Enterprise Biology Software: XV. Research (2014)

ROBERT P. BOLENDER

Enterprise Biology Software Project, P. O. Box 292, Medina, WA 98039-0292, USA

http://enterprisebiology.com

SUMMARY

Clinical diagnosis follows largely a subjective approach wherein signs and symptoms play a major role in identifying a disease. In disorders of the brain, however, these signs and symptoms often overlap and diagnosis can become a daunting task. Since complexity theory allows us to capture the rules and algorithms the brain uses to define itself in health and disease, it should be possible to assemble an objective model for diagnosis. The advantage of a data-driven approach is that we can design it as a complexity parallel to that of biology, one that encapsulates both the diagnostic skills and published data of experts. Using MRI data coming from thousands of patients reported in 117 papers (Internet Brain Volume Database (IBVD): Kennedy et al., 2012), we will figure out how to diagnose 27 different disorders of the brain correctly 100% of the time. In turn, this objective approach to diagnosis will trigger several unexpected outcomes - the genesis of a new gold standard, a shift from small data to big, and a better understanding of how to solve complex problems with large data sets. To make our move into big data, we will combine the properties of a spreadsheet with those of two database platforms, upgrade to a 64-bit operating system, and assemble algorithms (Appendix II). A proposed solution to the diagnosis problem will emerge from a series of tests applied to databases containing triplet (AX:BY:CZ) and quadruplet (AX:BY:CZ:DQ) markers. These tests will guide our solution to a complex problem by identifying database filters one after another, as we gather clues along the way. The results will show that this approach to problem solving offers not only a general solution to the problem of clinical diagnosis, but it also provides algorithms and new strategies for automation. In effect, by demonstrating an ability to diagnose phenotypes objectively, we now have the where with all to figure out - at any given point in time - what we are, were, or will be. We can do this because mathematical markers allow us to generalize biological data sets, which contain vast stores of diagnostic and predictive information. The current software package includes new and updated databases along with instructions for their use.

INTRODUCTION

A central challenge in transitioning from a descriptive to an evidence-based health care system includes the introduction of new technologies and theory structures that can deliver phenotypes with diagnostic and predictive properties. To this end, recent reports of the Enterprise Biology Software Project have described methods for translating published data into mathematical markers that can capture the complexity of phenotypes as quantitative patterns, which, in turn, can generalize both locally and globally (Bolender, 2011-2013). Such an approach seeks to identify an enterprise biology capa-

ble of operating seamlessly across all the basic and clinical sciences.

Our task in this report will be to explore the properties of mathematical markers as a diagnostic tool. To this end, we will use data from the Internet Brain Volume Database (IBVD) (Kennedy et al., 2012) to generate markers for known disorders of the human brain and then figure out how to use them to identify unknown disorders.

Since we know that the specificity of a mathematical marker increases in proportion to the number of variables therein, we will increase our level of play by making the transition from triplets (AX:BY:CZ), to quadruplets (AX:BY:CZ:DQ). This upgrade, however, requires a technological shift from small to big data, which introduces a new set of problems. Increased memory requirements compel us to move from a 32-bit to a 64-bit computing platform and to introduce a new suite of software tools. Moreover, we have to learn how to work with database tables containing rows of data numbering in the millions and to automate many of the diagnostic procedures previously done manually.

This is what to expect. At first, most of the diagnostic tests with quadruplet and triplet markers will fail because disorders of the brain share many of the same markers and, as a result, the markers of one disorder influence the diagnosis of another. Moreover, an unknown marker may or may not have a counterpart (duplicate) in the known group. These observations suggest that we need to pursue a strategy based on unique markers that can be shown to work flawlessly within a well-defined data space – the Internet Brain Volume Database. In effect, our goal becomes a software product with algorithms that can deliver the correct diagnosis 100% of the time.

Why set the bar at 100%? When biology triggers an algorithm to produce a specific disorder, it reconfigures it parts and connections according to a new set of instructions. Consequently, each disorder displays mathematical markers – and collections thereof – unique to the disorder. We can tap directly into the algorithm defining a given disorder by selecting only those markers that deliver the correct diagnosis 100% of the time. There is an additional benefit to this approach. We can argue that by applying this procedure to individual patients and to populations thereof, we can approach a general solution to the problem of diagnosis in clinical medicine. In turn, this general solution becomes universal in that it applies not only to MRI data, but also to all the data types that can form mathematical markers. This creates a world of possibility. A universal data set serves not only as a diagnostic tool but also becomes the foundation of a predictive vehicle capable of moving forward and backward in time. Such mobility may become especially helpful as we begin to search for the many and elusive connections that exist between phenotypes and genotypes.

As the methods and results section will show, finding markers that work at the 100% level requires extensive data processing with two 64-bit programs – Microsoft Excel and Access. Reconciling large data sets with these two interacting programs offers a host of new challenges for the reader that may become somewhat less intimidating by working through the examples given in Appendix II. Since big data are fundamental to complexity theory, learning how to diagnosis an unknown disorder with a data-driven approach offers the reader a demanding but rewarding experience.

A key question, however, remains unanswered. Will mathematical markers allow us to diagnose disorders of the brain – one patient at a time? The answer of course can only come from the authors with access to the original patient data in the IBVD.

METHODS AND RESULTS

The software package for 2014/2015 includes new databases and software tools for diagnosing disorders of the brain – using mathematical markers derived from published data. In addition, templates and worked examples will help to ease the transition from small to big data.

Enterprise Biology Software Package

The software includes eight screens offering ready access to programs, databases, and documents (Figure 1).

Figure 1. Enterprise Biology Software Package – 2014/15. The 4GB package contains 354 files stored in 9 folders.

The Game Plan

The report explores new strategies for diagnosing disorders of the brain using big data. They include increasing the specificity of mathematical markers by increasing the number of variables in play, identifying software synergies, and demonstrating the effectiveness of mathematical markers in solving difficult problems. In all cases, the IBVD will serve as our primary source of data (Figure 2).

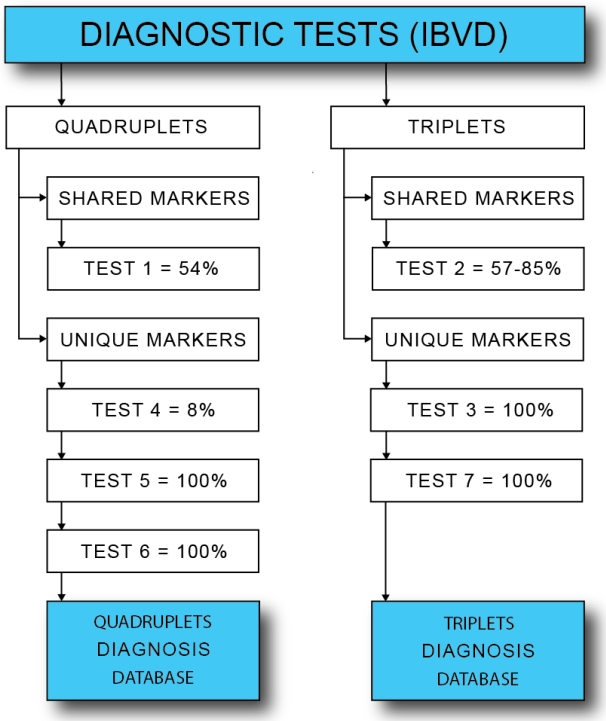


Figure 2. Using clinical data from the Internet Brain Volume Database (IBVD) of Kennedy et al., 2012, we will assemble mathematical markers and use them to diagnose disorders of the brain objectively. A diagnosis, which consists of comparing unknown markers to known standards, will use markers based on 8 (quadruplets) and 6 (triplets) variables. Tests 1 and 2 compare unknown markers to known markers (shared), whereas tests 3 to 7 compare unknown markers to known markers (unique). The tests will help us to design databases capable of diagnosing disorders of the brain – in a well-defined data set (IBVD) - with an accuracy of 100%. Such an outcome requires an approach consistent with the big data requirement of complexity theory. Note that running these tests required operations involving more than 15,000,000 mathematical markers.

Mathematical Markers

A mathematical marker includes parts (names) and connections (ratios) arranged as an alphanumeric string. It serves as a fundamental unit of biological complexity, according to the definitions developed for complexity theory (Bolender, 2012-2013; Appendix III). Moreover, each marker encapsulates the expertise of the physicians who collected the data and made the diagnosis. Although it may take but a few markers to identify an unknown disorder correctly, its characterization by thousands or millions of markers reveals its wide reaching effects.

Recall that mathematical markers increase the amount of information in a publication by permutating an original data set (Bolender, 2001-2013). By forming all possible data ratios, we maximize the likelihood of detecting differences between normal and abnormal data sets. As shown in Figure 3, forming mathematical markers quickly turns little numbers into big ones.

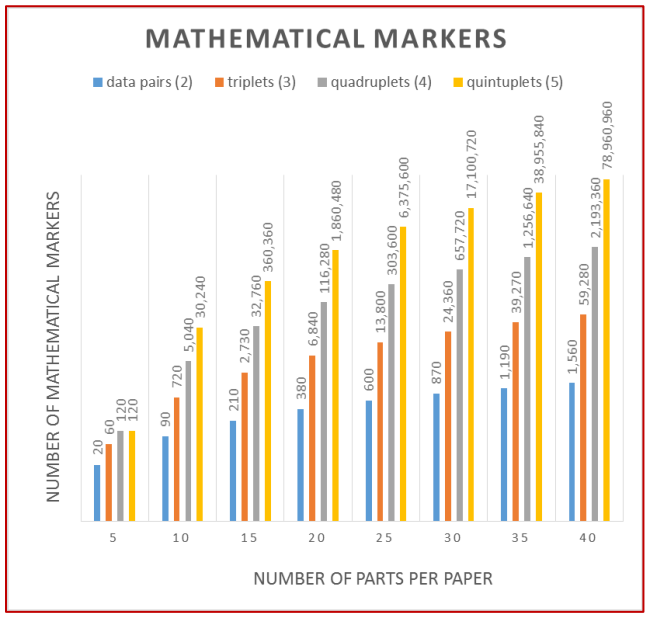


Figure 3. Original data sets increase the amount of information they contain by forming permutations. When expressed as mathematical markers these permutations define phenotypes both qualitatively (alpha string) and quantitatively (numeric string). Notice, for example, that the same 20 parts can produce 380 data pairs, 6,840 triplets, 116,280 quadruplets, and 1,860,480 quintuplets. Each data set represents the phenotype as a set of patterns with a different degree of specificity.

Our immediate task will be to extract information from a large clinical data set (IBVD) with the goal of replicating the clinical diagnoses of the original studies. By connecting the expertise of the physician in diagnosing a disorder with its phenotypic expression, we can begin to explore the properties of a data-driven approach to diagnostics in clinical medicine. Notice the deliberate shift in strategy. Instead of dispersing this expertise across largely inaccessible journals, the IVDB allows us to concentrate it with algorithms designed to solve a real-world problem. Specifically, we want to diagnose disorders of the brain by analyzing volume data derived from MRI head scans.

Quadruplet Markers

A quadruplet marker includes four named parts (A, B, C, D) each with an accompanying numerical value (X, Y, Z, Q). This defines the relationship of one part to another as a mathematical ratio (AX:BY:CZ:DQ). By dividing each numerical value by the value of X, X becomes equal to one (1:Y:Z:Q).

In moving from triplets to quadruplets, however, we move from small data to big. As shown in Figure 3, quadruplets quickly exceed the limits imposed by 32-bit Excel spreadsheets (2 GB of memory and ≤1,048,576 rows of data). Since working with mathematical markers includes shuttling data back and forth between spreadsheets and databases, a 32-bit technology allows us to operate comfortably with triplets, but not with quadruplets.

Moreover, working with quadruplet markers introduces a new set of problems. Recall that diagnosis - as practiced with mathematical markers - depends on matching unknown markers to known standards and tallying the results (Bolender, 2011-2013). This is done by adding an unknown set of markers to a table of known markers in a diagnosis database, sorting the markers alphabetically, and then scanning down the table and marking each duplicate marker (unknown=known standard) as it appears. Alternatively, we can transfer the database table to an Excel spreadsheet and identify the duplicates automatically using the conditional formatting option. In practice, however, sorting a large data set automatically can take hours.

These technology related problems quickly disappeared by shifting to a 64-bit platform and running Excel and Access together as a team. Appendix II includes worked examples to show how this was accomplished.

Diagnostic Tests

A diagnostic test depends on identifying a set of properties unique to a given disorder. Mathematical markers allow us to detect such properties in individual markers and in combinations thereof. Alt-

hough increasing the number of variables in a marker increases its power to distinguish one disorder from another, the cost of adding variables increases the workload exponentially (Figure 3).

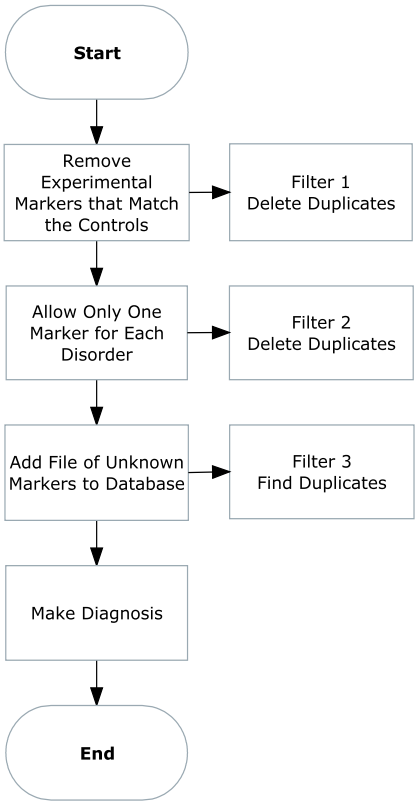
A diagnosis begins with an original set of mathematical markers that requires filtering. Some of the markers are unique (occur one or more times in one setting), whereas others are shared (occur one or more times in multiple settings). Since each category alone or in combination can produce a different outcome, each test shows us how a given outcome depends on our selection of filters. By applying a battery of tests, each result supplies clues, which, when taken together, guide us toward a solution. The filtering process becomes somewhat easier to follow by summarizing each test visually with an algorithm.

Test 1: Quadruplets (Shared Markers)

The first test posed the following question. By upgrading the markers from triplets to quadruplets, will this improve the ability of the database to diagnose unknown disorders?

Making Known Markers: We begin with the volume data of a given paper in the IBVD and use the names thereof to generate quadruplets - using the permutation function in Mathematica. Next, we import this list of quadruplets into an Excel worksheet as a text file (tab delimited) and use a template worksheet (Template_Quads.xlsx) to associate each part (name) with its numerical value (volume). After calculating ratios, we assign the decimal repertoire values as defined in the documents section of the software package (Forms: Worksheet - Connection Phenotype). The template worksheet performs all the concatenations and calculations automatically, thereby producing a table of quadruplet markers. This procedure is applied separately to control and experimental data sets – paper by paper. See Appendix II for a worked example.

Test 1 used the following algorithm.



Algorithm 1. Test 1.

The filters removed duplicate markers from the same paper (control marker = experimental marker) and from the database (experimental marker = experimental marker) so that a given marker could appear only once for a given disorder (Algorithm 1). This defined a diagnosis database for quadruplet markers, which was stored as a text file (Test1.txt).

Making Unknown Markers: To test the effectiveness of this diagnosis database, unknowns were prepared – one paper at a time - using the template mentioned above (Unknown-Test1.txt). Since these data came from patients that carry both normal and abnormal markers, a false positive will occur whenever a normal marker in the unknown corresponds to an abnormal marker in the diagnosis database (control (unknown) = experimental (known)). Although this uncertainty may always exist, it can be minimized (see Test 7).

Diagnosing Unknown Markers: The diagnostic procedure consisted of importing the database text file

(test1.txt) into an Access database, appending a text file containing the markers of a test paper (unknown), looking for matches (unknown = known), and tallying the results. The diagnosis went to the disorder with the largest number of identified unknowns. Obviously, the problem with this approach is that it is subject to a sampling bias. Since the same marker can appear in different disorders, the markers of one disorder can overwhelm those of another. In turn, this can lead to an incorrect diagnosis. The point of test 1 was to see if the quadruplet markers with their increased specificity could overcome the risk of this potential sampling bias.

Results: The MRI papers from the IBVD supplied about 12,000,000 quadruplet markers for the control and experimental data sets. Eliminating duplicates (control = experimental) at the level of individual papers reduced this number to 4,796,416, and finally to 589,945 after deleting duplicates (experimental = experimental for a given disorder). This produced a diagnosis database for the known quadruplet markers. Note that the filters assured that a given mathematical marker could occur only once for a given disorder, but that the same maker could occur in different disorders (Algorithm 1).

