Many medical diagnoses begin when an enterprising clinician notices a constellation of signs and symptoms across multiple patients (Jablensky). The correlated occurrence of signs and symptoms, called a syndrome, may indicate a disease or group of diseases, but there is no guarantee. The most optimistic assumption is that the syndrome captures a unitary etiological entity, but this is rarely true. The next assumption is that the syndrome harbors different diseases with similar clinical pictures but distinct etiologies and pathophysiologies. Parsing the syndrome via clinical and laboratory examination allows for patient stratification and targeted treatment prescription.

This assumption underlies reliance on contemporary psychosis diagnoses. Belief in parsing clinically defined psychiatric syndromes into distinct diseases anchored Robins and Guze's classical approach (REF). DSM-III framers (REF) and supporters (REF) adopted this view to back their revolutionary reconceptualization of psychiatric diagnoses, including serious psychiatric conditions like psychosis.

However, there may be little chance of carving diseases from current clinical psychosis diagnoses, like those in the DSM (REFs). Despite calls from the psychiatric establishment for patience with a gradual iterative process (Kendler), there are reasons for pessimism. Contemporary DSM diagnostic criteria derive from committee consensus; committees are not enterprising clinicians. Wisdom of crowds (REF), one form of group decision-making, assumes many independent judgments. Independence of thought is not the method used for constructing psychiatric diagnostic manuals. Collective decision-making, the method used for DSM construction, is prone to systematic errors and cognitive biases (REF). Consensus means compromise, which can distort the optimal solution. Current psychosis diagnoses may have wandered from the historically significant syndromes they were meant to operationalize (Carpenter). Meehl (REF) warned how conceptual drift, in the spirit of psychometric efficiency, can markedly and hopeless morph the original phenomenon of study. Regardless, 42 years of patience with DSM-III-type clinical psychosis diagnoses as gold standards has supported little in the way of significant treatment advances for persons with idiopathic psychosis.

It is believed that medical disciplines must abandon exclusive reliance on clinical definitions and incorporate extra-clinical data like laboratory tests to start cleaving individual diseases from a heterogeneous clinical stew (REF). Psychosis diagnostics have not advanced along such lines. This is surprising considering the DSM program for idiopathic psychosis is inconsistent with the available data across genetic, molecular, brain structure or function, physiological, clinical, and outcome research platforms (REFS). Indeed, a comprehensive clinical and biomarker research program demonstrated (n>700 psychosis cases) and replicated (n>700 psychosis cases) that DSM psychosis diagnoses fail to capture neurobiologically distinct entities (REFS). Across most measures, psychosis diagnoses describe a neurobiological continuum (schizophrenia<schizoaffective disorder<bipolar disorder<healthy persons), with considerable group overlap.

So current clinical psychosis diagnoses fail to approximate distinct neurobiological entities with pathology-specific treatment targets (Hyman). As an alternative, the Bipolar-Schizophrenia Network for Intermediate Phenotypes (B-SNIP; REF) searched for neurobiological homology within idiopathic psychosis cases, regardless of specific DSM diagnosis (REF). B-SNIP used variance across multiple psychosis-relevant laboratory measures to detect neurobiologically similar cases via numerical taxonomy. B-SNIP identified, replicated, cross-validated, and externally validated three transdiagnostic subgroups called psychosis Biotypes (REF; see Table 1). Biotype-1 and Biotype-2 share marked cognitive deficits, with Biotype-2 being marginally worse. Biotype-1's defining feature is weak neural response across multiple neurophysiological measures. Biotype-2's defining feature is excessive intrinsic neural brain activity. Biotype-3 cases are close to normal across cognitive, psychomotor, and neurophysiological functions despite having a DSM psychosis diagnosis. Biotypes do not fall along a simple unitary continuum of neurobiological severity.

Stratification by neurobiology may facilitate the search for specific etiology and improve treatment targeting (REF), although this possibility must be proven not preached. Developing precision therapeutics based on clinical features alone, though, is difficult because multiple causes can yield the same clinical features. But clinical evaluation through observation and interview is always the first step in medical diagnosis, with clinical presentation advising the selection of laboratory tests. Even though B-SNIP formed psychosis subgroups using laboratory measures, the defining features of psychosis Biotypes may still map to characteristic signs and symptoms. The present paper evaluates the hypotheses that: (i) neurobiologically homologous psychosis subgroups have clinical features that enlighten their differential identification, and (ii) those features do not simply recapitulate clinical characteristics that differentiate DSM psychosis diagnoses. To the extent clinical information differentiates B-SNIP Biotypes, clinical scientists in any environment could use those features to identify neurobiological subgroups. Such information would support and amplify etiological investigations and clinical applications of Biotypes compared to the available alternatives.

