Visualization of Variance in LA2K*

D.S. Parker, E. Congdon, R.M. Bilder
Consortium for Neuropsychiatric Phenomics (CNP)
University of California, Los Angeles

June 20, 2013

Abstract

The LA2K study conducted at UCLA was an investigation into the biological bases of traits such as memory and response inhibition phenotypes, and to explore whether they are linked to syndromes including ADHD, Bipolar disorder, or Schizophrenia. An aim of the study was in moving from traditional categorical approaches for psychiatric syndromes towards more quantitative approaches based on large-scale analysis of the space of human variation. It represents an application of phenomics — wide-scale, systematic study of phenotypes — to neuropsychiatry research.

This paper reports on a system for visualization of variance in the LA2K data. More specifically it describes a system for exploratory data analysis called VIVA LA2K for visual exploration of the data, with visualization methods that are aimed at ‘explanation’ of the variance through mathematical models of variance. An example of these methods is called VISOVA, a combination of visualization and analysis of variance, with the flavor of exploration associated with ANOVA in biomedical hypothesis generation. The system was designed for hypothesis exploration by interdisciplinary teams, and more specifically for grounding hypotheses in data — thereby enlisting forces of natural selection to improve the hypothesis pool.

Variance structure models include dimensional models, which usually have factors that can each represent a continuum of possibilities. These models have been used widely for characterizing dimensions of personality, and are becoming more widely used now for neuropsychiatric disorders. This article describes how dimensional models fit with variance structure models in VIVA LA2K for describing the architecture of disorders, and can be used for interactive improvement of neuropsychiatric hypotheses.

1 Introduction

1.1 The LA2K Study

The LA2K study [1, 3] was conducted at UCLA during 2008–2012, designed as a large-scale analysis for about 2000 volunteers from the Los Angeles metropolitan region. Behind its development was the hypothesis that there may be sufficient variance in healthy people along dimensions shared with people with psychopathology, that we might find common mechanisms with genetic links. For example, the genetic bases for variability in working memory in healthy people may also be the basis for working memory impairments commonly found in patients. LA2K focused on the evaluation of memory and response inhibition as central endophenotypes [9] having potential to serve as basic dimensions for neuropsychiatry. Ultimately this approach also could permit development of statistical models characterizing neuropsychiatric syndromes, without relying on traditional taxonomies and their discrete categories.

LA2K was developed in part, then, as a demonstration of the use of phenomics as a framework for neuropsychiatry, and hopefully a way to advance the pace of research. Phenotypes — detectable or measurable characteristics of an organism — are outward manifestations of interaction of its genotype and environment; phenomics is the systematic study of biological and behavioral phenotypes [3]. Because phenotypes are present in all scales of science, phenomics is interdisciplinary and intrinsically large-scale in scope. LA2K represents a 6-year effort aimed at the discovery of relationships that would go undetected with a smaller scope.

1.2 The LA2K Database

LA2K study results are stored as a relational database with about 50 tables together comprising over 2500 numeric variables (columns), where each column representing a single (but not independent) phenotype or measurement, and each table reflects an experimental protocol. Figure 1 shows names of cognitive tasks and personality rating scales represented in LA2K tables. With approximately 1300 subjects (originally intended to cover only healthy control subjects, not suffering from any major syndrome, but subsequently expanded to include some patients diagnosed with ADHD, Bipolar Disorder, and Schizophrenia) the database is the result of a large study with great opportunities for discovery.

In general terms, the LA2K database permits investigators to analyze the behavior of subjects in neurocognitive measures that reflect memory and response inhibition (recording reaction times and accuracy measures), as well as aspects of temperament, personality, and syndromal behavior. Each table in Figure 1 has data for the relevant LA2K subjects.

1.3 Dimensional Hypotheses about Human Variation

Dimensional models in neuropsychiatry are descriptions of behavior or mental state in terms of multiple independent factors that can each take on a continuum of possible values. For example, dimensional models of personality usually include dimensions of extraversion, neuroticism, and psychoticism [6].