Since a quadruplet marker contains four parts (names) with four connections (ratios), the extent to which they were shared across such a wide range of disorders was quite unexpected. Figure 4, for example, shows that the quadruplet database contained 2,538 markers for ADHD (red), but that ADHD shared its markers with at least 12 other disorders (blue). Appendix I includes similar histograms for 21 different disorders.

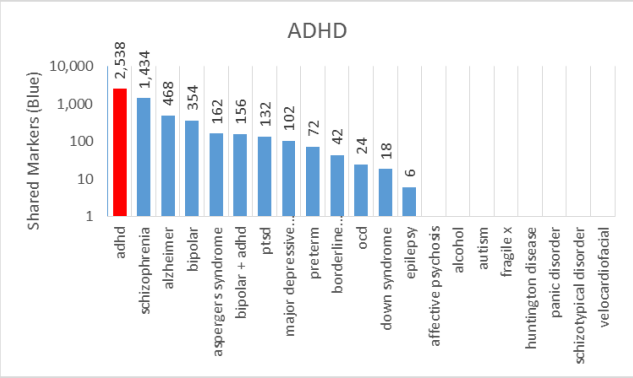


Figure 4. The diagnostic database contains 2,538 quadruplet markers for ADHD of which 1,434 also occurred in schizophrenia, 468 in Alzheimer, etc.

Figure 5 gives the frequency distribution of the quadruplet markers – by disorder - for the database used in Test 1. Notice that the markers range in number from 24 to 245,621 and that schizophrenia and bipolar disorder account for 96% of the markers.

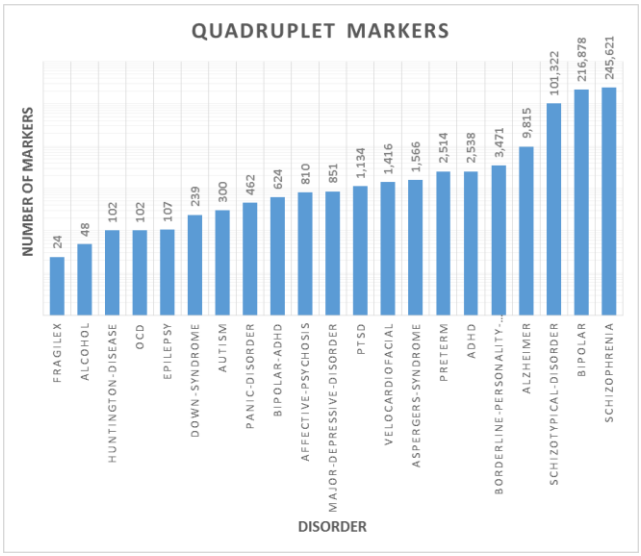


Figure 5. The histogram illustrates the frequency distribution of quadruplet markers in the diagnosis database across 21 disorders of the brain. Notice that most of the markers belong to schizophrenia and bipolar disorder.

To test the effectiveness of this first database as a diagnostic tool, data from thirteen IBVD papers were translated into quadruplet markers (unknowns) and run – one by one - against the knowns of the diagnosis database summarized in Figure 5. The results appear in Table 1.

Table 1. The table includes the results of running the data of 13 unknowns (publications) – one at a time - against a collection of known standards, both of which came from the IBVD. A result can be correct (YES), incorrect (NO), tied (TIE), or nonexistent (variables not in play). In spite of the more than 500,000 known markers being in play, the diagnosis was correct only about 50% of the time. Clearly, the correct diagnosis was frequently being overwhelmed by the data of other disorders (308, 329, 472. 587, 621, and 623). In two cases (308 and 472), the disorder being diagnosed failed to show even a single marker – the unknown variables were not in play. Notice that 5 of the 7 correct diagnoses came from the two disorders with the largest number of markers - schizophrenia (4) and bipolar (1).

DIAGNOSIS OF UNKNOWN QUADRUPLET MARKERS - DIAGNOSIS DATABASE (DDB-2A-1) - TEST 1														
PAPER ID (IBVD)	UNKNOWNS →	126	154	329	308	472	555	587	591	621	623	635	639	657
NUMBER OF PARTS IN PLAY	→	12	10	9	7	6	22	7	7	7	18	7	13	9
DISORDER	DUPLICATE MARKERS ↓	UNKNOWN MARKERS IDENTIFIED (GREEN=DIAGNOSIS)												
ADHD	2,538	0	0	0	24	0	0	0	6	0	0	0	14	0
AFFECTIVE-PSYCHOSIS	810	0	0	0	0	0	0	0	0	0	0	0	0	0
ALCOHOL	48	0	0	0	0	0	0	0	0	0	0	0	0	0
ALZHEIMER	9,815	0	0	6	6	0	23	69	6	6	234	30	70	0
ASPERGERS-SYNDROME	1,566	54	0	0	0	0	20	7	0	0	0	0	14	0
AUTISM	300	0	0	0	0	0	0	0	0	0	0	0	0	0
BIPOLAR	216,878	0	6	18	24	0	2	102	12	54	150	18	17	265
BIPOLAR-ADHD	624	0	0	0	0	0	0	0	0	0	0	0	0	0
BORDERLINE-PERSONALITY-DISORDE	3,471	0	12	12	0	12	14	0	6	12	0	72	6	0
DOWN-SYNDROME	239	0	0	0	0	0	0	0	66	150	0	30	0	0
EPILEPSY	107	0	0	24	0	30	0	0	66	24	0	0	0	18
FRAGILEX	24	0	0	0	0	0	0	0	0	0	0	0	0	0
HUNTINGTON-DISEASE	102	0	0	0	0	0	0	0	0	0	0	0	0	0
MAJOR-DEPRESSIVE-DISORDER	851	0	0	0	0	0	8	6	0	0	0	36	0	12
OCD	102	0	0	0	0	0	0	0	0	0	0	0	0	0
PANIC-DISORDER	462	0	0	0	0	0	0	129	0	2	150	0	0	0
PRETERM	2,514	0	0	0	20	0	18	0	12	21	30	40	41	0
PTSD	1,134	0	0	0	2	0	8	78	0	5	24	5	0	0
SCHIZOPHRENIA	245,621	108	12	0	0	0	62	66	0	18	54	0	87	204
SCHIZOTYPICAL-DISORDER	101,322	54	0	0	0	0	0	1	0	0	0	0	0	0
VELOCARDIOFACIAL	1,416	0	0	52	0	96	0	0	36	78	0	0	0	24
TOTAL MARKERS	589,944													
DIAGNOSIS	54% CORRECT	YES	YES	NO	NO	NO	YES	NO	YES	NO	NO	YES	YES	YES
DIAGNOSIS	45% CORRECT W/0 TIES		TIE						TIE					

In test 1, the diagnosis was correct only 54% of the time (Table 1). The results, however, indicated that several disorders masked the correct diagnosis (velocardiofacial, panic disorder, Down syndrome, and Alzheimer), but not schizophrenia. Although the increased specificity of the quadruplet markers played a role (e.g., no masking by schizophrenia), the number of parts not in play seemed to be a major limiting factor. In effect, the test (knowns vs. unknowns) was comparing incompatible samples. If this is the case, increasing the amount and mix of data in the diagnosis database (knowns) might produce a better result. Test 2 was designed to test this possibility.

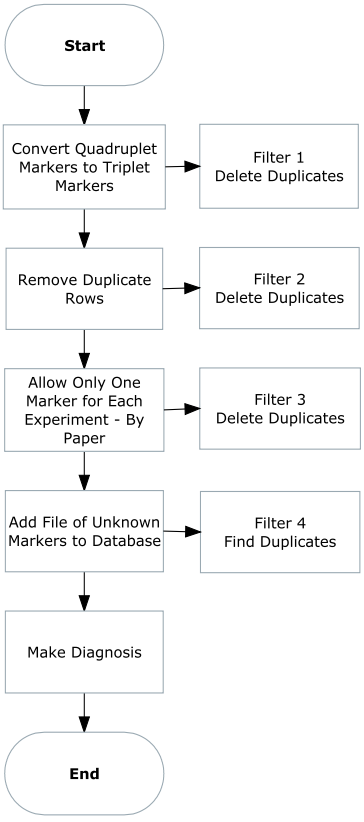
Test 2: Triplets (Shared Markers)

Test 2 consisted of downsizing the quadruplet markers of test 1 to triplets (AX:BY:CZ) and increasing the number of IBVD papers contributing markers to both the known and unknown data sets (Algorithm 2). Notice in Table 2, however, that test 2 failed at about the same level as test 1 - the diagnosis was correct only 57% of the time. Moreover, the triplet markers displayed a substantial loss of specificity, as shown by the strong masking effect by schizophrenia. If we remove this masking effect, the success of the test jumps to 86%.

Taken together, the results of tests 1 and 2 tell us that a database containing shared markers shows little promise as a diagnostic tool. Accordingly, test 3 used only unique markers.

Table 2. The diagnosis database of Test 2 included the same collection of parts used in Test 1, but this time they were used to generate triplet markers. The table shows a strong masking effect by schizophrenia, which led to a diagnostic score of only 57%. When the offending schizophrenia data were removed from the analysis, the score increased to 86%.

DIAGNOSIS OF UNKNOWN TRIPLET MARKERS - (T-DDB-3A) - TEST 2																								
PAPER ID (IBVD)	UNKNOWNS →	50	126	132	329	358	482	485	536	555	587	590	591	621	626	635	639A	639	657	667	724	777
NUMBER OF PARTS IN PLAY	→	27	12	9	9	8	7	9	11	22	7	6	7	7	10	7	13	13	9	12	7	7
DISORDER	MARKERS IN DB ↓	UNKNOWN MARKERS IDENTIFIED BY DISORDER (GREEN=DIAGNOSIS)																				
ADHD	249	29	0	0	0	0	2	0	0	0	0	0	2	0	0	0	7	2	2	0	0	0
AFFECTIVE-PSYCHOSIS	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ALCOHOL	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ALZHEIMER	4,391	26	0	0	1	0	0	0	0	46	7	3	1	0	2	0	15	12	1	0	0	0
ASPERGERS-SYNDROME	221	12	0	0	0	0	2	0	0	7	0	0	0	0	0	0	0	0	0	0	0	0
AUTISM	113	0	0	4	0	0	0	0	5	0	0	0	0	0	0	0	0	0	0	0	2	0
BIPOLAR	17,529	23	1	4	10	1	21	0	0	8	15	3	4	10	7	1	0	4	27	0	0	0
BIPOLAR-ADHD	87	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
BORDERLINE-PERSONALITY-DISORDER	271	12	1	0	1	1	1	0	0	8	1	0	1	1	3	1	6	0	1	0	0	0
DEVELOPMENTAL-DELAY	124	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
DOWN-SYNDROME	15	0	0	0	2	0	4	0	0	0	0	0	0	15	2	0	0	0	0	0	0	0
EPILEPSY	18	0	0	0	7	0	6	0	0	0	0	0	17	8	2	0	0	0	7	0	0	0
FRAGILEX	67	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
HUNTINGTON-DISEASE	236	0	0	4	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	2	0
MAJOR-DEPRESSIVE-DISORDER	3,381	11	0	0	0	0	2	3	0	14	0	0	0	0	0	0	6	3	0	0	0	0
OCD	30	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
PANIC-DISORDER	40	8	0	0	1	0	0	0	0	0	13	1	1	0	0	0	0	0	1	0	0	0
PRETERM	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
PTSD	103	20	0	0	0	0	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0
SCHIZOPHRENIA	99,166	446	169	4	10	17	26	25	1	174	46	12	16	20	49	20	85	72	93	7	4	1
SCHIZOTYPICAL-DISORDER	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
VELOCARDIOFACIAL	155	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
WILLIAMS-SYNDROME	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
TOTAL MARKERS	126,212																					
DIAGNOSIS	57% CORRECT	NO	YES	YES	NO	YES	NO	NO	YES	YES	YES	NO	YES	YES	NO	NO	NO	YES	YES	YES	NO	YES
DIAGNOSIS	56% CORRECT W/0 TIES			TIE					TIE			TIE										
MASKING BY SCHIZOPHRENIA	→	X			X			X				X			X	X	X				X	
DIAGNOSIS W/O SCHIZOPHRENIA	86% CORRECT	YES	YES	YES	NO	YES	NO	YES	YES	YES	YES	YES	YES	YES	YES	YES	NO	YES	YES	YES	YES	YES
	83% CORRECT W/O TIES			TIE					TIE			TIE										



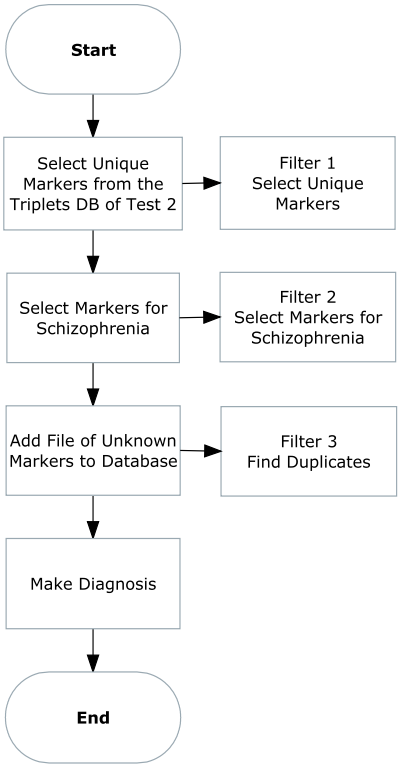
Algorithm 2. Test 2.

Test 3: Triplets (Unique Markers)

Recall that schizophrenia has two types of markers, those that it shares with other disorders and those unique to schizophrenia. When we select only the unique markers (the ones that occur only once) from the diagnosis database of triplets described in Test 2, we find 83,305 for schizophrenia, 2 for Alzheimer, 2 for bipolar, and 1 for major disruptive disorder (Algorithm 3). If, in turn, we run several unknowns against this new set of unique markers (knowns), the effectiveness of the diagnostic method jumps to 100% (Table 3).

Table 3. When the unknown markers were run against the database of markers unique to schizophrenia, each unknown was diagnosed correctly.

DIAGNOSIS OF UNKNOWN TRIPLET MARKERS - (T-DDB-3B) - TEST 3 MARKERS UNIQUE TO SCHIZOPHRENIA													
PAPER ID (IBVD)	UNKNOWNS →	126	154	358	555	587	621	623	639	657	667	669	777
NUMBER OF PARTS IN PLAY	→	12	10	8	22	7	7	18	13	9	12	9	7
DISORDER	MARKERS IN DB ↓	UNKNOWN MARKERS IDENTIFIED AS SCHIZOPHRENIA											
SCHIZOPHRENIA MARKERS	83,305	141	375	14	144	34	8	39	27	64	7	5	1
DIAGNOSIS	100% CORRECT	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES



Algorithm 3. Test 3.

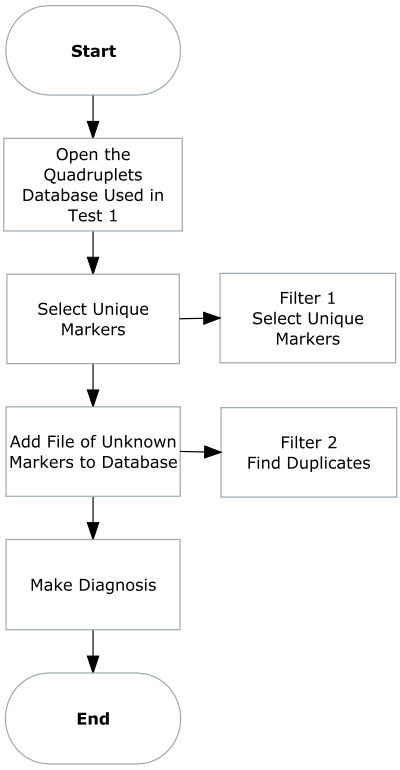
Test 4: Quadruplets (Unique Markers)

In Test 4, we return to the quadruplets and select only the unique markers - those that appear only once in the database (Algorithm 4). When several unknowns were run against the unique knowns of this database, the promising results seen in Table 3 failed to appear. Table 4 shows that the database of unique markers had a success rate of only 8%. Moreover, the markers of four papers (308, 587, 621, and 657) were not even in play. Although the markers were unique in the quadruplet database of knowns, they were not unique in the unknowns because the same marker occurred in more than one disorder. In effect, the unknowns were sharing similar markers.