Method

There were 1903 psychosis cases (see Table 1 for demographic information) with sufficient clinical and laboratory data for inclusion in this project. B-SNIP recruitment sites were in Athens GA (University of Georgia), Baltimore MD (Maryland Psychiatric Research Center), Boston MA (Beth Israel Deaconess Medical Center), Chicago IL (University of Illinois-Chicago and University of Chicago), Dallas TX (UT Southwestern Medical Center), Detroit MI (Wayne State University), and Hartford CT (Institute of Living). All subject recruitment, interviews, and laboratory data collection were completed at those locations. Cases were drawn from academic and community mental health centers, small towns with large universities, large cities, inner cities, rural regions, affluent and less affluent areas. B-SNIP recruited a research sample; nonetheless, the large study numbers and broad geographical recruitment foster generalizability of the outcomes across the range of midcourse idiopathic psychosis.

Clinical Evaluations

Full B-SNIP recruitment details and approaches are available in Tamminga et al. (REF). Briefly, clinically stable outpatient participants were administered the Structured Clinical Interview for the Diagnostic and Statistical Manual of Mental Disorders IV (DSM-IV-TR; REF). Psychosis cases were limited to schizophrenia, schizoaffective disorder, and bipolar disorder with psychosis because these are the idiopathic diagnoses with the highest prevalence in most settings, and studying more subgroups was deemed unfeasible. Cases were rated on the Birchwood Social Functioning (REF), Montgomery-Asberg Depression Rating (REF), Positive and Negative Syndrome (REF), and Young Mania Rating (REF) scales. All participants were also rated on the Hollingshead Two-Factor Socioeconomic Rating Scale (REF). Table 1 shows the clinical information by subgroup. Supplementary Table 1 provides a summary of medication information. In previous publications, we demonstrated that medication effects cannot account for group differences on biomarker features (REFS).

The extensive clinical information on each participant was reviewed in a best-estimate diagnostic meeting with at least two experienced research clinicians to establish the consensus diagnosis. Cross-site diagnostic conference calls were carried out monthly; they were chaired by two senior primary investigators and attended by the 2–4 trained clinical assessors at each site. From study start, there were face-to-face and virtual training sessions for all raters, with a requirement for reliability above 0.85. Each month, diagnostic conferences were held with in-depth diagnostic discussions. Each year, rater training was repeated to reestablish reliability. See Tamminga et al. (REF) for complete details.

Biotype Evaluations

Participants completed comprehensive laboratory evaluations within a few weeks of their clinical assessments. Papers on the individual laboratory tasks provide extensive data collection and analysis details (REFS). Details of biomarker quantification and biotyping procedures are in Clementz et al. (2022), including an extensive supplementary method.

The seven laboratory measures used for Biotypes creation were the Brief Assessment of Cognition in Schizophrenia (BACS; REFS), Stop-Signal Task (SST; REFS), pro- and anti-saccade tasks (saccades; REFS), auditory paired stimuli and oddball tasks (ERPs; REFS), and the 9-10 second inter-pair interval of the paired stimuli task (intrinsic EEG activity or IEA; REF). Within each laboratory measurement class (BACS, SST, saccades, ERP, IEA), principal component analysis reduced multiple variables within that class to an efficient and smaller variable set. This was done for two main reasons. First, for estimating a person’s true value on any theoretical construct, multiple independent measures are better than any single variable. Cognitive ability and personality tests, for instance, use many questions to estimate the trait of interest. Likewise, for example, a person’s neural response to stimulus salience is better estimated by many ERP measures than by a single voltage from a single sensor at one time point. Second, reducing the redundancy of measurements increases the accuracy of numerical taxonomy (REF). The outcome of this data reduction process was integrated variables called bio-factors. These bio-factors were used in numerical taxonomy (REF). Every subcomponent of the above procedure replicated in two independent samples, each sample of >700 psychosis cases and >200 healthy participants. The numerical taxonomy outcome cross-validated (REF), and multiple other comparisons supported the neurobiological validity of B-SNIP psychosis Biotypes (REFS). Figure 1 shows effect sizes in relation to main bio-factor classes for the exact same subjects plotted by either their Biotype (1A) or DSM diagnosis (1B). Figure 2 shows the distribution of cases within a DSM class that have each Biotype (left plot) or cases within a Biotype that have each DSM diagnosis (right plot).

