemia Medica, 23, 137-141.