Normally, such a result would signal failure and bring the testing to an abrupt close. Complexity theory, however, changes the rules. It assures us that every problem in biology has a solution provided we set it up correctly. We simply need to rethink our approach.

Table 4. The database of unique quadruplet markers was not effective as a diagnostic tool because it was sharing its unique markers with more than one of the unknown disorders. In effect, the known markers were unique to the knowns but not to the unknowns.

DIAGNOSIS OF UNKNOWN QUADRUPLET MARKERS - DIAGNOSIS DATABASE (QUADS-UNIQUE¹) - TEST 4													
PAPER ID (IBVD)	UNKNOWNS →	126	154	308	329	472	555	587	591	621	623	635	639	657
NUMBER OF PARTS IN PLAY	→	12	10	7	9	6	22	7	7	7	18	7	13	9
DISORDER	UNIQUE MARKERS ↓	UNKNOWN MARKERS IDENTIFIED (GREEN=DIAGNOSIS)												
ADHD	240,304	42	0	30	0	0	22	0	6	0	0	0	132	0
AFFECTIVE-PSYCHOSIS	1,008	0	0	0	0	0	0	0	0	0	0	0	0	0
ALCOHOL	1,212	0	0	0	0	0	0	0	0	0	0	0	0	0
ALZHEIMER	587,743	0	12	0	0	0	442	210	0	0	318	6	30	0
ASPERGERS-SYNDROME	337,914	42	0	0	0	0	22	0	0	0	0	0	0	0
AUTISM	10,797	0	0	0	0	0	12	0	0	0	0	0	0	0
BIPOLAR	770,306	12	24	48	192	0	40	90	6	54	102	12	312	0
BIPOLAR-ADHD	450	0	0	0	0	0	0	0	0	0	0	0	0	0
BORDERLINE-PERSONALITY-DISORDER	50,679	0	0	0	0	0	402	0	0	0	0	18	0	0
DOWN-SYNDROME	559	0	0	54	12	18	0	0	42	208	0	0	0	0
EPILEPSY	433	0	0	42	12	12	0	0	204	12	0	0	0	0
FRAGILEX	6,372	0	0	0	0	0	0	0	0	0	0	0	0	0
HUNTINGTON-DISEASE	22,410	0	0	0	0	0	0	0	0	0	0	0	0	0
MAJOR-DEPRESSIVE-DISORDER	44,325	0	6	0	0	0	16	0	0	0	0	6	0	0
OCD	38,685	0	0	0	0	0	0	0	0	0	0	0	0	0
PANIC-DISORDER	15,006	0	0	0	0	0	0	36	0	0	24	0	0	0
PRETERM	49,836	0	0	0	24	18	50	0	0	18	0	0	18	0
PTSD	64,362	12	0	0	0	0	826	0	0	0	24	6	0	0
SCHIZOPHRENIA	1,189,381	624	6	0	0	0	194	0	6	0	114	24	36	0
SCHIZOTYPICAL-DISORDER	0	0	0	0	0	0	0	0	0	0	0	0	0	0
VELOCARDIOFACIAL	73,446	0	0	0	0	132	30	0	18	18	0	42	0	0
TOTAL MARKERS	3,505,228	732	48	174	240	180	2056	336	282	310	582	114	528	0
DIAGNOSIS	8% CORRECT	YES	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO



Algorithm 4. Test 4.

What do we know so far? We know that a diagnosis succeeds when the all the markers (known and unknown) are unique (Table 3), but fails when one or both of the markers are shared (Tables 1, 2, and 4). Tests 1 and 2 failed because they used just shared markers. Test 3 succeeded because both the known and unknown markers were unique. Test 4 failed because one set of markers was unique (knowns), but the other set was shared (unknowns).

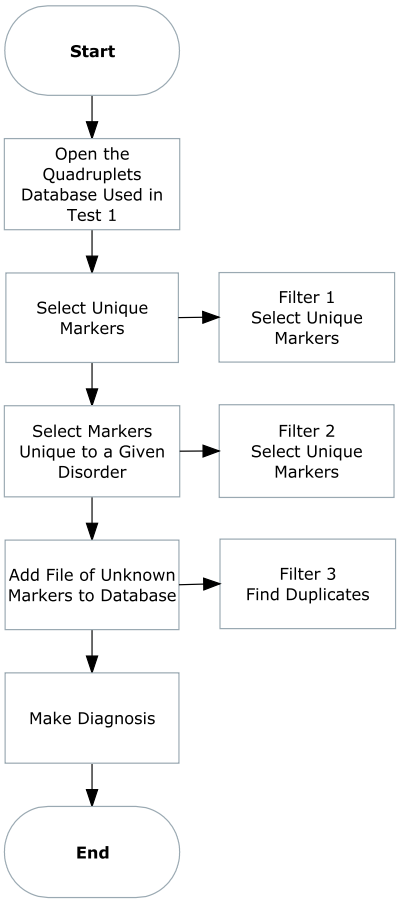
These results tell us what to do. We need to apply a set of filters that prevent or minimize sharing within - but not between - the known and unknown markers. In effect, we can diagnose a disorder of the brain by matching unknown to known markers, provided such markers are unique to the individual known and unknown data sets. In Test 5, we take the next step by applying a set of filters that improved the success of the diagnosis from 8% (Table 4) to 80% (Table 5).

Test 5: Quadruplets (Unique Markers)

Test 4 used one unique filter, whereas Test 5 used two (Algorithm 5). The first filter of Test 5 selected for unique markers, whereas the second filter selected for markers unique to a given disorder – paper by paper. The resulting markers served as the knowns in the diagnosis database used for Test 5. Table 5 indicates that this filtering algorithm leads to a better outcome, given the score of 80%. Notice that three of the unknowns (472, 587, and 657) were out of play (OOP) in that Filter 3 found no duplicates. Moreover, the unknown markers of papers 308 and 621 led to the incorrect diagnosis of epilepsy and that the correct diagnosis was out of play, as indicated by the absence of duplicates (0). This tells us that we may have reduced, but not eliminated the sampling and data compatibility issues.

Table 5. By increasing the uniqueness of the markers, we increase their ability to diagnose disorders correctly. Removing papers that are out of play (OOP) improved the results.

DIAGNOSIS OF UNKNOWN QUADRUPLET MARKERS - DIAGNOSIS DATABASE (QUADS-UNIQUE²'³) - TEST 5													
PAPER ID (IBVD)	UNKNOWNS →	126	154	308	329	472	555	587	591	621	623	635	639	657
NUMBER OF PARTS IN PLAY	→	12	10	7	9	6	22	7	7	7	18	7	13	9
DISORDER	UNIQUE MARKERS ↓	UNKNOWN MARKERS IDENTIFIED (GREEN=DIAGNOSIS)												
ADHD	0	0	0	0	0	0	0	0	0	0	0	0	0	0
AFFECTIVE-PSYCHOSIS	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ALCOHOL	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ALZHEIMER	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ASPERGERS-SYNDROME	0	0	0	0	0	0	0	0	0	0	0	0	0	0
AUTISM	0	0	0	0	0	0	0	0	0	0	0	0	0	0
BIPOLAR	312	0	0	0	0	0	0	0	0	0	0	0	312	0
BIPOLAR-ADHD	0	0	0	0	0	0	0	0	0	0	0	0	0	0
BORDERLINE-PERSONALITY-DISORDER	18	0	0	0	0	0	0	0	0	0	0	18	0	0
DOWN-SYNDROME	0	0	0	0	0	0	0	0	0	0	0	0	0	0
EPILEPSY	204	0	0	36	12	0	0	0	204	12	0	0	0	0
FRAGILEX	0	0	0	0	0	0	0	0	0	0	0	0	0	0
HUNTINGTON-DISEASE	0	0	0	0	0	0	0	0	0	0	0	0	0	0
MAJOR-DEPRESSIVE-DISORDER	0	0	0	0	0	0	0	0	0	0	0	0	0	0
OCD	0	0	0	0	0	0	0	0	0	0	0	0	0	0
PANIC-DISORDER	0	0	0	0	0	0	0	0	0	0	0	0	0	0
PRETERM	0	0	0	0	0	0	0	0	0	0	0	0	0	0
PTSD	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SCHIZOPHRENIA	932	625	6	0	0	0	194	0	0	0	114	0	4	0
SCHIZOTYPICAL-DISORDER	0	0	0	0	0	0	0	0	0	0	0	0	0	0
VELOCARDIOFACIAL	0	0	0	0	0	0	0	0	0	0	0	0	0	0
TOTAL MARKERS	0	0	0	0	0	0	0	0	0	0	0	0	0	0
DIAGNOSIS 1 (QUADS-UNIQUE²)	80% CORRECT	YES	YES	NO	YES	OOP	YES	OOP	YES	NO	YES	YES	YES	OOP
DIAGNOSIS 2 (QUADS-UNIQUE³)	100% CORRECT	YES	YES	OOP	YES	OOP	YES	OOP	YES	OOP	YES	YES	YES	OOP



Algorithm 5. Test 5.

Test 6: Quadruplets (Unique Markers)

To get to a diagnostic score of 100%, we must deal with the issues of sampling and compatibility. Such issues can be eliminated by moving to a closed system wherein all the markers are unique and each marker can assume the role of either a known or unknown. We create a closed system by including only those markers coming from the IBVD. In effect, we identify two worlds, one filled with knowns (MRI data stored in the IBVD) and the other with unknowns (MRI data not stored in the IBVD). One world exists as a certainty (100%), the other as an uncertainty (?%).

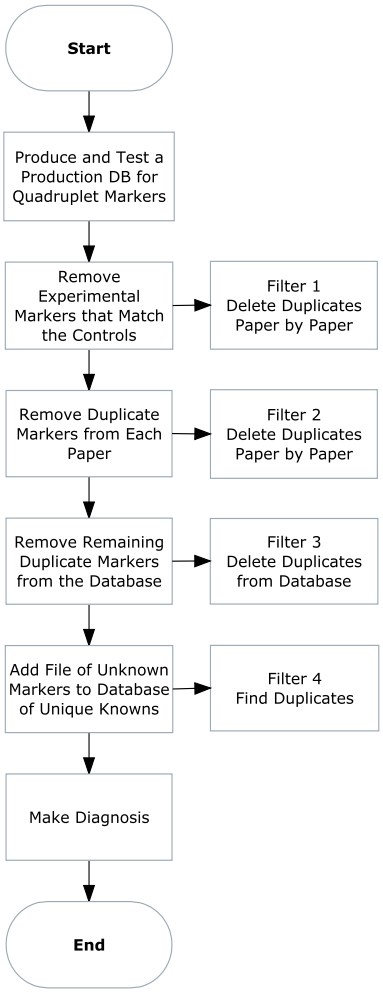
When a marker from the unknown world enters the known world, we can predict its diagnosis with a given probability – that can be determined empirically. The preliminary results of Test 3, however, suggest that this prediction can be correct – at least in this one example – 100% of the time.

Diagnosis Database (Quadruplets): Since the results of the tests indicated that only unique mathematical markers could give the correct results 100% of the time, the diagnosis database for quadruplets (MRI_Q_DIAG_100) now contains just such markers (Algorithm 6). Table 6 summarizes the composition of the database, which includes data from 75 papers and 3.6 million unique markers. If a marker is generated from any one of these 75 papers and run against this database, the only possible outcome is a correct diagnosis.

Table 6. With the appropriate filters applied, a quadruplet database of unique markers can diagnose a disorder correctly 100% of the time.

QUADRUPLET DATABASE - UNIQUE MARKERS - 2014			
DISORDER	**MARKERS**	**PAPERS**	**DIAGNOSIS**
ADHD	240,286	3	YES
AFFECTIVE-PSYCHOSIS	1,008	1	YES
ALCOHOL	1,212	1	YES
ALZHEIMER	587,131	2	YES
ASPERGERS	338,064	3	YES
AUTISM	10,797	6	YES
BIPOLAR	771,734	11	YES
BIPOLAR-ADHD	450	1	YES
BORDERLINE PERSONALITY DISORDER	50,417	2	YES
DOWN-SYNDROME	541	1	YES
DYSLEXIA	63	1	YES
EPILEPSY	2,821	2	YES
FRAGILEX	6,372	1	YES
HUNTINGTON-DISEASE	22,410	2	YES
KLINEFELTER-SYNDROME	9,036	1	YES
MAJOR DEPRESSIVE-DISORDER	44,277	3	YES
OCD	38,685	1	YES
PANIC-DISORDER	14,982	1	YES
PRETERM	49,800	1	YES
PTSD	64,153	2	YES
SCHIZOPHRENIA	1,324,205	27	YES
VELOCARDIOFACIAL	73,326	2	YES
QUADRUPLET TOTALS	**3,651,770**	**75**	**100%**

Table 6 offers a gentle wake-up call. If the IBVD is representative of the clinical literature, then three or fewer papers (3=3, 2=6, and 1=10) are representing 86% (19/22) of the disorders. At some point, such small sample sizes will compromise our ability to diagnose and predict.



Algorithm 6. Test 6.

Test 7: Triplets (Unique Markers)

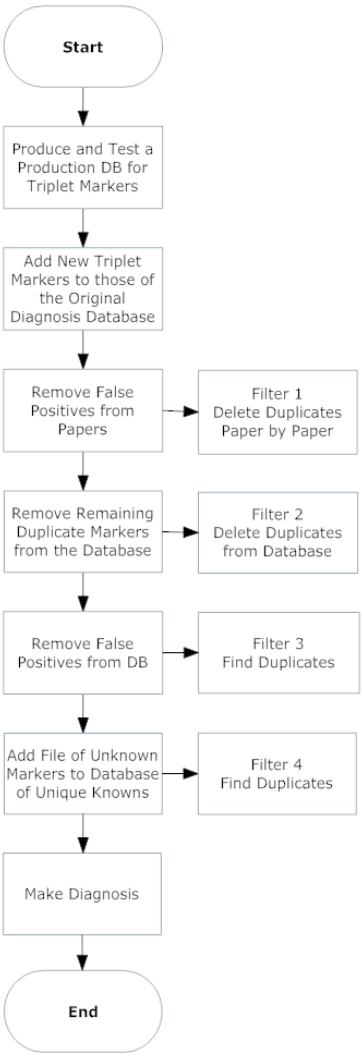
Diagnosis Database (Triplets): Test 7 applies the procedure described for the quadruplet markers of Test 6 to triplet markers (Algorithm 7). Once again, the diagnosis of unknown markers was correct 100% of the time (Table 7).

However, we still need one more filter to minimize the effect of false positives that may occur when we use the diagnosis database to predict a disorder with an unknown set of markers – beyond those of the IBVD. This would include, for example, data coming from an individual patient. Recall that a marker of a disorder becomes a false positive whenever a control marker duplicates it. These duplications occur at two levels - papers and databases. We can remove false positives (C=E) from a given paper by identifying duplicates between normal (C) and abnormal (E) markers. Once a diagnosis database is built, it can be run against the original database of normal markers to delete the remaining false positives (C=E for all papers) in the database. This database filter, for example, removed an additional 31,275 false positives from the MRI-T-Diag-100 database of Test 7. Remember that when working within a complexity, we are always dealing with both local and global issues.

Notice in Tables 6 and 7 that the markers characterizing 22-27 disorders of the brain came from a relatively small number of papers - 75 for quadruplets and 117 for triplets. Given the tools included in the software package, the task of increasing the number of papers in play from hundreds to thousands now becomes a realistic goal.

Table 7. When filtered appropriately, a triplet database of unique markers can diagnose a disorder correctly 100% of the time.

TRIPLET DATABASE - UNIQUE MARKERS - 2014			
DISORDER	**MARKERS**	**PAPERS**	**DIAGNOSIS**
ADHD	27,499	6	YES
AFFECTIVE-PSYCHOSIS	494	2	YES
AGING	120	1	YES
ALCOHOL	984	2	YES
ALZHEIMER	28,568	5	YES
ASPERGERS	16,574	4	YES
AUTISM	2,921	11	YES
BIPOLAR	46,839	16	YES
BIPOLAR-ADHD	34	1	YES
BORDERLINE PERSONALITY DISORDER	2,847	2	YES
DEVELOPMENTAL-DELAY	948	2	YES
DOWN-SYNDROME	63	1	YES
DYSLEXIA	210	1	YES
EPILEPSY	276	2	YES
FRAGILEX	1,502	2	YES
HUNTINGTON-DISEASE	2,692	3	YES
INTRAUTERINE-GROWTH-RESTRICTION	676	1	YES
KLINEFELTER-SYNDROME	1,232	1	YES
MAJOR DEPRESSIVE-DISORDER	4,046	7	YES
OCD	3,091	1	YES
PANIC-DISORDER	1,858	2	YES
PRETERM	4,023	3	YES
PSYCHOPATHIC	938	1	YES
PTSD	5,288	2	YES
SCHIZOPHRENIA	114,878	35	YES
VELOCARDIOFACIAL	7,490	2	YES
WILLIAMS	948	1	YES
TRIPLET TOTALS	**277,039**	**117**	**100%**



Algorithm 7. Test 7.