<table>
<thead>
<tr>
<th>LA2K Domain/Test Abbrev</th>
</tr>
</thead>
<tbody>
<tr>
<td>Consent/Screening/Diagnosis/Clinical Rating Scales</td>
</tr>
<tr>
<td>Adult ADHD Interview (module from KSADS-PL) AAI</td>
</tr>
<tr>
<td>Hopkins Symptom Checklist 25 HSCI25</td>
</tr>
<tr>
<td>Structured Clinical Interview for DSM-IV/Axis I/Patient version SCID-IP</td>
</tr>
<tr>
<td>Personality/Temporality/Symptom Questionnaires</td>
</tr>
<tr>
<td>The Temperament and Character Inventory TCI</td>
</tr>
<tr>
<td>The Chapman Scales – Physical Anhedonia RPAS</td>
</tr>
<tr>
<td>The Chapman Scales – Social Anhedonia RSAS</td>
</tr>
<tr>
<td>The Chapman Scales – Perceptual Aberrations PAS</td>
</tr>
<tr>
<td>Eckblad and Chapman’s Hypomanic Personality Scale HPS</td>
</tr>
<tr>
<td>Golden and Meehl’s 7-Item Schizoid Scale G+M</td>
</tr>
<tr>
<td>Munich Chromotype Questionnaire MCTQ</td>
</tr>
<tr>
<td>Akiskal’s Bipolar II Scale BP-II</td>
</tr>
<tr>
<td>Barratt Impulsivity Scale ISPQ</td>
</tr>
<tr>
<td>Eysenck’s Impulsivity Venturesome and Empathy Inventory IVE-R</td>
</tr>
<tr>
<td>MPQ (Control-Impulsivity items) MPQ-CI</td>
</tr>
<tr>
<td>The Dickman Scale of Functional vs Dysfunctional Impulsivity DSFDI</td>
</tr>
<tr>
<td>Neurocognitive Measures</td>
</tr>
<tr>
<td>Spatial &amp; Verbal Memory and Manipulation Tests SMNM/VMMNM</td>
</tr>
<tr>
<td>Remember-Know Paradigm RK</td>
</tr>
<tr>
<td>Scene Recognition Task SR</td>
</tr>
<tr>
<td>California Verbal Learning Test CVLT-II</td>
</tr>
<tr>
<td>WMS-III Spatial Span WMS-SS</td>
</tr>
<tr>
<td>WMS-III Digit Span WMS-DS</td>
</tr>
<tr>
<td>WMS-III Visual Reproduction (part III [immed/delayed recall]) WMS-VR3</td>
</tr>
<tr>
<td>WMS-III Letter Number Sequencing WMS-LNS</td>
</tr>
<tr>
<td>Stop-Signal Task SST</td>
</tr>
<tr>
<td>Conner’s CPT II CPT-II</td>
</tr>
<tr>
<td>Reversal Learning PRLE</td>
</tr>
<tr>
<td>Task Set Switching TS</td>
</tr>
<tr>
<td>Stroop Test SCWT</td>
</tr>
<tr>
<td>Attention Networks Task ANT</td>
</tr>
<tr>
<td>Delay Discounting DD</td>
</tr>
<tr>
<td>Balloon Analog Risk Task BART</td>
</tr>
</tbody>
</table>

Dimensional models are increasingly used in characterizing personality disorders and mental disorders. The established DSM-IV [28] taxonomy and method for mental disorders has some dimensional aspects. More recently many disorders have been grouped into Internalizing and Externalizing dimensions [21], and these have become part of the ‘meta-structure’ of DSM-V [29], the new Diagnostic and Statistical Manual of mental health. DSM-V includes dimensional classification — quantitative approaches to assessment and diagnosis [30], as a step beyond Chinese menu diagnosis and categories or rubrics that do not always fit, and toward quantitative assessments of severity and treatment response. The stakes involved in this evolution are breathtaking, as DSM-IV is a cornerstone of the mental health system.