Statistical Methods

B-SNIP Biotypes are multivariate biomarker-derived subgroups among mostly persons with midcourse idiopathic psychosis, although there is a range of cases from 15-62 years of age at all stages of illness. The psychosis cases have in common clinical indication of reality distortion (delusions, hallucinations, or disorders of thought). For creating neurobiological subgroups, however, B-SNIP used laboratory measures at a middle level of analysis in the causal chain (between more molecular and more clinical levels). Despite that starting point, this project investigates the hypotheses that B-SNIP Biotypes have distinctive clinical characteristics that do not recapitulate DSM psychosis diagnoses. Clinical features came from the 7 subscales of the Birchwood, 10 items of the MADRS, 30 items of the PANSS, and 11 items of the YMRS, for a total of 58 clinical features (see Supplementary Table 2).

For the current analyses, an important distinction is the constraint on the number of features that may be used per sample. The model can be built from a relatively large set of items (i.e., the 58 clinical features). But to reduce the burden of patient and clinician, the model can be restricted to a smaller number of items, while still maintaining high classification accuracy. By using decision trees or ensembles of decision trees (i.e., a random forest), the number of estimators (trees) in each sample and the maximum depth (i.e., number of items, or features, assessed) of each tree can be constrained.

Constraints will incur a performance cost. It is known that random forests and their related architectures produce significant improvement over classical decision trees (Breiman, 2001). In our application, we built an ensemble with a limited number of estimators, each of which have a constrained depth. The result is a weighted combination of the component estimators, using weights optimized from the whole sample and the all 58-clinical feature calibration data (for example see Gibbons et al., 2013).

As demonstrated previously (Gibbons et.al. 2021, 2022) the “extra-trees” algorithm maximizes area-under-the-curve (AUC) for the constraint described above. The Extra-Trees method (extremely randomized trees, Geurts et al. 2006) drops the random forest idea of using bootstrap copies of the learning sample, and instead selects a cut-point at random, which leads to increased accuracy (due to smoothing) and decreased computational burden. To provide out-of-sample validation, we create an empirical distribution for AUC, where each individual replicate is generated from a random split of the dataset into training and validation subsets. This approach is similar to “leave-p out cross-validation,” except we specify a dataset fraction instead of a fixed p (validation fraction = 0.5, training fraction = 0.5). The process is repeated until there is no significant change in the AUC distribution under a Kolmogorov-Smirnov (KS) test. Since this split can be modeled as a random draw from the complete dataset, confidence bounds for AUC are computed directly from percentiles of the empirical distribution.

In the present context, we have three groups, BT1, BT2, and BT3, which is a multi-class identification problem with three Biotype classes. We used the 58 clinical features at the beginning to develop a low-burden adaptive classifier that can be used by non-B-SNIP sites to derive a Biotype classification based on clinical information alone. We computed a one-vs-all extremely randomized ensemble of decision trees for each of the three Biotypes. While the out-sample AUCs obtained for each individual Biotype may be high, the end-to-end performance for the three-Biotype problem is typically lower. In constructing these trees, we enforced a maximum depth of 10, with 2 distinct estimators in our “forest”, implying that at most 20 features were used to decide the Biotype of any given case. The average number of features used for the overall classification of a case into one of the three Biotypes typically will be much smaller than the maximum of 20 features allowed. The outcome of this procedure, which we call Algorithmic Diagnostics for Efficient Prescription of Treatments (ADEPT) is described below.

Results

Overall Accuracy as a Function of Items

First, we evaluated overall classification accuracy of the model as a function of the number of items included in the three Biotype classification problem. Figure 3 shows mean accuracy (with 99% confidence intervals) for classifying a case into one of the three Biotypes as a function of the number of items used (the 58 clinical features). Overall accuracy of correctly assigning an individual case to their Biotype membership peaks at .765 with 32 or more items included in the classification decision.