The Software Package

The software package includes a new MRI database (MRI_2014.db) containing separate databases for quadruplet (MRI-Q-DIAG-100%) and triplet (MRI-T-DIAG-100%, MRI-T-BIG, AND MRI-T-SMALL) markers – all derived from the IBVD. Note that the DVD includes separate copies of the Sybase and Microsoft databases. The templates, databases, and worked examples provided in the software package offer the reader a step-by-step approach to clinical diagnosis, one that operates comfortably within the framework of biological complexity and big data.

DISCUSSION

One of the most delightful consequences of applying complexity theory to biology is that one never knows what will happen until it happens. Surprises, it would appear, become a part of the complexity package. Another, somewhat curious consequence is that every biological problem seems to have a solution. Once constructed, a parallel complexity becomes a mirror into which we can look to see what biology is doing. This means that by simply engaging complexity, we get to enjoy biology as a first rate colleague.

The solution to the diagnosis problem described herein, however, may seem odd at first reading because it follows a different set of rules. Instead of using the signs and symptoms of a disorder to make a diagnosis, it uses unique markers taken from diagnosed patients to define a given disorder as a unique phenotype. Moreover, by applying an appropriate set of filters to a parallel complexity (in this case a diagnosis database), we can be assured that a diagnosis made within this complexity will be correct 100% of the time. Such an outcome occurs because a convergence exists between biology and our parallel complexity (Appendix III).

Bear in mind that the diagnostic method described herein allows us to operate from a position of strength. By encapsulating the expertise of many skilled clinicians into a set of unique markers, the power of that expertise becomes universally available. In effect, the leveraging power of technology becomes enormous.

The Tests

The report put complexity theory to the test by asking a hard question: “Can we develop a data-driven approach to diagnosing disorders of the brain?” It qualifies as a hard question for two reasons. First, the properties of these disorders overlap considerably and second, the playing field shifts from small data to big. Moreover, a solution to one depends on a solution to the other.

The report offers new insight into the mechanism of complex problem solving. To arrive at a solution to our diagnosis problem, we had to apply a series of filtering algorithms to the original databases of mathematical markers. This process, which began with a long string of failures, identified – incrementally - the filters needed to improve the diagnostic outcome. We can take this exhaustive approach to problem solving because our shift to big data introduced extensive automation. This automation came largely from the synergies that developed between Excel spreadsheets and Access databases. However, given the restrictions imposed by these programs related to memory, clipboard capacity, and number of rows per table, automating the procedures was itself a challenge (Appendix II).

Shared Markers: All the tests based on shared (duplicate) markers received failing grades. The assumption that the diagnosis goes to the disorder attracting the largest number of markers proved to be incorrect because the duplicate markers of other disorders often occurred more frequently. Although the increased specificity of the quadruplet markers eliminated some of this masking effect (e.g., no masking by schizophrenia in Test 1), it was not enough to overcome masking by other disorders. In short, the results of tests 1 and 2 eliminated shared markers as a reliable diagnostic tool.

Unique Markers: The results of tests 3, 5, 6, and 7 indicated that diagnosing disorders of the brain with mathematical markers required filters with multiple levels of uniqueness. By aggregating uniqueness, we eventually arrived at a filtering algorithm that produced the correct diagnosis 100% of the time (Tests 6 and 7). This result depended on a willingness to embrace a closed system, one that guaranteed the uniqueness of the markers and the consistency of the outcome. When everything is known, the questions become the answers and the answers the questions. Big data played an important role in this unusual approach to problem solving in that it allowed us to filter our way to a solution.

False Positives: A diagnostic procedure increases its reliability by removing distractors from patient data that might otherwise lead to an incorrect result. In a data-driven approach, false positives become a major distractor. In our case, they exist whenever one mathematical marker duplicates another at a place where mischief can result. These places exist in individual papers and in diagnosis databases whenever control markers duplicate experimental (C=E) and in diagnosis databases where different disorders share the same marker (E=E). By eliminating these false positives, a diagnosis can be correct - 100% of the time (Tests 6 and 7).

We know that a patient presenting with a disorder of the brain carries both normal and abnormal markers – in roughly equal proportions. This means that an unknown data set includes markers that could be acting as false positives. Although most of these will not be in play because of the filters applied to the diagnosis database, a residual population of normal markers will continue to exist as false positives. We can eliminate many of these remaining false positives by running the unknown markers against the database of control markers (MRI-T-SMALL) to remove all duplicates – before running it against the diagnosis database (MRI-T-DIAG-100).

The Disease Process

Now that we know how to diagnose disorders of the brain objectively, our attention can shift to the disease process. Since MRI data coming from living patients frequently generalize both locally (within a given paper or lab) and globally (across many papers and many labs), we can read the rules that biology is using to make, remodel, and repair itself.

Notice in Table 8 that quadruplet markers displayed 10 different duplicate sets, whereas triplet markers displayed 63 (Figure 7). The counts of duplicates identify the amount and range of the generalizations. Counts of data sets identify global rules in that biology uses the same combinations of parts and connections repeatedly. This modular arrangement allows us to approach the disease process as a mathematical puzzle. Individual markers, which rep-

resent snippets of larger rules, can be concatenated into networks displaying higher levels of order. Using this approach, we can begin to model disorders of the brain quantitatively using the wide range of data types found in the literature.

Quadruplet Markers: Table 8 and Figure 6 summarize the distribution of duplicate markers – and groups thereof – in normal patients and in those diagnosed with disease. Even though the alphanumeric string of the quadruplet markers contained eight variables, 21% of the markers formed duplicates with 2 to 11 copies each. The shift in the frequency distribution of the groups from 2 to 3 and 4 copies suggests that the disease process increases connectivity. This remodeling event may signal the activation of a shared mechanism for the onset of a disease.

Table 8. The distributions of quadruplet markers suggest that the brain responds to the disease process by increasing connectivity. Markers shifted from 2 copies per group to 3 and 4. Of the 13,360,056 quadruplet markers, 2,802,799 (21%) were duplicates.

Duplicates	Normal		Disease		Normal	Disease
	Total	Groups	Total	Groups		
2	832246	416123	1447108	723554	91.47%	76.45%
3	56139	18713	319635	106545	6.17%	16.89%
4	14616	3654	114984	28746	1.61%	6.07%
5	3660	732	8135	1627	0.40%	0.43%
6	1368	228	1722	287	0.15%	0.09%
7	462	66	840	120	0.05%	0.04%
8	768	96	288	36	0.08%	0.02%
9	432	48	270	30	0.05%	0.01%
10	60	6	0	0	0.01%	0.00%
11	66	6	0	0	0.01%	0.00%

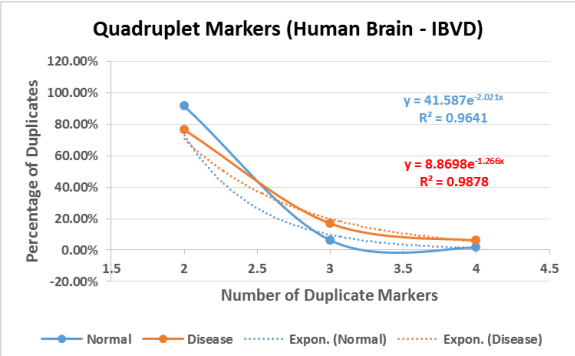


Figure 6. Most of the markers have 2, 3, or 4 duplicates in the normal (99.3%) and disease (99.4%) data sets. Notice, however, that in disease we find an increase in specificity by shifting the distribution of duplicates from 2 to 3 and 4. In both cases, the curves follow an exponential rule. Recall that such a rule was found earlier with ladder equations (Bolender, 2004).

Triplet Markers: The original database of triplets (MRI_T_Small.accdb) included 381,476 duplicate markers, which represented 47.2% of the total (Figure 7). The number of duplicate markers ranged from 2 to 64.

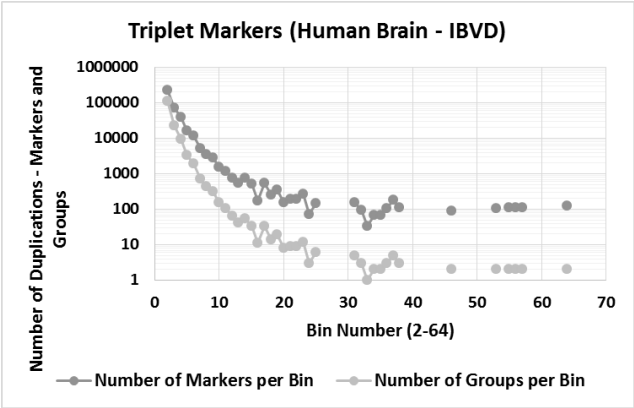


Figure 7. The distribution of triplet markers (C+E) shows duplications ranging from 2 per group to 64.

The data in Figure 7 were collected with an Access database by modifying the SQL script used to select duplicates (right click tab, select SQL View). Figure 8 illustrates the method. In the original script, the query used <…Having Count(*)>1)))> - top panel. The modified script (middle panel) selects only those markers with 10 duplicates <…Having Count(*)=10)))>. The script was run (right click tab, select Datasheet View) to view the results (bottom panel).

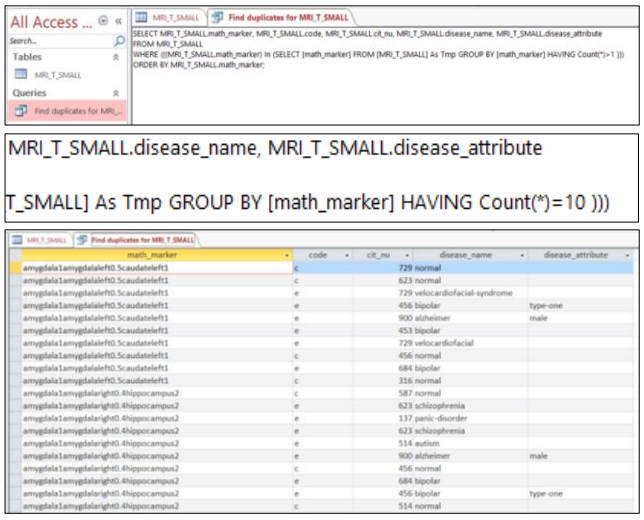


Figure 8. Top: SQL script for finding duplicates. Middle: Modified script to collect markers with a set number of duplicates (10). Bottom: The output table identifies the markers with 10 duplicates (1520 total markers, 152 groups in bin 10).

Figures 8 and 9 illustrate ways in which we can combine the strengths of the Microsoft and Sybase databases - within the framework of complexity theory – to work out the patterns and relationships between brain disorders. The query by example (QBE) frontend to the original triplet database provides ready access to all the data contained therein (Figure 9).

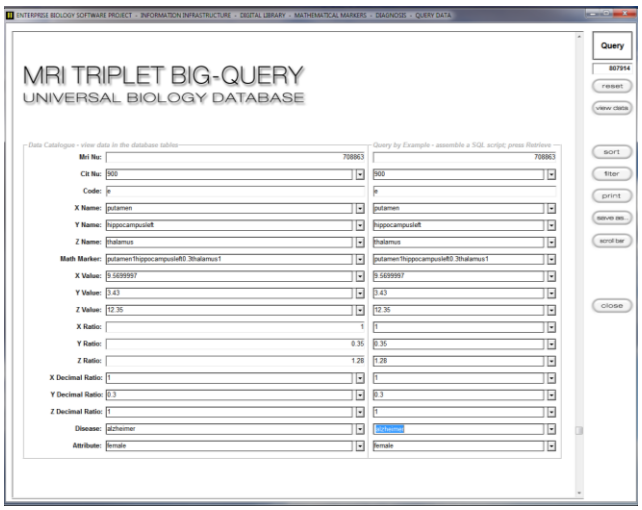


Figure 9. The query by example (QBE) front-end of the triplet database greatly simplifies the task of finding highly specific information - quickly.

Caveats

Online Access to Published Data: Problem solving throughout the biology enterprise will depend - increasingly - on open access to large amounts of data online. This becomes unavoidable as our investigative models for biology shift from simple to complex – as they must. The new diagnostic procedures described herein could not exist, for example, in the absence of the Internet Brain Volume Database. Regrettably, such databases are few in number and often hard to find.

Some facts are indisputable. Complexity is a big data game and in the absence of such data, we cannot become players. The single, greatest threat to our success as a science going forward is the largely unchallenged construction of paywalls around our data. Try to run a literature search on PubMed or Highwire and the severity of this threat becomes obvious. Curiously, the solution to this problem is both simple in design and easy to accomplish. We need to publish our data simultaneously in both paywalled journals and in open access databases. As such, win-lose becomes win-win. In our case, the success of the IBVD as an open access model for publishing data is demonstrated by the fact that we now have a new collection of databases for diagnosing disorders of the brain.

Heterogeneity: Given the eclectic makeup of the diagnosis databases, the results of the tests seem quite remarkable. Data came from patients with different disorders, severities, ages, genders, treatments, and sample sizes – using different methods of data collection and analysis. A more heterogeneous group of patient data is difficult to imagine. In spite of these presumed shortcomings, the mathematical markers were still able to deliver the correct diagnosis. Moreover, the number of duplicate markers one finds by scrolling through the original databases demonstrates the remarkable ability of biology to maintain the stoichiometry of its parts in such diverse settings. This pattern of order persists relentlessly in data pair, triplet, and quadruplet markers. Wherever we look, the same rules remain in play and in plain sight.

Individual Patients: The big unknown remains the diagnosis of a single patient. Since the testing protocol relied exclusively on average patient data, nothing can be said about its application to individual patients. Individual patient data were simply not available to test. All we can do is surmise that the assignment of decimal repertoire values to the ratios of the original data provides enough of a buffer that will work to our advantage with individual patient data. Once again, data access becomes the major limiting factor in resolving such issues.

Opportunities

Complexity theory allows us to unify data across the biology enterprise with mathematical markers, which can standardize and connect most types of published data (Bolender, 2001-2014). By translat-

ing the biology literature into big data, problem solving can become automated and exhaustive.

Clinical Diagnosis: By translating the IBVD into mathematical markers, it can serve as a gold standard for diagnosing disorders of the brain objectively (Figure 10). This approach creates a built in support structure wherein we can rely on data coming from expert investigators to guide the outcome of a diagnosis.

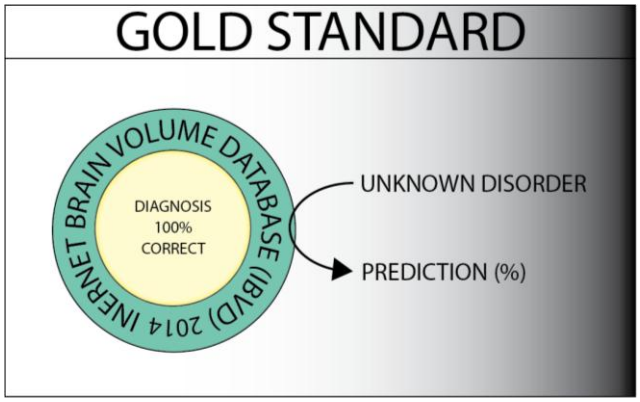


Figure 10. Databases allow us to generate a host of new applications from the biology literature. The Internet Brain Volume Database, for example, currently serves as a gold standard for diagnosing disorders of the brain objectively.

An objective approach to diagnosis becomes a key to opening numerous doors to progress in the life sciences. Recent reports, for example, indicate that disorders, predispositions, treatments, and exposures can leave quantitative tracks throughout an organism, especially in the brain (see, for example, Cecil et al. 2008, Guido et al. 2013, Herting et al. 2014, Khan et al. 2011, Strassburger et al. 1997, and Tiehuis et al. 2008). Such information when combined with technology could spark new industries.

Clinical medicine may soon have its own "iPhone" revolution along with a wave of innovative applications. A handheld device, for example, with an array of sensors might pick up enough information to assemble a diagnostic and predictive phenotype by simply comparing samples to known standards. Since everything in a biological complexity is connected, such outcomes seem quite likely. Once again, the database becomes the solution. In fact, something exciting may soon happen. A startup – called Butterfly Network, Inc. – plans to introduce handheld scanning devices for MRI and ultrasound with built in diagnostics. If these devices provide volume data, then the data set of an individual could be analyzed – at least provisionally - within the existing framework of a diagnosis database (MRI-T-DIAG-100).