Still, categories such as ADHD, Bipolar Disorder, or Schizophrenia are often said to be inadequate for classification, as they rest on inaccurate descriptions, and dimensional models have raised as a possible resolution. The NIMH Strategic Plan aims at branching from existing diagnostic categories for mental disorders, in part because the categories lack validity and because they limit incorporation of new scientific results. A criticism often leveled against the DSM is that it is not aligned with any scientific model of neuropsychiatric disorders [10]. A related criticism is that different diagnoses overlap significantly, in some cases using different terminology for the same concept. The variance structure models described below can sidestep both these criticisms.

Dimensional models are growing increasingly sophisticated [22]. They are often taken to be general linear models (GLMs [26]) or structural equation models (SEMs [23]), permitting characterization of variance structure with a set of functional equations. An emphasis on dimensional models has been developed within the NIMH Research Domains Criteria (RDoC) project [7, 8]. RDoC encourages exploration of dimensional models in diagnosis, permitting continuous variables of function, ranging from behavior down to neurobiology.

1.4 Wide-Phenome Analysis

The LA2K data was designed around a schema spanning 7 levels of neuroscience, from genome to syndrome, centering on memory and response inhibition phenotypes [3]. The levels were designed to facilitate development of new models for syndromes.

Figure 2 shows a correlation matrix for a set of over 150 LA2K variables as a consensus representation of much of its variance. The matrix exhibits limited pockets of positive correlation, reflecting the careful selection of these variables so as to be independent. Generally the matrix shows low correlation outside these pockets, and some anti-correlation between variables at the start (reaction time variables) and variables at the end (total raw score variables).

The dendrograms in Figure 2 show block structure, and suggest how the variables could be partitioned into clusters representing dimensions of significant variance. A benefit of using a broad range of phenotypes (‘wide phenome’) approach like this is that we can attempt to explain the variance in the population with models. For example, dimensional models such as PCA [27] can be used with this matrix to show that its first 9 eigenvalues explain 33% of the variance, and the first 20 explain 50%. Naively, the matrix might be summarized with about 20 dimensions.

However, decomposition can capture much more variance structure if we condition on features of the population. Multi-level models permit conditioning on group or factor features such as gender or age, or syndrome like ADHD, BP, or SZ. This ‘hybrid’ blend of categorical and quantitative modeling permits explanation of differences in variance for different population groups. An example is shown in Figure 3, a dimensional model in which the variance structure will change if we condition on either the ADHD, BP, or SZ group. Multi-level variance structure of this kind is fundamental to analysis of variance (ANOVA) [26], and linear modeling that spans groups.

Principal Components Analysis [27], shown in Figure 3, is a popular method for exploring variance structure. Group structure
of controls and patients is visible in projections along the first three components. Models like these can help map out the space of human variation, and this could lead to sound diagnostics for neuropsychiatric syndromes.

![Figure 3: The projection of the LA2K data on the first 3 principal components (PC1, PC2, PC3) for the LA2K correlation matrix in Figure 2. Two views are offered — pairwise projections and a 3D view — showing the relative positions of healthy controls (light orange), ADHD patients (blue), Bipolar patients (red), and Schizophrenia patients (purple). Schizophrenia patients have extreme coefficient values on the first two principal component, and both BP and ADHD patients on the second and third. Very roughly, the three components respectively emphasize total raw scores in tasks, scores measuring psychological stress and mental health, and scores reflecting working memory performance.](http://datamining.cs.ucla.edu/cgi-bin/la2k)

An initial motivation for LA2K was to support transdisciplinary collaborations in studying wide phenotype spaces of human variation, in order to obtain new characterizations of neuropsychiatric syndromes. A corresponding objective for ViVA LA2K is to develop interactive methods for explaining variance across human populations. ViVA LA2K is novel in directly linking visualization with variance structure models, permitting interactive visual exploration of dimensional representations of disorder. The point here is that these goals, and goals set for dimensional visualization with variance structure models, permitting interactive visual exploration of dimensional representations of disorder. The point here is that these goals, and goals set for dimensional models, fit the variance structure models described below.

2 ViVA LA2K

This project initially began as a server aimed at sharing of exploratory data analysis scripts for the LA2K data. These evolved, and at some point it became clear that they emphasized visual analogues of ANOVA — including what we call ViSOVA — holding some variables fixed, varying others, and permitting visualization of the resulting ‘response’.