Accuracy of Individual Biotype Assignment

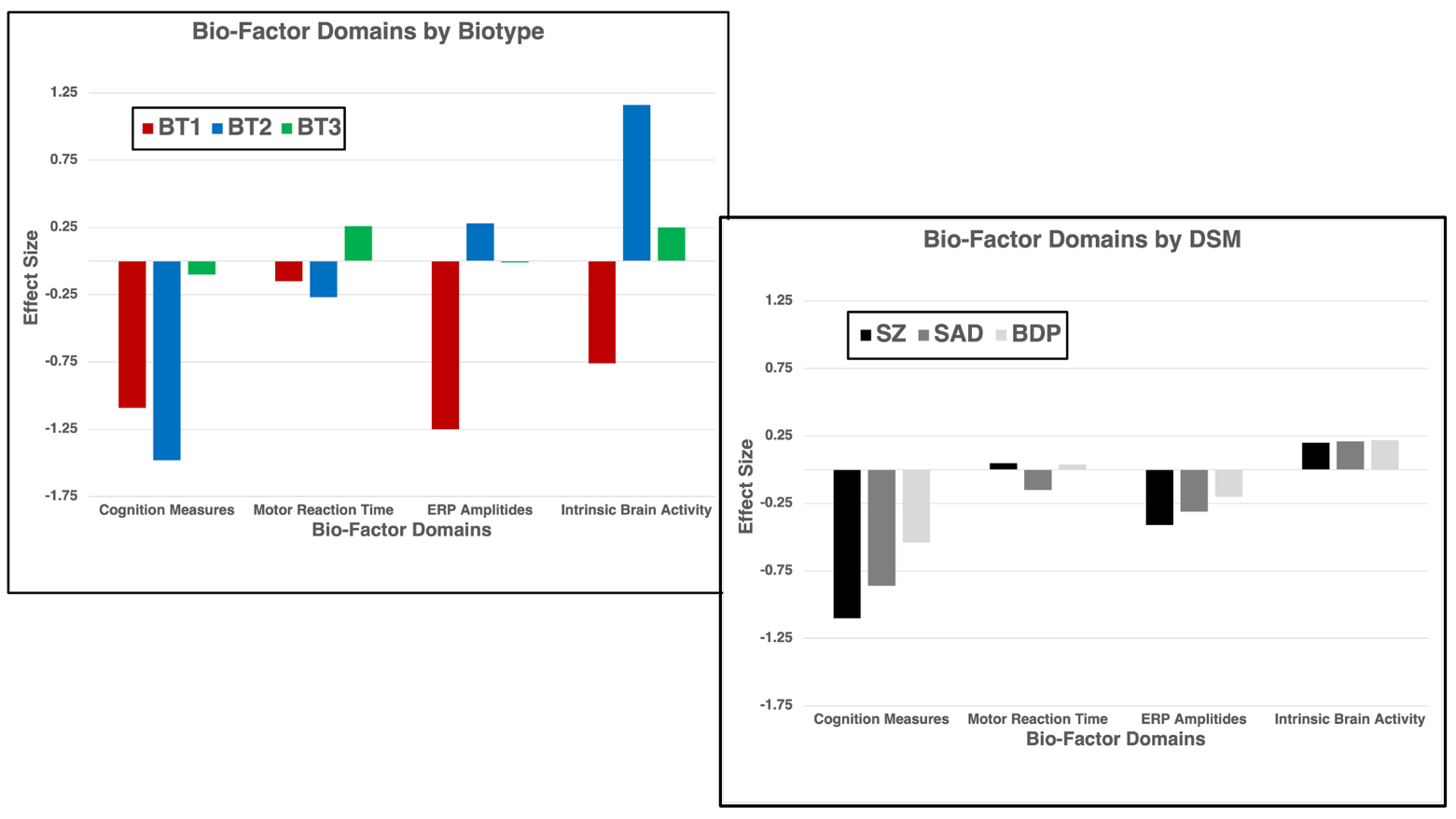
Next, we evaluated the one-vs-all classification success (e.g., BT1 vs BT2 or BT3) extremely randomized ensemble of decision trees for each of the three Biotypes. In constructing these trees, we enforced a maximum depth of 10, with 2 distinct estimators in our “forest”, implying that at most 20 features were used to decide the Biotype of any given case. The mean number of items needed to obtain optimal classification accuracy was XX (SD) for BT1, XX (SD) for BT2, and XX (SD) for BT3. The sensitivity, specificity, and AUC obtained for each Biotype in out-of-sample runs is shown in Figure 4. The classification of one Biotype versus all others has high sensitivity and specificity as illustrated in the ROC curves (and 99% confidence intervals), with corresponding AUCs also uniformly high (>0.85).