Big Data: Technology has reached the point where we can accumulate and analyze very large data sets. Our health care systems, which deal almost exclusively with events occurring at the level of phenotypes, are currently trying to figure out how to use enormous amounts of patient data constructively.

Figure 11 considers the phenotype of an individual over a lifetime of ten decades. A comprehensive set of markers collected, for example, at ten-year intervals will provide a diagnostic set that becomes – retrospectively – perfectly predictive. When collected from a large numbers of individuals, such information provides a global resource that can assign predictions to the diagnosis of an individual phenotype – at any point in time. In effect, the diagnosis of one patient becomes the predictor of another.

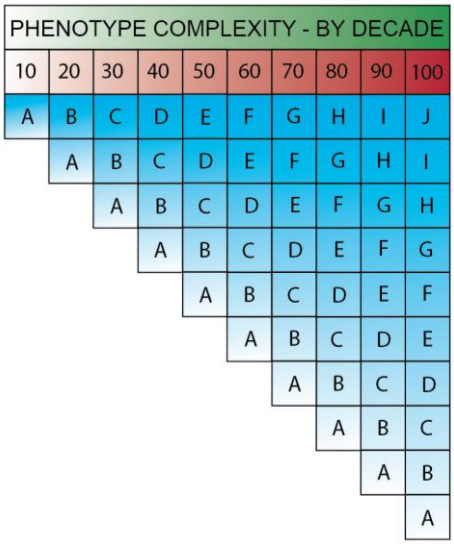


Figure 11. Like trees, phenotypes accumulate a history of our lives that we can record, read, and interpret with mathematical markers. It allows us to evaluate our past and current state and to predict our future.

Abnormal Brain Phenotypes: Disorders change the patterns that define a phenotype, which we can capture with mathematical markers. By unfolding the brain, for example, into its component parts and connections, we quickly discover that many of the same patterns appear across a wide range of different disorders (this report, Bolender, 2012 and 2013).

	SHARED MATHEMATICAL MARKERS				
DISORDER A	A1:B4:C3	D1:E7:F4	G1:H2:I8	J1:K5:L2	X1:Y1:Z4
DISORDER B		D1:E7:F4	G1:H2:I8	J1:K5:L2	X1:Y1:Z4
DISORDER C			G1:H2:I8	J1:K5:L2	X1:Y1:Z4
DISORDER D				J1:K5:L2	X1:Y1:Z4
DISORDER E					X1:Y1:Z4
	1	2	3	4	5

Figure 12. Disorders of the brain can be unfolded into collections of shared and unique markers. Repairing the marker in row 5, for example, might result in widespread and dramatic benefits.

Such an observation triggers new possibilities. If disorders share similar etiologies, then they may also share similar solutions and treatments. This suggests that by shifting our focus from treating symptoms to identifying and repairing the underlying abnormalities, we may end up solving a host of different problems simultaneously (Figure 12). The cost effectiveness of such an approach could be enormous.

The point, which now seems inescapable, is that disorders of the brain involve enormous complexity. Working out the underlying patterns will no doubt require large data sets drawn from a wide range of disciplines. Hunting for such patterns becomes both a compelling and worthwhile adventure because we will be reinventing biology as a quantitative science.

Reality Check

Modern day biology suffers from grievous flaws. In biology, everything connects within and across species, defining vast complexities and interrelationships. In biology as we practice it, little or nothing is connected. Biology runs on complexity, we run on reductionism. Biology uses its data (parts and connections) to generate emergent properties that create dazzling outcomes. We use largely data isolated from biology to look for significant differences that ultimately require the context of complexity to explain and understand. Biology plays by the rules of nature, we play - all too often - by our own rules. Biology already has most of the solutions, whereas we are still trying to figure out how to set up the problems.

We lack a critical understanding. As a product of nature, biology like physics and chemistry is a mathematical discipline. It operates by well-defined and thoroughly tested rules that can be captured mathematically. By allowing biology to develop as a descriptive science, however, we have constructed unwittingly an artificial wall between our common languages of mathematics. Biology is speaking mathematics as it creates the complexities that give rise to emergent properties. In contrast, we often speak with the throttled data of countless methods that may or may not have anything to do with biology (Bolender, 2013).

The promise of mathematical markers as the syntax of a common language derives from the fact that they order data exactly the same way that biology orders its parts – according to stoichiometric rules. When we capture biology quantitatively as a phenotype, these markers can combine to generate parallel complexities capable of producing their own emergent properties. By applying this construct to the biology literature, we can use published data to read biology mathematically as a complexity. As described in this report, our newly acquired ability to diagnose disorders of the brain is an emergent property coming to us from biology by way of the IBVD.

Concluding Comments

Complexity theory allows us to explore the relationship of diagnosis to prediction in biology. Diagnosis defines a phenotype at a given point in time, whereas prediction extrapolates the phenotype in time

into the future or back to the past. Diagnosis is the key. If we do not know what we are at a given point in time, we cannot know what we were or what we are likely to become. In effect, diagnosis and prediction lie at the heart of the biology enterprise.

By making the transition from small to big data, we now have databases capable of diagnosing disorders of the brain with a reliability of 100% - within the boundaries defined by the IBVD. This represents a promising first step.

If we wish to become more effective as a science, we need to move our methods and thinking into the realm of complexity. To do this, we want to become privy to and play by the same rules and algorithms that biology uses to run its business. Then, and only then, can we begin to tackle the truly difficult problems. Few would argue that biology knows many of the most profound secrets in our universe, but even fewer would admit that until we begin to flush them out mathematically, we could never be more than a descriptive science. As we make the inevitable shift to mathematics and complexity, everything will change dramatically – mostly to our advantage.

REFERENCES

Bolender, R. P. 2001a Enterprise Biology Software I. Research (2001) In: Enterprise Biology Software, Version 1.0 © 2001 Robert P. Bolender

Bolender, R. P. 2002 Enterprise Biology Software III. Research (2002) In: Enterprise Biology Software, Version 2.0 © 2002 Robert P. Bolender

Bolender, R. P. 2003 Enterprise Biology Software IV. Research (2003) In: Enterprise Biology Software, Version 3.0 © 2003 Robert P. Bolender

Bolender, R. P. 2004 Enterprise Biology Software V. Research (2004) In: Enterprise Biology Software, Version 4.0 © 2004 Robert P. Bolender

Bolender, R. P. 2005 Enterprise Biology Software VI. Research (2005) In: Enterprise Biology Software, Version 5.0 © 2005 Robert P. Bolender

Bolender, R. P. 2006 Enterprise Biology Software VII. Research (2006) In: Enterprise Biology Software, Version 6.0 © 2006 Robert P. Bolender

Bolender, R. P. 2007 Enterprise Biology Software VIII. Research (2007) In: Enterprise Biology Software, Version 7.0 © 2007 Robert P. Bolender

Bolender, R. P. 2008 Enterprise Biology Software IX. Research (2008) In: Enterprise Biology Software, Version 8.0 © 2008 Robert P. Bolender

Bolender, R. P. 2009 Enterprise Biology Software X. Research (2009) In: Enterprise Biology Software, Version 9.0 © 2009 Robert P. Bolender

Bolender, R. P. 2010 Enterprise Biology Software XI. Research (2010) In: Enterprise Biology Software, Version 10.0 © 2010 Robert P. Bolender

Bolender, R. P. 2011 Enterprise Biology Software XII. Research (2011) In: Enterprise Biology Software, Version 11.0 © 2011 Robert P. Bolender

Bolender, R. P. 2012 Enterprise Biology Software XIII. Research (2012) In: Enterprise Biology Software, Version 12.0 © 2012 Robert P. Bolender

Bolender, R. P. 2013 Enterprise Biology Software XIII. Research (2013) In: Enterprise Biology Software, Version 13.0 © 2013 Robert P. Bolender

Cecil K.M., Brubaker C.J., Adler C.M., Dietrich K.N., Altaye M., Egelhoff J.C. , Wessel S., Elangovan I., Hornung R., Jarvis, K., and B.P. Lanphear . (2008) Decreased Brain Volume in Adults with Childhood Lead Exposure. PLoS Med 5(5): e112. doi:10.1371/journal.pmed.0050112

Guido K.W. Frank G.K.W., Shott, M.E., Hagman J.O., abd T.T. Yang. (2013). Localized Brain Volume and White Matter Integrity Alterations in Adolescent Anorexia Nervosa. Journal of the American Academy of

Child & Adolescent Psychiatry 52: Issue 10, 1066–1075.e5.

Herting M.M., Gautam P., Spielberg J.M., Kan E, Dahl R.E., and E. R. Sowell. (2014) The role of testosterone and estradiol in brain volume changes across adolescence: A longitudinal structural MRI study. Human Brain MappingVolume 35, Issue 11, 5633–5645.

Kennedy D. N., Hodge S. M., Gao Y., Frazier J. A., and C. Haselgrove. (2012) The internet brain volume database: a public resource for storage and retrieval of volumetric data. Neuroinformatics. Apr; 10(2):129-40.

Khan O., Bao F., Shah M., Caon C., Tselis A., Bailey R., Silverman B., Zak I., Resnick S.M., Goldszal A.F., Davatzikos C., Golski S., Kraut M.A., Metter E.J., Bryan R.N., and Zonderman A.B. (2012) Effect of disease-modifying therapies on brain volume in relapsing-remitting multiple sclerosis: results of a five-year brain MRI study. J Neurol Sci. 312(1-2):7-12. doi: 10.1016/j.jns.2011.08.034. Epub 2011 Sep 13.

Strassburger T.L., Lee H.C., Daly E.M., Szczepanik J., Krasuski J.S., Mentis M.J., Salerno J.A., DeCarli C., Schapiro M.B., and G.E. Alexander (1997) Interactive effects of age and hypertension on volumes of brain structures. Stroke. 28(7):1410-7.

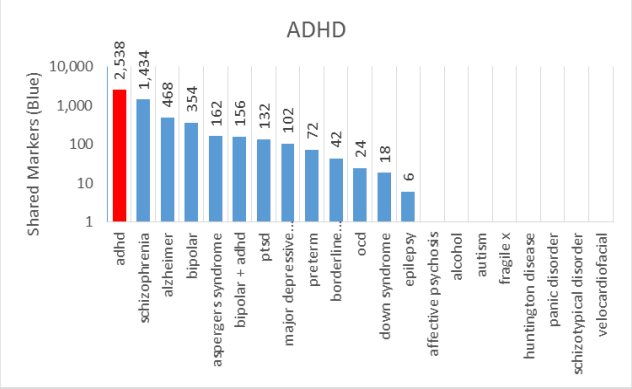
Tiehuis A.M., van der Graaf Y., Visseren F.L., Vincken K.L., Biessels G.J., Appelman A.P., Kappelle L.J., and W.P. Mali (2008) Diabetes increases atrophy and vascular lesions on brain MRI in patients with symptomatic arterial disease. Stroke. 39(5):1600-3. doi: 10.1161/STROKEAHA.107.506089. Epub 2008 Mar 27.

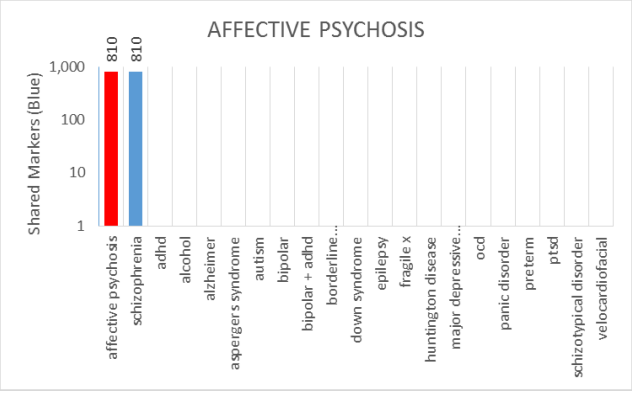
APPENDIX I

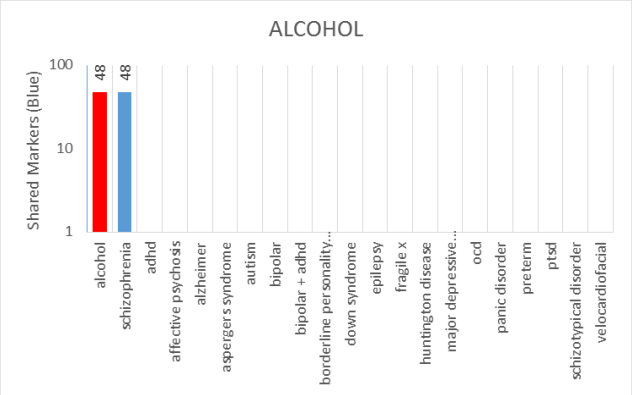
Distributions of Quadruplet Markers

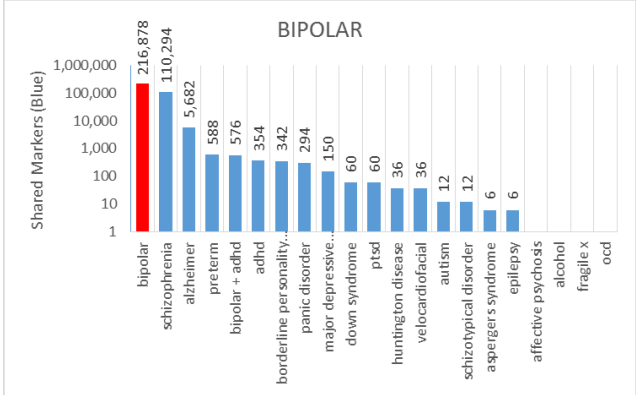
Each figure below indicates that a given disorder carries a distinct set of quadruplet markers (red), most of which appear in other disorders (blue). In effect, this widespread sharing of similar sets of components suggests a modular basis for the formation of disorders in the brain. Using a common pool of parts and connections, the brain appears to be rearranging modules derived therefrom to produce new patterns with new emergent properties.

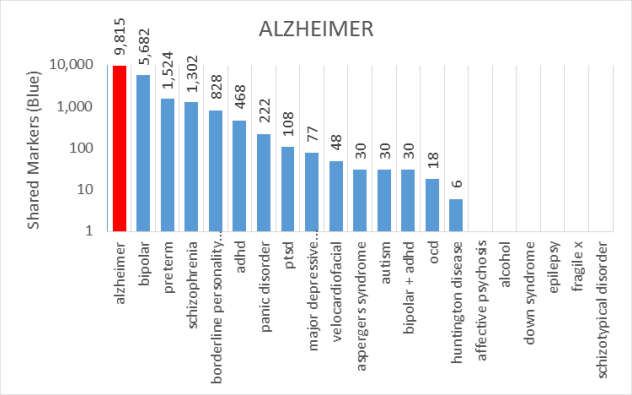
The question yet unanswered deals with the motivation behind these disorders. Is the brain trying out new combinations of modules to become more successful or is it simply responding to mistakes?