2.1 ViVA LA2K Architecture

ViVA LA2K is actually three related systems, each providing a different mode of exploration:

- **LA2K Atlas** — ‘gallery’ of results for predefined hypotheses ([http://datamining.cs.ucla.edu/cgi-bin/la2k_atlas.cgi](http://datamining.cs.ucla.edu/cgi-bin/la2k_atlas.cgi))
- **LA2K Viewer** — simple menu-based hypothesis exploration ([http://datamining.cs.ucla.edu/cgi-bin/la2k_viewer.cgi](http://datamining.cs.ucla.edu/cgi-bin/la2k_viewer.cgi))
- **LA2K Explorer** — advanced development of hypotheses ([http://datamining.cs.ucla.edu/cgi-bin/la2k_explorer.cgi](http://datamining.cs.ucla.edu/cgi-bin/la2k_explorer.cgi)).

These can be used in sequence, progressing from initial passive orientation to active exploration. With each, investigators gradually refine a rough hypothesis, which is a variance structure model — explaining variance in the LA2K dataset.

The architecture common to these three systems aims at making exploration as effortless as possible. Succeeding at this implies many things: automated data cleaning, implementation of group tables, creation of tables and fields wherever needed, development of web infrastructure, and introduction of features aimed at simplifying interaction and rising above the complexity of the LA2K dataset. ViVA LA2K consists of about 50,000 lines of Python code, which is approximately equally split between web interface and data management.

Experience with ViVA LA2K has highlighted benefits of this architecture for visualization of variance:

- implementing visualization with a server is advantageous for pooling effort (such as in data cleaning and visualization scripts), and for maintaining best practices and standards (such as verification of distributions).
- using a universally-understood vocabularies of visualization and exploratory data analysis (EDA [23] and ANOVA) can be an effective medium for trans-disciplinary work, particularly among scientists with little programming experience.

The interface of **LA2K Explorer** is the most sophisticated: hypotheses can involve any subset of the 2500 phenotype variables in LA2K, and can condition on any of 60 group structures in 18 predefined subsets of LA2K (as well as on all LA2K experimental protocols). The other two interfaces provide less flexibility, but all three interfaces allow specification of a variance structure model.

2.2 Variance Structure Models

**Variance structure** is a common term in data analysis. It typically refers to patterns of variation in statistical models related to distributions (variance of individual variables), covariance and correlation of variables, and more general equational relationships among variables. A **variance structure model** is a mathematical expression of these relationships.

Important classes of variance structure models include general linear models (GLMs [26]) and structural equation models (SEMs [25]). These permit characterization of variance structure in terms of a set of functional equations that are often linear in form. However they permit nonlinear interactions (nested models) and conditioning on non-numeric variables or factors that take discrete values (multi-level models).

Variance structure is often exposed by an incremental process of decomposition or factoring, yielding a hierarchy or graph of components that together form an overall model. For example, Principal Components Analysis (PCA) [27] is a linear algebraic model of structure in covariance matrices. From the standpoint of decomposition, this process also can differentiate clusters of similar variables from others, and extract hierarchical block structure.

In ViVA LA2K, a variance structure model $M: Y \sim X \mid G: (S)$ is a specification of a rough hypothesis with five items:

- a subset $S$ of the data (a population of subjects that is meaningful for analysis).
• a grouping \( G \), specifying a set of class names (factor levels) defining several subpopulations.

• zero or more dependent variables \( Y \) from LA2K tables. (If zero, all variables are independent.)

• one or more independent variables \( X \) from LA2K tables.

• a model of variance \( M \) (a mathematical model and/or visualization method for explaining variance).

In particular, the model

\[
\text{VISOVA: ReactionTime} \sim \text{Age} \mid \text{Gender} : (\text{LA2K control})
\]

relates reaction time to age for two gender groups, female and male, which define population subsets — in this case restricted to control subjects. Multi-level models with this kind of conditioning are basic to ANOVA, explaining variance across the different values of a factor, such as the Female and Male values of Gender.