Pipeline for the Overall Classification Problem

The average number of features used for the overall classification of a patient into one of the three biotypes was 12 (SD) (i.e., adaptive administration of 12 items out of the 58 items used to derive the original biotypes). In our out-of-sample tests, we obtain a median end-to-end accuracy of 61.5%.

Application to Classification into DSM Subgroups

Discussion

Figure 1

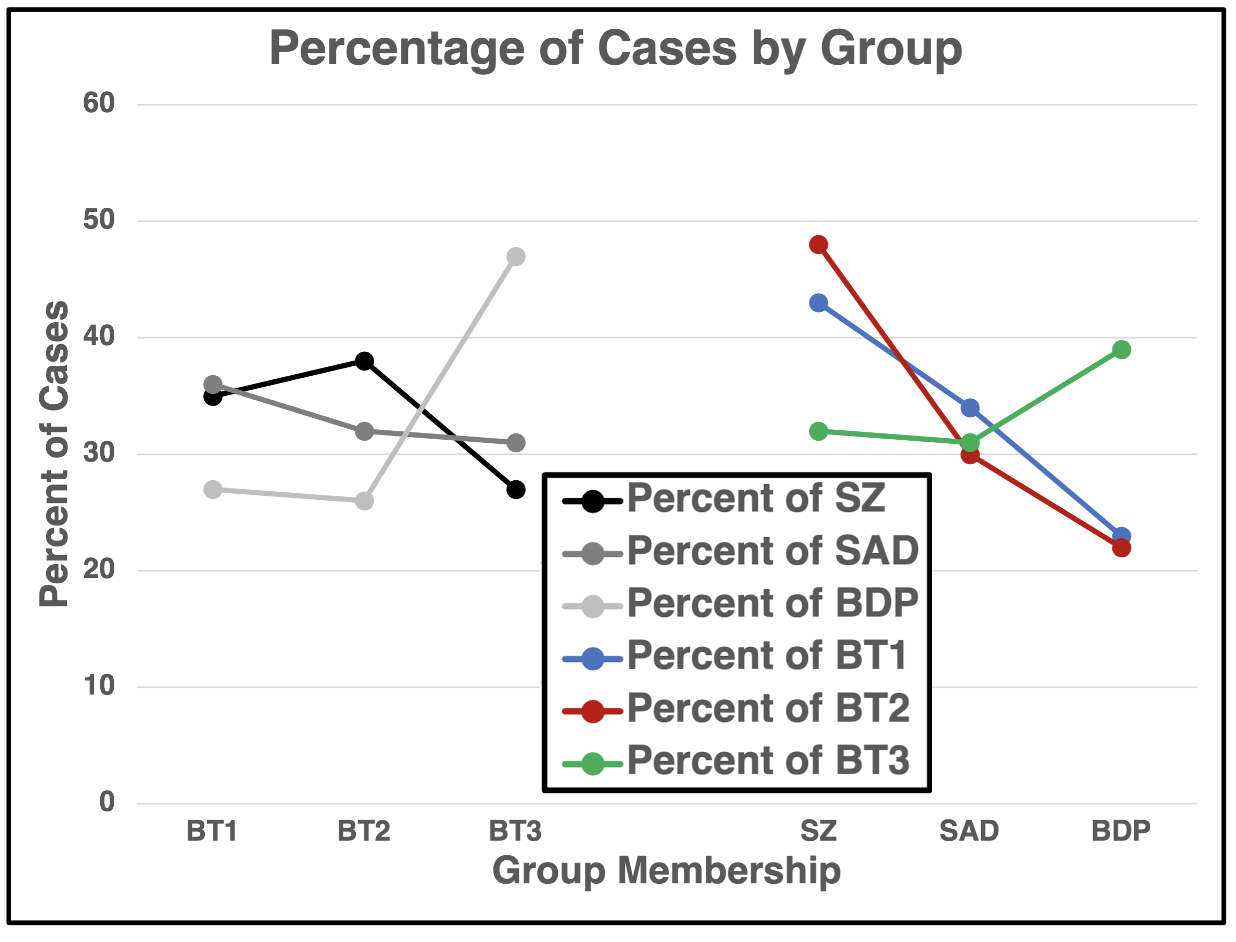
Figure 2

Figure 3

Chart, line chart

Description automatically generated

Chart, line chart

Description automatically generatedFigure 4

Chart

Description automatically generatedChart, scatter chart

Description automatically generated