Published in final edited form as:
Clin Pharmacol Ther. 2014 August ; 96(2): 144–146. doi:10.1038/clpt.2014.100.

Interpreting the CYP2D6 Results From the International Tamoxifen Pharmacogenetics Consortium

MA Province[1], **RB Altman**[2], and **TE Klein**[2]

[1]Division of Statistical Genomics, Washington University School of Medicine, St. Louis, Missouri, USA

[2]Department of Genetics, Stanford University, Stanford, California, USA

Abstract

Meta-analysis of the entire analyzable cohort of 4,935 tamoxifen-treated breast cancer patients by the International Tamoxifen Pharmacogenetics Consortium (ITPC) (criterion 3) revealed no CYP2D6 effect on outcomes but strong heterogeneity across sites.[1] However, a *post hoc*–defined subgroup (criterion 1: postmenopausal, estrogen receptor positive, receiving 20 mg/day tamoxifen for 5 years; *n* = 1,996) did find statistically significant effect of CYP2D6 on both invasive disease–free survival as well as breast cancer–free interval, with little site heterogeneity. How should we interpret these discrepant findings?

> "Statistics: The only science that enables different experts using the same figures to draw different conclusions."

—Evan Esar, humorist

> "Data do not give up their secrets easily. They must be tortured to confess."

—Jeffrey Hooper, Bell Laboratories

> "Facts are stubborn, but statistics are more pliable."

—Mark Twain, humorist

If the ITPC investigators had been clever enough to have defined the various subgroup criteria *a priori* (i.e., before any data analysis), then it would be easier to interpret the results. In such an alternative universe, we would reason as follows. The full ITPC sample (criterion 3) shows no significant association between CYP2D6 and treatment response. However, in the (now *a priori*–defined) criterion 1 subset, CYP2D6 shows a significant effect on outcome. Either the subgroup analysis is a false-positive type I error or it is a true-positive finding and the lack of replication in the entire data set may be due to CYP2D6's being just one of several important biological pieces of the puzzle. Other genes or factors

Correspondence: TE Klein (teri.klein@stanford.edu).

CONFLICT OF INTEREST
R.B.A. is founder, scientific consultant, and stockholder for Personalis, Inc. T.E.K. is a scientific consultant and stockholder for Personalis, Inc.

that may be equally or more important in determining outcome might dominate in other subsets of breast cancer patients. This would explain the heterogeneity of results seen in the literature, as well as that seen in the ITPC itself.

Unfortunately, as we explained in detail in the Methods section of Province and colleagues' paper,[1] we did *not* define the criterion 1 subgroup until after some preliminary analyses had been conducted that did not recapitulate previously published results in much the same data. This lack of replication prompted further investigation of the data, a change in analysis strategy from a mega-analysis to a meta-analysis, and several rounds of additional data cleaning that identified suspicious values, which in turn required going back to each of the 12 source sites to collect more data and either confirm or correct the data, followed by additional analyses. Such an iterative process is actually not particularly unusual in retrospective studies like the ITPC, in which data must be harmonized *post hoc* across multiple source studies that used different measurement protocols, instruments, and procedures in originally collecting the data. Because no one person or group had seen the data from all 12 studies in one place before, the harmonization and cleaning process was a learning and illuminating experience to all involved. Furthermore, the act of analysis is always one of profound quality control as well as exploration. It reveals aspects of the data that were not appreciated in the initial cleaning procedures—no matter how carefully designed and skillfully executed. Once discovered, that information cannot be ignored but must be reported. Although the end result is often a better analysis of the data, the process is not always straightforward. Thus, it can be difficult, if not impossible, to do formal statistical accounting of the exact number of multiple comparisons of tests and analyses conducted along the way so as to properly correct for inflation of P values.

Formally, corrections for multiple comparisons are partly a question of numbers. The Bonferroni correction accurately adjusts the significance levels for multiple comparisons when the precise number of statistical tests done is known (and they are independent). At one extreme, in prospective clinical trials, a single, clear, primary hypothesis is tested, the associated P value of which can be taken at face value with no corrections for "multiplicities." If there are preliminary analyses of the accumulating data, they are preplanned and formally adjusted for multiple comparisons so as to allow the possibility of early termination without inflating the significance level of the final inference of that one hypothesis test at the end of the trial. At the other extreme, millions of statistical tests may be conducted, and they may even be partially dependent tests, such as in genome-wide association studies (GWAS), but an accurate correction can still be obtained because the number of tests is well defined and the way in which the tests are correlated is mathematically tractable to being undone (e.g., Gao *et al.*[2]).