Multi-level models are important in neuropsychiatry for other reasons, including their basic connections with nosology and diagnosis. Among these, a fundamental aspect of conditioning is that it makes hypotheses differential — they consider not only a base hypothesis with a single level value, but also alternative level values. Without this conditioning hypotheses are difficult to falsify, and difficult to make mutually exclusive. As a result they are difficult to verify or contradict, and they all can be ‘right’. This lack of exclusivity, and the difficulty of grounding hypotheses in data, impedes progress [5].

The specification of these five items by selection is managed entirely with a large hierarchical menu; altogether the menu for LA2K Explorer includes over 3200 individual options. Together the five selection types specify hypothesized effects in the data in terms of a model of variance structure.

2.3 Visualization of Variance

ViVA LA2K integrates variance structure modeling with visualization. It can be used to generate a selection of visualizations for a rough hypothesis. The visualizations are compiled into reports of related results in a way that permits exploratory collaborative hypothesis development.

Variance structure models such as \( M: Y \sim X \mid G : (S) \) resemble the formulas \( Y \sim X \mid G \) supported by linear models and ANOVA functions in the S and R statistical computing environments [26]. They are also a mainstay of trellis graphics [24], an influential visualization toolkit in these environments. In trellis graphics, the values (factor levels) of conditioning variables define an array or grid (‘trellis’) of similar visualizations, allowing side-by-side comparison across these values. Thus the conditioning variables yield automatic construction of grids of visualizations.

This side-by-side display is a powerful basic tool for visualizing variance, although it is not presented or developed in this way. A new emphasis of ViVA LA2K is in directly linking visualization with variance structure models, permitting interactive visual exploration of dimensional representations of disorder.

How does ViVA LA2K differ from existing visualization systems? For example, the Clinical suite of tools in Spotfire [31] [32], a widely-used visualization framework, provides summary statistics, relation plots, and many other types of charts. Most statistical computing environments provide these functions as well. Is ViVA LA2K fundamentally different, or more powerful in some way? Although it doesn’t provide more functions or analytical power, ViVA LA2K is novel and has some unique strengths.

First, ViVA LA2K focuses on visualization of variance. It integrates variance structure models with visualization, permitting exploration of hypotheses across groups; this is unique. For example, Figures 4 and 5 below show examples of VISOVA (to be outlined below) — a novel integration of ANOVA, clustering, and parallel coordinates. Also it generates reports with sets of visualizations about variance, rather than single plots.

Second, ViVA LA2K goes to lengths to make exploration effortless. It consists of about 50,000 lines of Python code, which is equally split between web interface and back-end data management. The latter involves automation of updates in a data extraction and cleaning pipeline, implementation of group structures and subpopulations, creation of useful extra tables and fields, and visualization functions that have been improved over time. These things can be implemented in any environment, but LA2K is a complex dataset, and ViVA LA2K provides three natural exploration modes (Atlas, Viewer, Explorer). This is particularly beneficial to large interdisciplinary research projects, which include investigators with diverse backgrounds and skills.

3 Visualization of Variance in LA2K

To illustrate the use of ViVA LA2K, we sketch scenarios of its use. These applications illustrate points that motivated the development of LA2K, such as the value of a wide-phenome approach. However the main point they illustrate is that exploration can improve hypotheses. Although it cannot validate or prove hypotheses, it can ground them in data — and by doing this introduce forces of natural selection that yield better hypotheses.

This section offers a few examples to illustrate what ViVA LA2K has to offer for hypothesis development and visualization of variance. More examples are available at http://www.hypweb.org. The LA2K Atlas (http://datamining.cs.ucla.edu/cgi-bin/la2k_atlas.cgi) also offers complete gallery-like collections of visualizations for predefined hypotheses on tables in the LA2K database.

3.1 Sample Hypothesis: Age and Reaction Time

Consider a rough hypothesis that reaction time (RT) is affected by age. We can explore this hypothesis by considering all LA2K Mean RT values (joined across tables). Figure 4 shows a VISOVA (parallel coordinates/ANOVA) visualization for a table containing all key MEAN RT variables, essentially a sequence of variance structure models VISOVA: ReactionTime \sim Age \mid Gender : (LA2K control), in which each column shows the effect of age on one RT variable. This axis labels shows all of the MEAN RT variables included and the columns representing variables are clustered according to correlation similarity. In other words, each column represents the value range of a single variable, and trajectories across the columns give the sequence of RT values for a single subject.