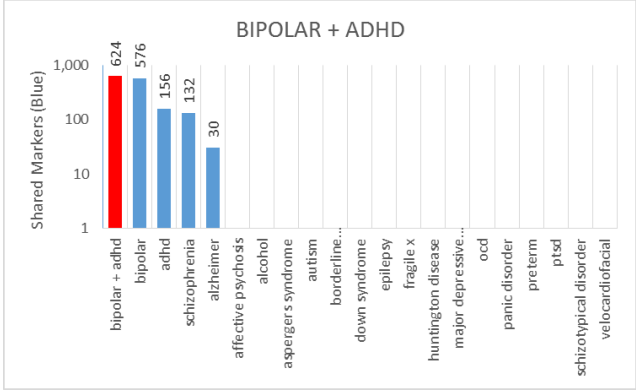


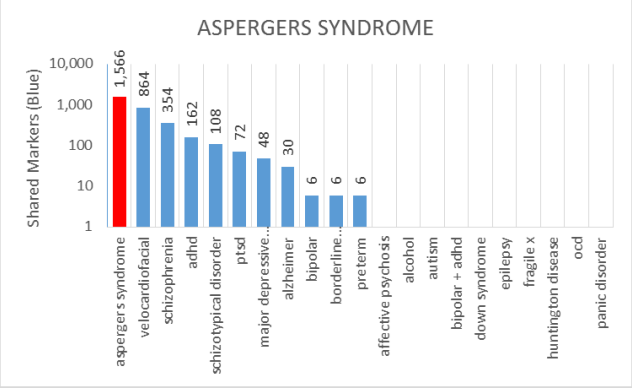


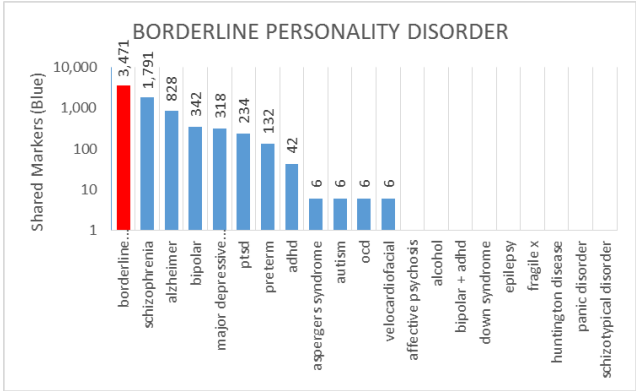

ALCOHOL
Shared Markers (Blue)
100
10
1
48
48
alcohol
schizophrenia
adhd
affective psychosis
alzheimer
aspergers syndrome
autism
bipolar
bipolar + adhd
borderline personality...
down syndrome
epilepsy
fragile x
huntington disease
major depressive...
ocd
panic disorder
preterm
ptsd
schizotypical disorder
velocardiofacial

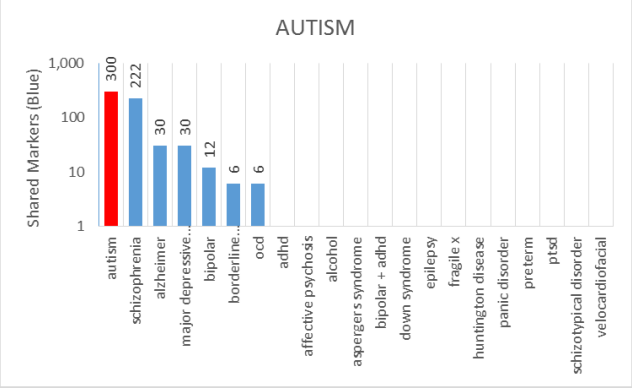

BIPOLAR
Shared Markers (Blue)
1,000,000
100,000
10,000
1,000
100
10
1
216,878
110,294
5,682
588
576
354
342
294
150
60
60
36
36
12
12
6
6
bipolar
schizophrenia
alzheimer
preterm
bipolar + adhd
adhd
borderline personality...
panic disorder
major depressive...
down syndrome
ptsd
huntington disease
velocardiofacial
autism
schizotypical disorder
aspergers syndrome
epilepsy
affective psychosis
alcohol
fragile x
ocd

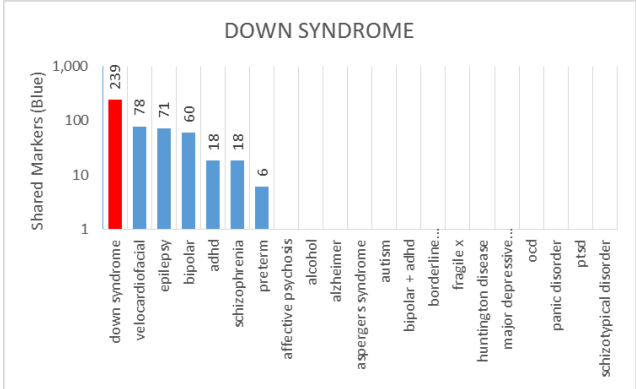

ALZHEIMER
Shared Markers (Blue)
10,000
1,000
100
10
1
9,815
5,682
1,524
1,302
828
468
222
108
77
48
30
30
30
18
6
alzheimer
bipolar
preterm
schizophrenia
borderline personality...
adhd
panic disorder
ptsd
major depressive...
velocardiofacial
aspergers syndrome
autism
bipolar + adhd
ocd
huntington disease
affective psychosis
alcohol
down syndrome
epilepsy
fragile x
schizotypical disorder

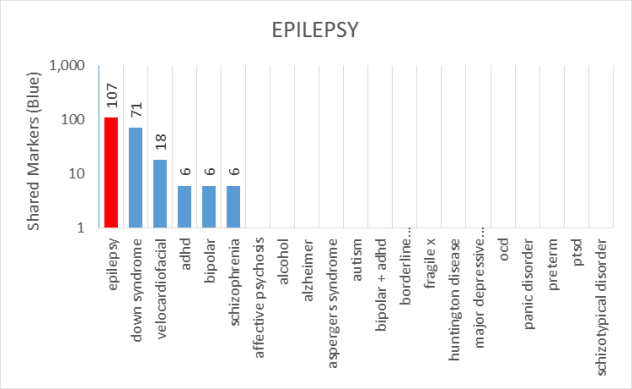

BIPOLAR + ADHD
Shared Markers (Blue)
1,000
100
10
1
624
576
156
132
30
bipolar + adhd
bipolar
adhd
schizophrenia
alzheimer
affective psychosis
alcohol
aspergers syndrome
autism
borderline...
down syndrome
epilepsy
fragile x
huntington disease
major depressive...
ocd
panic disorder
preterm
ptsd
schizotypical disorder
velocardiofacial

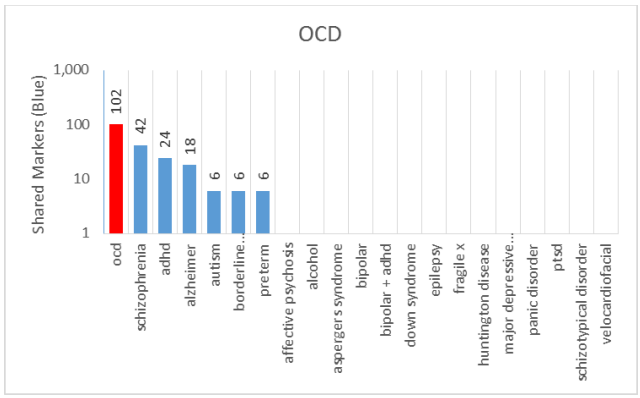

ASPERGERS SYNDROME
Shared Markers (Blue)
10,000
1,000
100
10
1
1,566
864
354
162
108
72
48
30
6
6
6
aspergers syndrome
velocardiofacial
schizophrenia
adhd
schizotypical disorder
ptsd
major depressive...
alzheimer
bipolar
borderline...
preterm
affective psychosis
alcohol
autism
bipolar + adhd
down syndrome
epilepsy
fragile x
huntington disease
ocd
panic disorder

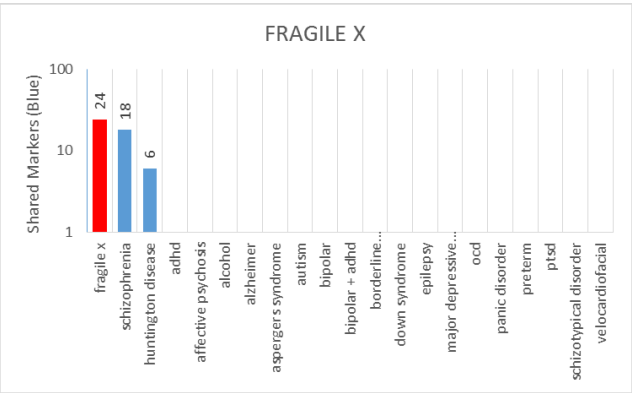

BORDERLINE PERSONALITY DISORDER
Shared Markers (Blue)
10,000
1,000
100
10
1
3,471
1,791
828
342
318
234
132
42
6
6
6
6
borderline...
schizophrenia
alzheimer
bipolar
major depressive...
ptsd
preterm
adhd
aspergers syndrome
autism
ocd
velocardiofacial
affective psychosis
alcohol
bipolar + adhd
down syndrome
epilepsy
fragile x
huntington disease
panic disorder
schizotypical disorder

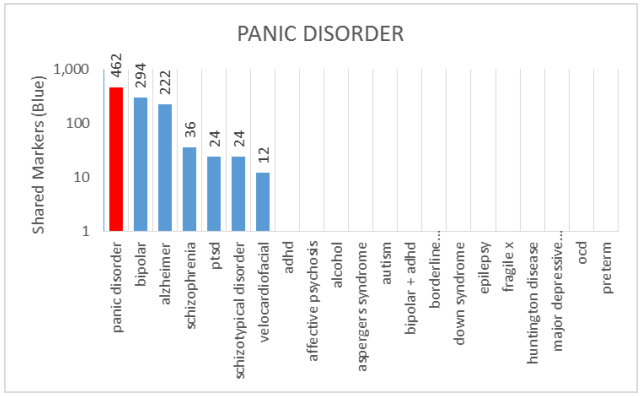

AUTISM
Shared Markers (Blue)
1,000
100
10
1
300
222
30
30
12
6
6
autism
schizophrenia
alzheimer
major depressive...
bipolar
borderline...
ocd
adhd
affective psychosis
alcohol
aspergers syndrome
bipolar + adhd
down syndrome
epilepsy
fragile x
huntington disease
panic disorder
preterm
ptsd
schizotypical disorder
velocardiofacial

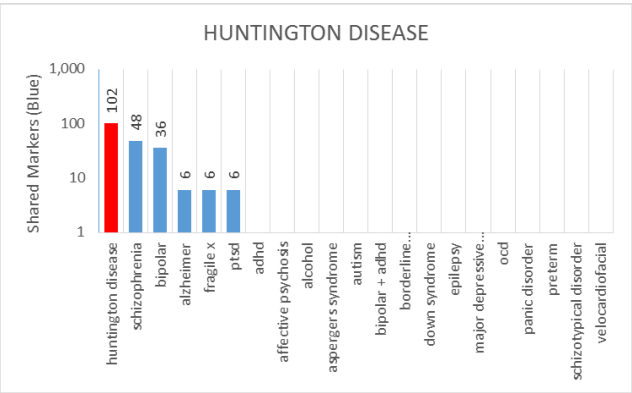

DOWN SYNDROME
Shared Markers (Blue)
1,000
100
10
1
239
78
71
60
18
18
6
down syndrome
velocardiofacial
epilepsy
bipolar
adhd
schizophrenia
preterm
affective psychosis
alcohol
alzheimer
aspergers syndrome
autism
bipolar + adhd
borderline...
fragile x
huntington disease
major depressive...
ocd
panic disorder
ptsd
schizotypical disorder

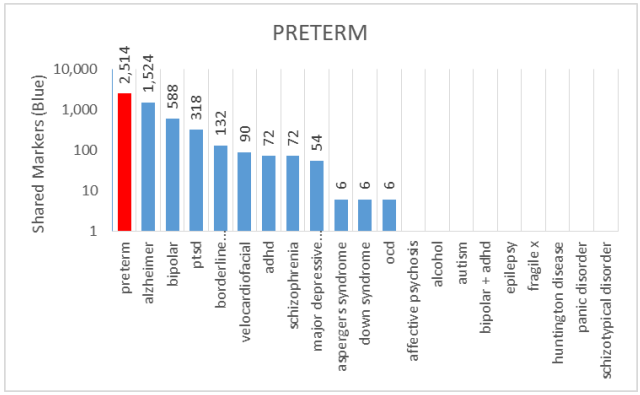

EPILEPSY
Shared Markers (Blue)
1,000
100
10
1
107
71
18
6
6
6
epilepsy
down syndrome
velocardiofacial
adhd
bipolar
schizophrenia
affective psychosis
alcohol
alzheimer
aspergers syndrome
autism
bipolar + adhd
borderline...
fragile x
huntington disease
major depressive...
ocd
panic disorder
preterm
ptsd
schizotypical disorder

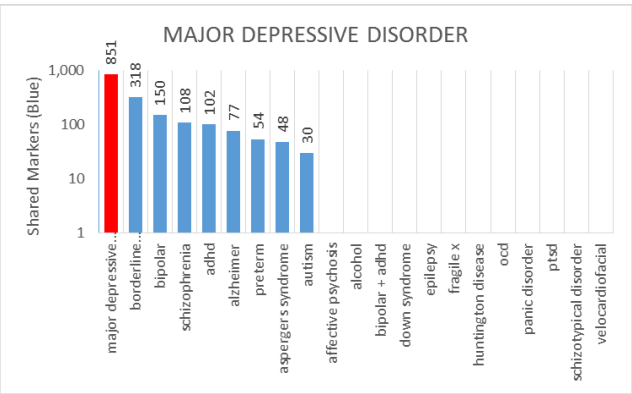

OCD
Shared Markers (Blue)
1,000
100
10
1
102
42
24
18
6
6
6
ocd
schizophrenia
adhd
alzheimer
autism
borderline...
preterm
affective psychosis
alcohol
aspergers syndrome
bipolar
bipolar + adhd
down syndrome
epilepsy
fragile x
huntington disease
major depressive...
panic disorder
ptsd
schizotypical disorder
velocardiofacial

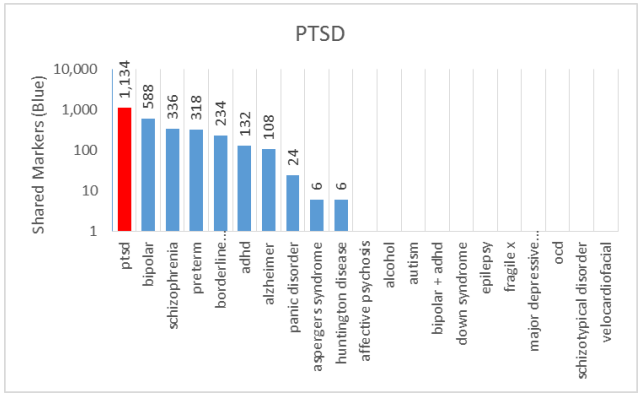

FRAGILE X
Shared Markers (Blue)
100
10
1
24
18
6
fragile x
schizophrenia
huntington disease
adhd
affective psychosis
alcohol
alzheimer
aspergers syndrome
autism
bipolar
bipolar + adhd
borderline...
down syndrome
epilepsy
major depressive...
ocd
panic disorder
preterm
ptsd
schizotypical disorder
velocardiofacial

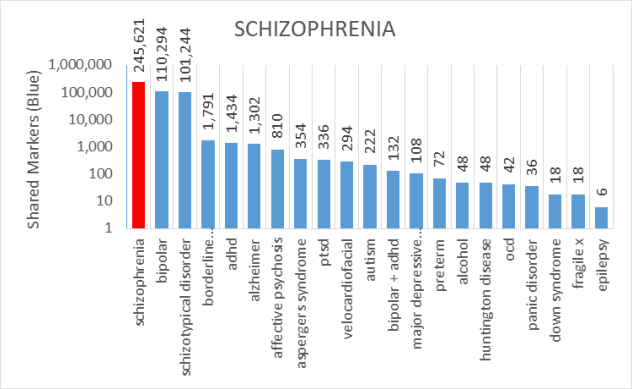

PANIC DISORDER
Shared Markers (Blue)
1,000
100
10
1
462
294
222
36
24
24
12
panic disorder
bipolar
alzheimer
schizophrenia
ptsd
schizotypical disorder
velocardiofacial
adhd
affective psychosis
alcohol
aspergers syndrome
autism
bipolar + adhd
borderline...
down syndrome
epilepsy
fragile x
huntington disease
major depressive...
ocd
preterm

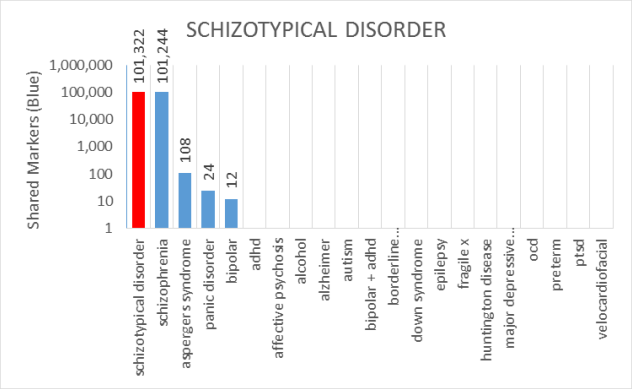

HUNTINGTON DISEASE
Shared Markers (Blue)
1,000
100
10
1
102
48
36
6
6
6
huntington disease
schizophrenia
bipolar
alzheimer
fragile x
ptsd
adhd
affective psychosis
alcohol
aspergers syndrome
autism
bipolar + adhd
borderline...
down syndrome
epilepsy
major depressive...
ocd
panic disorder
preterm
schizotypical disorder
velocardiofacial