Because of the extreme numbers of tests conducted in a GWAS, and the correspondingly high chance of a "winner's curse" false-positive finding, journals usually require that any findings not only exceed such strict genome-wide Bonferroni significance criteria but also be replicated in one or more completely independent data sets before the authors can claim a positive association. GWAS data collection has been criticized by some for setting the bar of proof too high, perhaps missing important signals lurking below such thresholds. Nonetheless, it has been enormously successful in identifying and statistically validating thousands of complex trait variants, some with quite small effect sizes that would have been out of reach of the earlier "candidate gene" experiments of the pre-GWAS era. In the statistical literature, this process of generating a large number of statistical tests and selecting the most extreme ones has been referred to as "hypothesis generation" as opposed to formal "hypothesis testing," to make the term sound more respectful than "fishing expedition," which derived from the legal field[3] and has negative connotations. When performed on an industrial, more automated scale, it is sometimes called "exploratory data analysis," "data mining," or "model selection." Regardless of what it is called, we believe that hypothesis generation should not be looked on with disdain but instead as an important part of the scientific process—a valuable complement to formal hypothesis testing. Indeed, even outside of the GWAS realm, where hypothesis generation works very well, there are concrete examples of important scientific discoveries that were first reported as "hypothesis generations," including the well-established association between apolipoprotein E genotype and Alzheimer's disease (the initial reports of which were dismissed by journal reviewers as biologically implausible and probably a type I error, until independent replication proved otherwise).

Even though we can't determine the exact number of tests conducted, certainly the number of analyses performed by the ITPC are much closer to the "one" of the prospective clinical trial (requiring only $P < 0.05$ α-levels) than they are to the millions of tests conducted in GWAS fishing expeditions (requiring $P < 10^{-8}$ α-levels).

Additionally, multiple repeat analyses of the same hypothesis test using updated or corrected versions of the data should not be penalized at the same Bonferroni rate as truly independent tests (but exactly how much should they be penalized?). However, there is a deeper problem here in interpreting the ITPC results than the simple numbers game of not knowing what Bonferroni denominator to use to correct for the number of hypotheses tested along the way. By defining subgroups guided by positive preliminary analyses, the ITPC passed over from formal hypothesis testing to hypothesis generation in ways that cannot be formally corrected by any known correction. Even though criterion 1 has high face validity, there may be many other subsets of the data that also have face validity, each of which may or may not show a positive association. How do we assess whether we are fitting to noise or signal in such process? Do we correct for all possible subsets that can be formed? Alternatively, should we correct for only subsets that test positive? Rather than being defined by an automated machine learning procedure—whose propensity to fit to noise we could formally evaluate in repeated simulation experiments—the subgroups were defined by experts using their clinical knowledge and prior experience. We could devote a lot of effort to attempting to "adjust" these results and not gain much insight.

So where does this leave us, and how do we interpret the ITPC results? We have an overall negative study, so it is clear that the CYP2D6 is at best one of several factors influencing outcome following adjuvant tamoxifen treatment. We have generated a hypothesis defining criterion 1, which has clinical face validity and does show a positive association to outcome in the ITPC. Unfortunately, it appears impossible to formally adjust the significance levels to account for the "generation" process. But we believe that it is important to report those findings as best we can and to inform the readership as to exactly how they were generated.

At the same time, we are extremely sensitive to the concerns of the CODATA-ICSTI Task Group on Data Citation Standards and Practices, which notes that "there is even evidence that researchers engage in data dredging, model fishing, or other methodological peccadilloes in search of results that are 'significant.'"[4] This is why we made a full disclosure of the (admittedly flawed) process by which these subgroups were defined; furthermore, we have exceeded the Task Group's recommendations for data citation by making all the ITPC data and all analysis programs publicly available. To complete the hypothesis-generation experiment, our generated hypothesis about criterion 1 should be taken into an independent data set in which it can be formally tested—as we called for in the conclusion of our paper.[1] In the end, the field will determine the truth. Since the formation of the ITPC, additional studies on the subject have already emerged, and more are expected to come. We hope the ITPC will have played a part in discovering the truth about the role of CYP2D6 in tamoxifen therapy.

References

1. Province MA, et al. *CYP2D6* genotype and adjuvant tamoxifen: meta-analysis of heterogeneous study populations. Clin. Pharmacol. Ther. 2014; 95:216–227. [PubMed: 24060820]
2. Gao X, Becker LC, Becker DM, Starmer JD, Province MA. Avoiding the high Bonferroni penalty in genome-wide association studies. Genet. Epidemiol. 2010; 34:100–105. [PubMed: 19434714]
3. Holmes OW. Ellis vs. Interstate Commerce Commission, 237 US 434, no. 712, US Supreme Court. 1915
4. Socha YM. CODATA-ICSTI Task Group on Data Citation Standards and Practices. Out of cite, out of mind: the current state of practice, policy, and technology for the citation of data. Data Sci. J. 2013; 12:1–75.