VISOVA displays extend parallel coordinates with group structure. Individual group averages for the variables are also superimposed as thicker lines along with standard error bars; the colors reflect the 7 age ranges used as groups. The thick red line at the
Figure 4: Mean reaction time values across the LA2K database, from all tasks, show similar effects of age on reaction time. The averages by age group are increasing and consistently ordered across all variables, showing a progressive slowing of reactions with age. (This is an example of VISOA, an integration of parallel coordinates plots [24] with analysis of variance: if horizontal trajectories in the diagram were removed, and made to emphasize the group averages for each variable (column), this could be a MANOVA display.) click to explore

top shows the average for subjects aged 46–50, while the blue line near the bottom shows the average for subjects aged 21–25. The gradual increase of RT over these age groups is consistent across all Mean RT variables, suggesting an ongoing process of decline (progressive increase in mean reaction time) with age, regardless of task or type of RT measure (e.g., congruent vs. incongruent RT). This continuum in this effect suggests a dimensional model, and that a quantitative model of decline is a possibility.

The pattern here appears strong, encouraging exploration of still stronger hypotheses. For example, the same variance structure model could be specialized with other factors — such as demographic factors like gender or ethnicity, and behavioral factors like smoking habits. Any of about 60 predefined factors in ViVA LA2K could be used. Changing the hypothesis above to incorporate new group structure or include other variables can be done immediately with the interaction menu. The interface is designed to encourage exploration of the hypothesis evolutionary landscape.

### 3.2 Sample Hypothesis: Timing and Schizophrenia

The LA2K database includes a large number of temporal phenotypes, many in the form of reaction time (RT) measurements. Hypotheses regarding the characterization of Schizophrenia in terms of interval timing have emerged recently [13, 14]. Various deficits in temporal processing are associated with symptoms of the disorder. The database has many temporal variables; for example ViVA LA2K can explore results about RT CV (CV = coefficient of variation = $\sigma/\mu$) for reaction measures. Assuming that poor timing control is reflected by higher RT CV values, this gives a way of checking whether there is a distinction between controls and Schizophrenia patients. An advantage of using CV is that it is less sensitive to the underlying reaction time distribution. Some related published work is in [11], including support for using CV in measuring reaction time [12].

Figure 5 shows a VISOA display of CV values that cut across tasks in LA2K, with the variance structure model VISOA: $\sim$ Reaction Time CV $| G : (LA2K$ control or patient$)$, where $G$ is an ad hoc group constructed to include control subjects and Schizophrenia patients. In this plot, Schizophrenia patients have uniformly higher CV values, and with the exception of a few variables in the middle of this plot, the differences are greater than the standard error (error bars). click to explore.

Figure 5: Comparison between controls (blue) and Schizophrenia patients (red) on the Coefficient of Variation (CV = $\sigma/\mu$) for Reaction Time measures that cut across tasks in LA2K. Schizophrenia patients have uniformly higher CV values, and with the exception of a few variables in the middle of this plot, the differences are greater than the standard error (error bars). click to explore.

3.3 Sample Hypotheses: Genetics and Ethnicity

Next we offer two examples in which ViVA LA2K raised doubt about hypotheses. Both involve ethnicity.

The very recent study [15] suggested possible links between health and ethnicity. This conclusion was based on a discovery that a region of the human genome that codes for many antibodies has sections that can be absent, and this variation can depend on ethnicity. Thus, ethnicities might have different health profiles.

LA2K was designed in part for analysis of Hispanic ethnicity, and about 40% of the LA2K subject population is Hispanic. LA2K also has some variables related to health, offering a way to ground the health hypothesis in data. However, when used to explore differences of health by ethnicity, ViVA LA2K showed no visible health differences between Hispanic and non-Hispanic subjects.