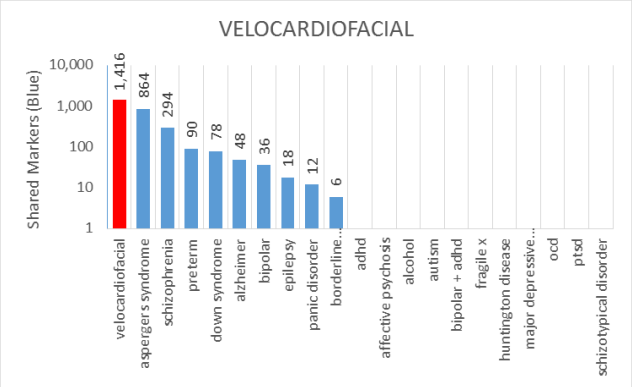

PRETERM
Shared Markers (Blue)
10,000
1,000
100
10
1
2,514
1,524
588
318
132
90
72
72
54
6
6
6
preterm
alzheimer
bipolar
ptsd
borderline...
velocardiofacial
adhd
schizophrenia
major depressive...
aspergers syndrome
down syndrome
ocd
affective psychosis
alcohol
autism
bipolar + adhd
epilepsy
fragile x
huntington disease
panic disorder
schizotypical disorder


MAJOR DEPRESSIVE DISORDER
Shared Markers (Blue)
1,000
100
10
1
851
318
150
108
102
77
54
48
30
major depressive...
borderline...
bipolar
schizophrenia
adhd
alzheimer
preterm
aspergers syndrome
autism
affective psychosis
alcohol
bipolar + adhd
down syndrome
epilepsy
fragile x
huntington disease
ocd
panic disorder
ptsd
schizotypical disorder
velocardiofacial


PTSD
Shared Markers (Blue)
10,000
1,000
100
10
1
1,134
588
336
318
234
132
108
24
6
6
ptsd
bipolar
schizophrenia
preterm
borderline...
adhd
alzheimer
panic disorder
aspergers syndrome
huntington disease
affective psychosis
alcohol
autism
bipolar + adhd
down syndrome
epilepsy
fragile x
major depressive...
ocd
schizotypical disorder
velocardiofacial







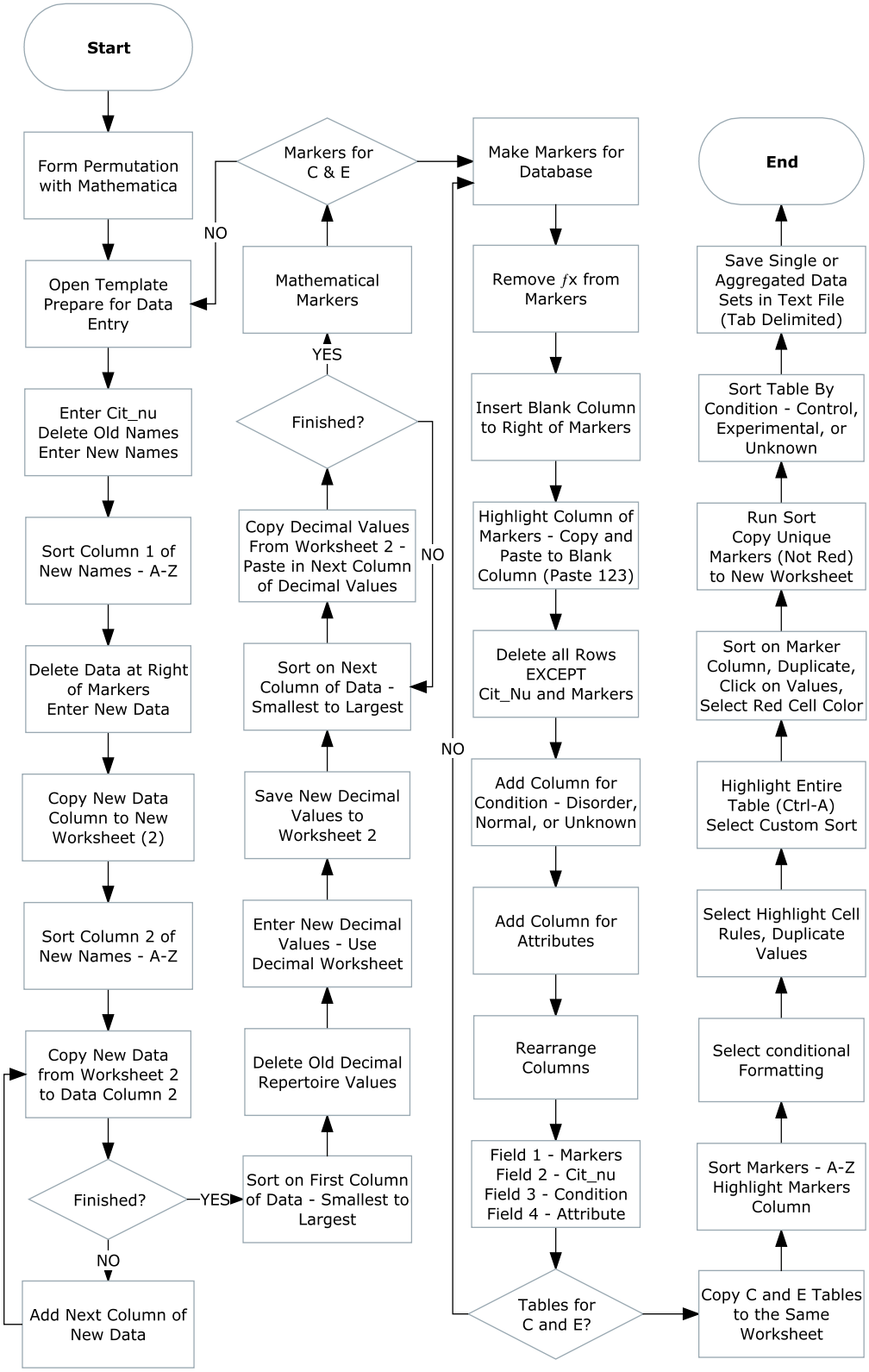
APPENDIX II

Algorithms

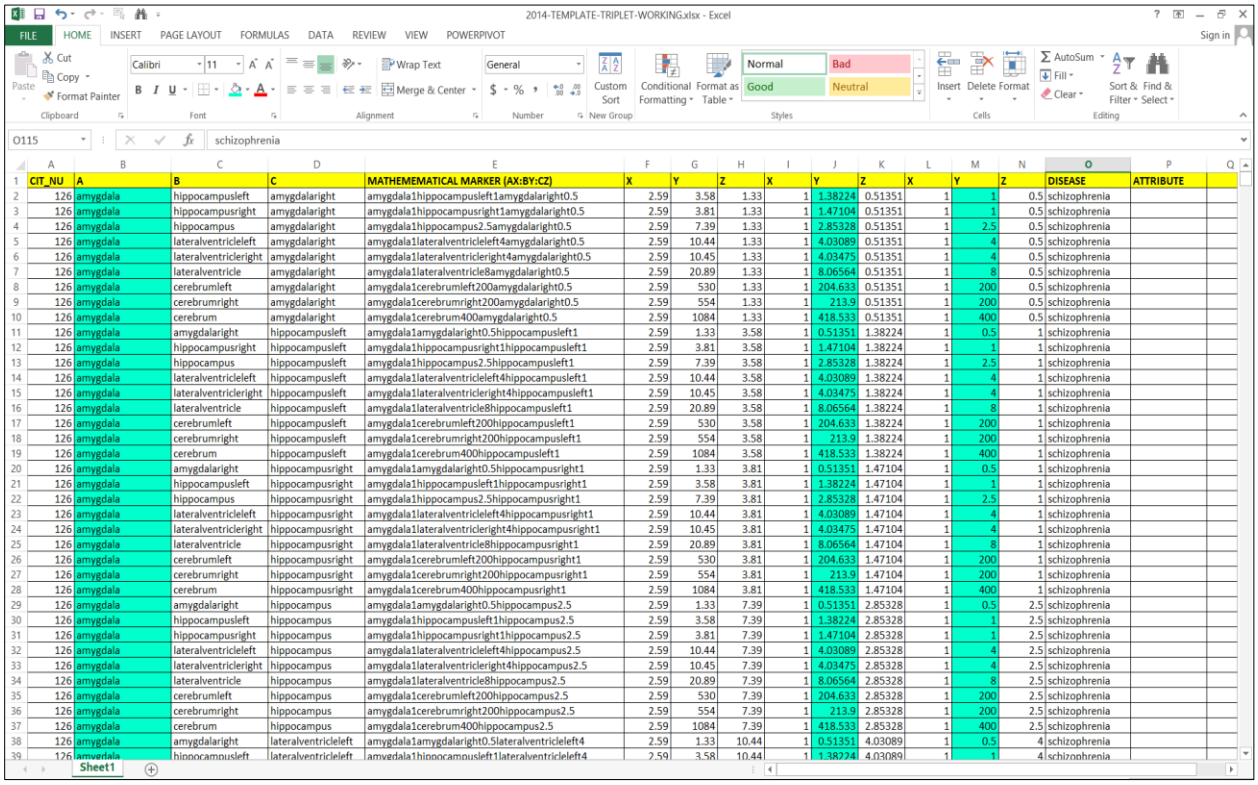
Figuring out how to assemble a diagnostic database from the biology literature involves a number of steps and software programs. For convenience, we can summarize the process with three algorithms: making mathematical markers, populating a database, and diagnosing an unknown.

Summary: We begin with the permutation function of Mathematica that allows us to generate alpha strings from lists of parts (IBVD). An Excel template simplifies the task of populating these strings with data to produce mathematical markers, which, in turn, are filtered, aggregated, and saved as tab delimited text files. When imported into Access these files become databases, which, for example, can undergo additional filtering to produce diagnostic tools. A diagnosis consists of appending a text file of unknown markers to the diagnosis database and then matching unknowns to knowns.

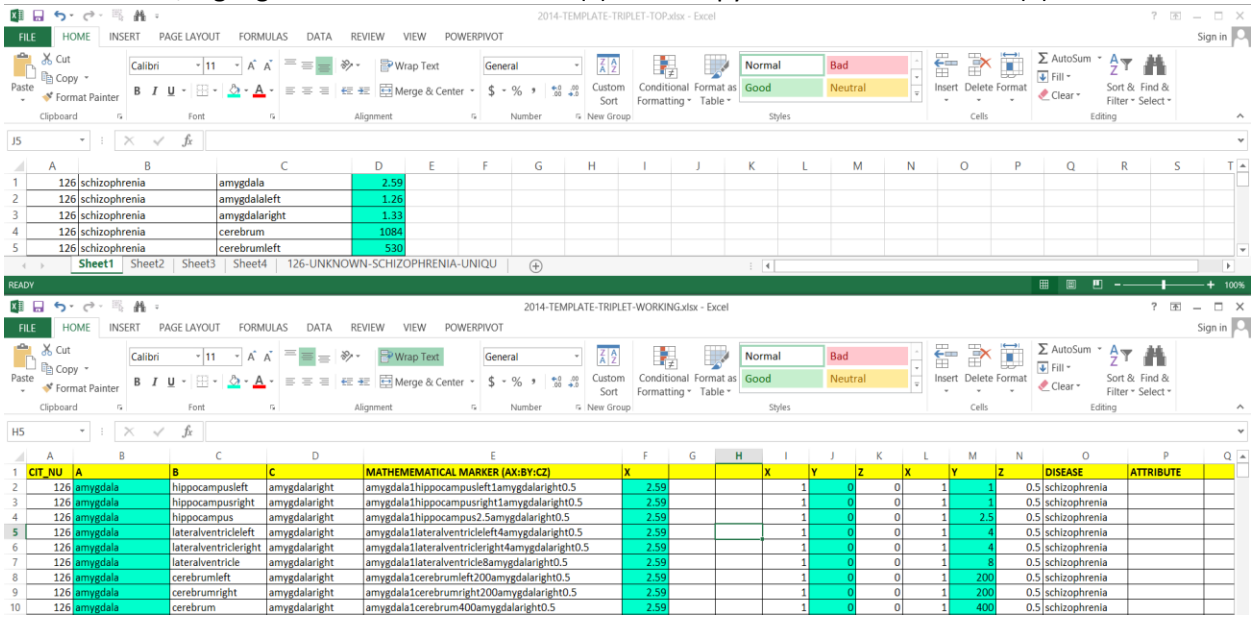
Making Mathematical Markers



The triplet template includes a data set entered for publication 126 of the IBVD. Begin by entering the citation numbers, delete the contents of the three parts columns (A, B, C), replace them with the new ones generated with Mathematica, delete the contents of three columns X, Y, and Z (F, G, H), and assign new values to the parts.



Use a split screen to arrange the worksheet with the data values above the template. Begin by copying the data for amygdala (2.59) from the top sheet and pasting it in the lower sheet (row 2, column X (F)). Click on the data field, move the pointer to the lower right hand corner of the box, hold the left button down, pull on the corner to fill all the boxes below with the value for the amygdala. Repeat this procedure for all the remaining data. When finished, highlight the column of data – X (F) – and copy it to a new worksheet (2) in the same workbook.



Fill the remaining two columns (G and H) using the column of data stored in worksheet 2, as follows. The Highlight the entire screen (Ctrl-A), sort on column B, copy the column in worksheet 2 and paste it into column G (Label the column Y). Repeat the procedure for the third data column (H (Z)).

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	CIT_NU	A	B	C	MATHEMEMATICAL MARKER (AX:BY:CZ)	X			X	Y	Z	X	Y	Z	DISEASE	ATTRIBUTE
2	126	amygdala	hippocampusleft	amygdalaright	amygdala1hippocampusleft1amygdalaright0.5	2.59			1	0	0	1	1	0.5	schizophrenia	
3	126	amygdala	hippocampusright	amygdalaright	amygdala1hippocampusright1amygdalaright0.5	2.59			1	0	0	1	1	0.5	schizophrenia	
4	126	amygdala	hippocampus	amygdalaright	amygdala1hippocampus2.5amygdalaright0.5	2.59										
5	126	amygdala	lateralventricleleft	amygdalaright	amygdala1lateralventricleleft4amygdalaright0.5	2.59										
6	126	amygdala	lateralventricleright	amygdalaright	amygdala1lateralventricleright4amygdalaright0.5	2.59										
7	126	amygdala	lateralventricle	amygdalaright	amygdala1lateralventricle8amygdalaright0.5	2.59										
8	126	amygdala	cerebrumleft	amygdalaright	amygdala1cerebrumleft200amygdalaright0.5	2.59										
9	126	amygdala	cerebrumright	amygdalaright	amygdala1cerebrumright200amygdalaright0.5	2.59										
10	126	amygdala	cerebrum	amygdalaright	amygdala1cerebrum400amygdalaright0.5	2.59										
11	126	amygdala	amygdalaright	hippocampusleft	amygdala1amygdalaright0.5hippocampusleft1	2.59										
12	126	amygdala	hippocampusright	hippocampusleft	amygdala1hippocampusright1hippocampusleft1	2.59										
13	126	amygdala	hippocampus	hippocampusleft	amygdala1hippocampus2.5hippocampusleft1	2.59										
14	126	amygdala	lateralventricleleft	hippocampusleft	amygdala1lateralventricleleft4hippocampusleft1	2.59										
15	126	amygdala	lateralventricleright	hippocampusleft	amygdala1lateralventricleright4hippocampusleft1	2.59										
16	126	amygdala	lateralventricle	hippocampusleft	amygdala1lateralventricle8hippocampusleft1	2.59										
17	126	amygdala	cerebrumleft	hippocampusleft	amygdala1cerebrumleft200hippocampusleft1	2.59										
18	126	amygdala	cerebrumright	hippocampusleft	amygdala1cerebrumright200hippocampusleft1	2.59										
19	126	amygdala	cerebrum	hippocampusleft	amygdala1cerebrum400hippocampusleft1	2.59										
20	126	amygdala	amygdalaright	hippocampusright	amygdala1amygdalaright0.5hippocampusright1	2.59			1	0	0	1	0.5	1	schizophrenia	
21	126	amygdala	hippocampusleft	hippocampusright	amygdala1hippocampusleft1hippocampusright1	2.59			1	0	0	1	1	1	schizophrenia	

Sort
Add Level | Delete Level | Copy Level | Options... | My data has headers
Column: Sort by B | Sort On: Values | Order: A to Z
OK | Cancel