To put this in perspective, it is not hard to find ethnic differences that could have links to health. For example, Figure 6 shows clear differences for a few physiological variables (Height, Weight, and BMI), so variance structure is evident in these variables. How-
ever, this variance does not extend to the LA2K Health Score — a sum across 21 items asking about diagnoses with, or treatment for, a number of serious medical illnesses. No pattern is apparent in Covariance: Health ~ BMI | Ethnicity : (LA2K control), and exploration did not find any health profile linked to ethnicity. Follow-up inspection of immune-related loci is possible, but these negative outcomes encourage other directions for hypothesis evolution.

LA2K also was designed for genotype-phenotype analysis, and includes SNP genotyping of its subjects. GWAS (genome-wide association studies) has been the focus, but exploration of particular genes is possible. For example, a specific variants of the DRD4 gene associated with Novelty Seeking (NS) behavior. A well-known hypothesis linking the two is that NS was important for human migration out of Africa 50,000 years ago [16], and a careful recent analysis [17] confirmed an association between the DRD4 2R and 7R polymorphisms and migratory distance.

A hypothesis for LA2K then might be that subjects of Hispanic ethnicity could have higher NS scores. Initially, grounding this hypothesis in LA2K showed the opposite: NS-related summary variables from the TCI yielded lower values for Hispanic controls. However, the LA2K ‘ethnicity’ classification is self-reported, and controls were screened for ADHD; if we consider both patients and control subjects, Figure 7 shows that the total NS measure from the TCI inventory covaries with many other personality and symptom scales — as well as Age (on the far left) and Ethnicity (on the far right).

Thus VISOSA: ~ Ethnicity | NS : (LA2K control and patient) is both encouraging and discouraging as a model: because Age appears anti-correlated with Novelty Seeking, the older the population we consider, the more reduced the effect we may see. Similarly, the more Harm Avoidant the population (a trait of older individuals, as well as e.g., Female MDD subjects discussed below), the more muted the effect.

PCA can check this further, as shown in Figure 8. We can see the Ethnicity axis is aligned with the Novelty Seeking gradient.
to the upper right. This offers some evolutionary support for the hypothesized association between Hispanic Ethnicity and Novelty Seeking (and possibly DRD4 polymorphisms), but again also illustrates how it would fail to hold. Without carefully-formalized restrictions on on population, hypotheses here risk being meaningless.

3.4 A Challenge: Gender and Disorder Prevalence

As study datasets grow larger, systems like ViVA LA2K may become important for developing better neuropsychiatric diagnostics. An example of some possible applications appeared recently, in an investigation of differences in disorder incidence rates by Gender [21]. This work specifically considered patterns of comorbidity in the large \((n = 43,093)\) National Epidemiologic Survey on Alcohol and Related Conditions (NESARC). The results offer an overall statistical outline or architecture for disorders, clarifying how they impact men and women differently.

Many studies have found that women have a higher incidence of internalizing (mood and anxiety) disorders, while men have a higher incidence of externalizing (antisocial and substance use) disorders [18]. In fact this is reflected in the overall internalizing-externalizing dimensional model of DSM-V [29]. This difference is easy to check with ViVA LA2K; Figure 9 shows that the prevalence of MDD (Major Depressive Disorder, an internalizing disorder) and AAD (Alcohol Abuse and Dependence, an externalizing disorder) in LA2K both follow the expected gender-specific liability profile.

Recently a dimensional model of comorbidity among these disorders appeared in [21], with the surprising claim that the comorbidity prevalence of disorders is identical in both genders. That is, once the latent gender-specific liability levels are conditioned for, the structure of common disorders is gender-invariant. This result suggests a single overarching disorder structure. It also bypasses a key problem in analyzing comorbidity — in making the assumption that the categories of the disorders are valid. If the disorders are dimensional, comorbidity should be dimensional also.

Reported levels of disorder prevalence by gender have varied [18, 19, 20], although the outline of internalizing and externalizing liability has been confirmed. These confirmations were important in the arduous DSM-V development process, which has required years of deliberations and field trials [29]. The ability to explore the structure of human variation in large databases like NESARC, with systems like ViVA LA2K, could be a way to increase consensus and incorporate scientific models into practice.

Figure 10 might indicate support for the gender-invariance claimed in [21]. The left image shows the psychological stress/mental health profile for all MDD (Major Depressive Disorder) subjects, and the right shows this for all AAD (Alcohol Abuse and Dependence) subjects, among LA2K controls — where these groups were defined based on a previous diagnosis. The profile is a set of summary scores from the HSCL-25 Hopkins System Checklist, and ASRS ADHD Self-Report Scale, giving a broad assessment of mental health. In both images affected males are shown in red, and females in blue; similarly control males are orange, and control females are green. Age and Smoking are strongly correlated with both MDD and AAD. The average profiles for MDD and AAD are similar across the spectrum of Hopkins and ASRS scores, and the distinctions between male and female subjects are not pronounced. This grounding in data has not yet endangered the gender-invariance hypothesis.

4 Conclusion

Phenomics and phenotype databases like LA2K are natural settings for hypothesis exploration. In biomedicine, hypotheses often concern variance structure — patterns of variation in variables when group and population structure are controlled. The abundance of extensions for ANOVA in the statistical and biomedical literature [26] show the importance of variance structure. However ViVA LA2K is the first system to provide VISOVA, facilitating exploration with a combination of parallel coordinates visualization and ANOVA.

In ViVA LA2K, a variance structure model is an assertion \(M: Y \sim X \mid G: (S)\) controlling variables \(G\) in population \(S\), representing variance \(M\) among the variables \(Y\) and \(X\). For example, in VISOVA: Reaction Time \(\sim\) Age | Gender: (LA2K control) represents changes in reaction time by age between male and female populations. There should be significant difference in the association between variables \(X\) when conditioned on the different groups. Other models in this paper have relied on Covariance matrices and PCA [27].

ViVA LA2K provides a way, even for scientists with little programming experience, to ‘try hypotheses on for size’ by grounding them in data — in any subset of the 2500 variables in LA2K, conditioning on any of 60 group structures in 18 predefined subsets of LA2K (as well as on all LA2K experimental protocols), using any of a large set of standard variance visualization schemes, without concerns about implementation or details of data cleaning (because these steps are provided by ViVA LA2K). To avoid confusion about scientific validity of the results, or Bonferroni correction, it is also intentionally limited to hypothesis exploration without confirmation. The exploration process is one of rapid evolution under selection, with stronger hypotheses surviving.

Amidst the deluge of data in which scientists now find themselves, it is vital to integrate relevant information with increasingly complex hypotheses. ViVA LA2K is an example of ways science can expand from established hypothesis-based processes to more data-driven, discovery-based processes that can benefit from the abundance of information.

References

Figure 9: Internalizing disorders such as MDD (Major Depressive Disorder) are known to have higher prevalence in women, and externalizing disorders such as AAD (Alcohol Abuse and Dependence) to have higher prevalence in men. ViVA LA2K permits rapid verification of this in LA2K controls. Group sizes for MDD are on the left, and AAD on the right (normal females: green, affected females: blue; normal males: orange, affected males: red).

Figure 10: As a reflection of ‘comorbidity’ of MDD (Major Depressive Disorder) and AAD (Alcohol Abuse and Dependence) with other psychological disorders, we can look at their average profiles against other mental health indicators and measures of psychological stress. The figures here display the psychological stress profiles of MDD (left) and AAD (right) across summary scores from the Hopkins System Checklist (HSCL-25). They include two summary scores from the Adult ADHD Self-Report Scale (ASRS) in order to track associations with ADHD, and Age and Smoking (Cigs) are also added as checks. The parallel coordinates displays here show similar average profiles. It has been known for some time that internalizing disorders such as MDD (Major Depressive Disorder) have higher prevalence in women than in men, and externalizing disorders such as AAD (Alcohol Dependence) have higher prevalence in men than women (normal females: green, affected females: blue; normal males: orange, affected males: red).


[28] DSM-IV (Diagnostic and Statistical Manual of Mental Disorders, IV): http://www.psych.org/MainMenu/Research/DSMIV.aspx

[29] DSM-V (Diagnostic and Statistical Manual of Mental Disorders, V): http://www.dsm5.org/