The resulting screen will have data stored in columns X, Y, and Z. Delete the contents of rows M and N.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	CIT_NU	A	B	C	MATHEMEMATICAL MARKER (AX:BY:CZ)	X	Y	Z	X	Y	Z	X			DISEASE	ATTRIBUTE
2	126	cerebrum	amygdalaleft	amygdala	cerebrum1amygdalaleftamygdala	1084	1.26	2.59	1	0.00116	0.00239	1			schizophrenia	
3	126	cerebrum	amygdalaleft	amygdala	cerebrum1amygdalaleftamygdala	1084	1.26	2.59	1	0.00116	0.00239	1			schizophrenia	
4	126	cerebrum	amygdalaleft	amygdala	cerebrum1amygdalaleftamygdala	1084	1.26	2.59	1	0.00116	0.00239	1			schizophrenia	
5	126	cerebrum	amygdalaleft	amygdala	cerebrum1amygdalaleftamygdala	1084	1.26	2.59	1	0.00116	0.00239	1			schizophrenia	
6	126	cerebrum	amygdalaleft	amygdala	cerebrum1amygdalaleftamygdala	1084	1.26	2.59	1	0.00116	0.00239	1			schizophrenia	
7	126	cerebrum	amygdalaleft	amygdala	cerebrum1amygdalaleftamygdala	1084	1.26	2.59	1	0.00116	0.00239	1			schizophrenia	
8	126	cerebrum	amygdalaleft	amygdala	cerebrum1amygdalaleftamygdala	1084	1.26	2.59	1	0.00116	0.00239	1			schizophrenia	
9	126	cerebrum	amygdalaleft	amygdala	cerebrum1amygdalaleftamygdala	1084	1.26	2.59	1	0.00116	0.00239	1			schizophrenia	
10	126	cerebrum	amygdalaleft	amygdala	cerebrum1amygdalaleftamygdala	1084	1.26	2.59	1	0.00116	0.00239	1			schizophrenia	
11	126	cerebrum	amygdalaleft	amygdalaright	cerebrum1amygdalaleftamygdalaright	1084	1.26	1.33	1	0.00116	0.00123	1			schizophrenia	
12	126	cerebrum	amygdalaleft	amygdalaright	cerebrum1amygdalaleftamygdalaright	1084	1.26	1.33	1	0.00116	0.00123	1			schizophrenia	
13	126	cerebrum	amygdalaleft	amygdalaright	cerebrum1amygdalaleftamygdalaright	1084	1.26	1.33	1	0.00116	0.00123	1			schizophrenia	
14	126	cerebrum	amygdalaleft	amygdalaright	cerebrum1amygdalaleftamygdalaright	1084	1.26	1.33	1	0.00116	0.00123	1			schizophrenia	
15	126	cerebrum	amygdalaleft	amygdalaright	cerebrum1amygdalaleftamygdalaright	1084	1.26	1.33	1	0.00116	0.00123	1			schizophrenia	
16	126	cerebrum	amygdalaleft	amygdalaright	cerebrum1amygdalaleftamygdalaright	1084	1.26	1.33	1	0.00116	0.00123	1			schizophrenia	
17	126	cerebrum	amygdalaleft	amygdalaright	cerebrum1amygdalaleftamygdalaright	1084	1.26	1.33	1	0.00116	0.00123	1			schizophrenia	
18	126	cerebrum	amygdalaleft	amygdalaright	cerebrum1amygdalaleftamygdalaright	1084	1.26	1.33	1	0.00116	0.00123	1			schizophrenia	
19	126	cerebrum	amygdalaleft	amygdalaright	cerebrum1amygdalaleftamygdalaright	1084	1.26	1.33	1	0.00116	0.00123	1			schizophrenia	
20	126	cerebrum	amygdalaleft	cerebrumleft	cerebrum1amygdalaleftcerebrumleft	1084	1.26	530	1	0.00116	0.48893	1			schizophrenia	
21	126	cerebrum	amygdalaleft	cerebrumleft	cerebrum1amygdalaleftcerebrumleft	1084	1.26	530	1	0.00116	0.48893	1			schizophrenia	

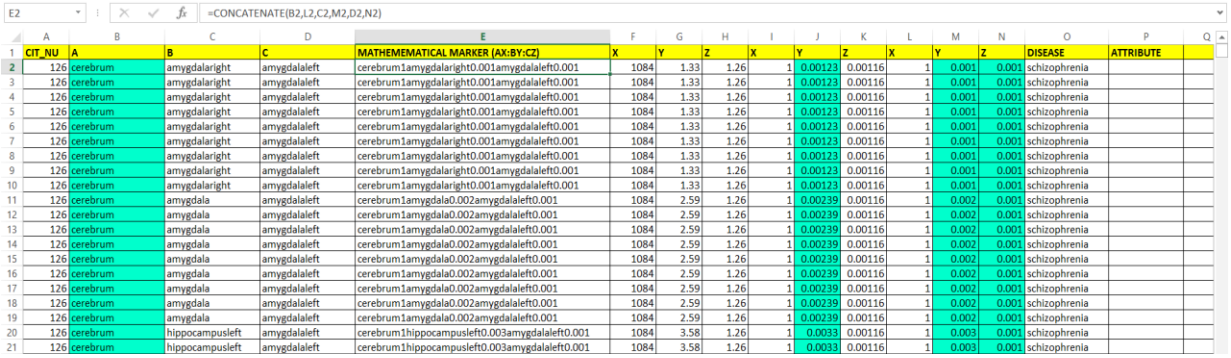
Highlight screen, sort on column Y (J), and then translate the numbers in column J to decimal repertoire values in column M - use the connection_phenotype_worksheet.pdf in the Forms section of Documents in the software package. When completed, store the completed column in worksheet 2. Finally, highlight the screen, sort on column Z (K), copy the column from worksheet 2, and paste it in column N.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	CIT_NU	A	B	C	MATHEMEMATICAL MARKER (AX:BY:CZ)	X	Y	Z	X	Y	Z	X	Y		DISEASE	ATTRIBUTE
2	126	cerebrum	amygdalaleft	amygdala	cerebrum1amygdalaleft0.001amygdala	1084	1.26	2.59	1	0.001162	0.002389	1	0.001		schizophrenia	
3	126	cerebrum	amygdalaleft	hippocampusleft	cerebrum1amygdalaleft0.001hippocampusleft	1084	1.26	3.58	1	0.001162	0.003303	1	0.001		schizophrenia	
4	126	cerebrum	amygdalaleft	hippocampusright	cerebrum1amygdalaleft0.001hippocampusright	1084	1.26	3.81	1	0.001162	0.003515	1	0.001		schizophrenia	
5	126	cerebrum	amygdalaleft	hippocampus	cerebrum1amygdalaleft0.001hippocampus	1084	1.26	7.39	1	0.001162	0.006817	1	0.001		schizophrenia	
6	126	cerebrum	amygdalaleft	lateralventricleleft	cerebrum1amygdalaleft0.001lateralventricleleft	1084	1.26	10.44	1	0.001162	0.009631	1	0.001		schizophrenia	
7	126	cerebrum	amygdalaleft	lateralventricleright	cerebrum1amygdalaleft0.001lateralventricleright	1084	1.26	10.45	1	0.001162	0.00964	1	0.001		schizophrenia	
8	126	cerebrum	amygdalaleft	lateralventricle	cerebrum1amygdalaleft0.001lateralventricle	1084	1.26	20.89	1	0.001162	0.019271	1	0.001		schizophrenia	
9	126	cerebrum	amygdalaleft	cerebrumleft	cerebrum1amygdalaleft0.001cerebrumleft	1084	1.26	530	1	0.001162	0.48893	1	0.001		schizophrenia	
10	126	cerebrum	amygdalaleft	cerebrumright	cerebrum1amygdalaleft0.001cerebrumright	1084	1.26	554	1	0.001162	0.51107	1	0.001		schizophrenia	
11	126	cerebrum	amygdalaleft	amygdalaright	cerebrum1amygdalaleft0.001amygdalaright	1084	1.26	1.33	1	0.001162	0.001227	1	0.001		schizophrenia	
12	126	cerebrum	amygdalaleft	hippocampusleft	cerebrum1amygdalaleft0.001hippocampusleft	1084	1.26	3.58	1	0.001162	0.003303	1	0.001		schizophrenia	
13	126	cerebrum	amygdalaleft	hippocampusright	cerebrum1amygdalaleft0.001hippocampusright	1084	1.26	3.81	1	0.001162	0.003515	1	0.001		schizophrenia	
14	126	cerebrum	amygdalaleft	hippocampus	cerebrum1amygdalaleft0.001hippocampus	1084	1.26	7.39	1	0.001162	0.006817	1	0.001		schizophrenia	
15	126	cerebrum	amygdalaleft	lateralventricleleft	cerebrum1amygdalaleft0.001lateralventricleleft	1084	1.26	10.44	1	0.001162	0.009631	1	0.001		schizophrenia	
16	126	cerebrum	amygdalaleft	lateralventricleright	cerebrum1amygdalaleft0.001lateralventricleright	1084	1.26	10.45	1	0.001162	0.00964	1	0.001		schizophrenia	
17	126	cerebrum	amygdalaleft	lateralventricle	cerebrum1amygdalaleft0.001lateralventricle	1084	1.26	20.89	1	0.001162	0.019271	1	0.001		schizophrenia	
18	126	cerebrum	amygdalaleft	cerebrumleft	cerebrum1amygdalaleft0.001cerebrumleft	1084	1.26	530	1	0.001162	0.48893	1	0.001		schizophrenia	
19	126	cerebrum	amygdalaleft	cerebrumright	cerebrum1amygdalaleft0.001cerebrumright	1084	1.26	554	1	0.001162	0.51107	1	0.001		schizophrenia	
20	126	cerebrum	amygdalaleft	amygdalaright	cerebrum1amygdalaleft0.001amygdalaright	1084	1.26	1.33	1	0.001162	0.001227	1	0.001		schizophrenia	
21	126	cerebrum	amygdalaleft	amygdala	cerebrum1amygdalaleft0.001amygdala	1084	1.26	2.59	1	0.001162	0.002389	1	0.001		schizophrenia	

Worksheet 2 displays the columns of data used for data entry (A=original values, C=decimal repertoire values).

	A	B	C
1	X		Y
2	2.59		0.001
3	2.59		0.001
4	2.59		0.001
5	2.59		0.001
6	2.59		0.001
7	2.59		0.001
8	2.59		0.001
9	2.59		0.001
10	2.59		0.001
11	2.59		0.001
12	2.59		0.001
13	2.59		0.001
14	2.59		0.001
15	2.59		0.001
16	2.59		0.001

The completed data entry screen appears below. The next step consists of producing a text file that can be imported into Access to become a database. Triplet markers, which concatenate three parts (A, B, C) with three values (X, Y, Z), must first be converted into text strings. Highlight column F, right click, select Insert, highlight column E, copy column E, and paste it into the newly created column.



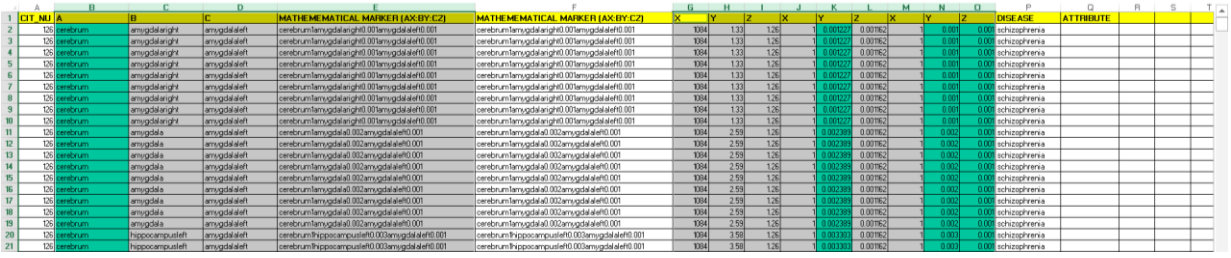
E2 =CONCATENATE(B2,L2,C2,M2,D2,N2)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	CIT_NU	A	B	C	MATHEMEMATICAL MARKER (AX:BY:CZ)	X	Y	Z	X	Y	Z	X	Y	Z	DISEASE	ATTRIBUTE	
2	126	cerebrum	amygdalaright	amygdalaleft	cerebrum1amygdalaright0.001amygdalaleft0.001	1084	1.33	1.26	1	0.00123	0.00116	1	0.001	0.001	schizophrenia		
3	126	cerebrum	amygdalaright	amygdalaleft	cerebrum1amygdalaright0.001amygdalaleft0.001	1084	1.33	1.26	1	0.00123	0.00116	1	0.001	0.001	schizophrenia		
4	126	cerebrum	amygdalaright	amygdalaleft	cerebrum1amygdalaright0.001amygdalaleft0.001	1084	1.33	1.26	1	0.00123	0.00116	1	0.001	0.001	schizophrenia		
5	126	cerebrum	amygdalaright	amygdalaleft	cerebrum1amygdalaright0.001amygdalaleft0.001	1084	1.33	1.26	1	0.00123	0.00116	1	0.001	0.001	schizophrenia		
6	126	cerebrum	amygdalaright	amygdalaleft	cerebrum1amygdalaright0.001amygdalaleft0.001	1084	1.33	1.26	1	0.00123	0.00116	1	0.001	0.001	schizophrenia		
7	126	cerebrum	amygdalaright	amygdalaleft	cerebrum1amygdalaright0.001amygdalaleft0.001	1084	1.33	1.26	1	0.00123	0.00116	1	0.001	0.001	schizophrenia		
8	126	cerebrum	amygdalaright	amygdalaleft	cerebrum1amygdalaright0.001amygdalaleft0.001	1084	1.33	1.26	1	0.00123	0.00116	1	0.001	0.001	schizophrenia		
9	126	cerebrum	amygdalaright	amygdalaleft	cerebrum1amygdalaright0.001amygdalaleft0.001	1084	1.33	1.26	1	0.00123	0.00116	1	0.001	0.001	schizophrenia		
10	126	cerebrum	amygdalaright	amygdalaleft	cerebrum1amygdalaright0.001amygdalaleft0.001	1084	1.33	1.26	1	0.00123	0.00116	1	0.001	0.001	schizophrenia		
11	126	cerebrum	amygdala	amygdalaleft	cerebrum1amygdala0.002amygdalaleft0.001	1084	2.59	1.26	1	0.00239	0.00116	1	0.002	0.001	schizophrenia		
12	126	cerebrum	amygdala	amygdalaleft	cerebrum1amygdala0.002amygdalaleft0.001	1084	2.59	1.26	1	0.00239	0.00116	1	0.002	0.001	schizophrenia		
13	126	cerebrum	amygdala	amygdalaleft	cerebrum1amygdala0.002amygdalaleft0.001	1084	2.59	1.26	1	0.00239	0.00116	1	0.002	0.001	schizophrenia		
14	126	cerebrum	amygdala	amygdalaleft	cerebrum1amygdala0.002amygdalaleft0.001	1084	2.59	1.26	1	0.00239	0.00116	1	0.002	0.001	schizophrenia		
15	126	cerebrum	amygdala	amygdalaleft	cerebrum1amygdala0.002amygdalaleft0.001	1084	2.59	1.26	1	0.00239	0.00116	1	0.002	0.001	schizophrenia		
16	126	cerebrum	amygdala	amygdalaleft	cerebrum1amygdala0.002amygdalaleft0.001	1084	2.59	1.26	1	0.00239	0.00116	1	0.002	0.001	schizophrenia		
17	126	cerebrum	amygdala	amygdalaleft	cerebrum1amygdala0.002amygdalaleft0.001	1084	2.59	1.26	1	0.00239	0.00116	1	0.002	0.001	schizophrenia		
18	126	cerebrum	amygdala	amygdalaleft	cerebrum1amygdala0.002amygdalaleft0.001	1084	2.59	1.26	1	0.00239	0.00116	1	0.002	0.001	schizophrenia		
19	126	cerebrum	amygdala	amygdalaleft	cerebrum1amygdala0.002amygdalaleft0.001	1084	2.59	1.26	1	0.00239	0.00116	1	0.002	0.001	schizophrenia		
20	126	cerebrum	hippocampusleft	amygdalaleft	cerebrum1hippocampusleft0.003amygdalaleft0.001	1084	3.58	1.26	1	0.0033	0.00116	1	0.003	0.001	schizophrenia		
21	126	cerebrum	hippocampusleft	amygdalaleft	cerebrum1hippocampusleft0.003amygdalaleft0.001	1084	3.58	1.26	1	0.0033	0.00116	1	0.003	0.001	schizophrenia		

When copied, the markers in column F will not match those in column E – click on the first Paste Values labeled 123 and they will.

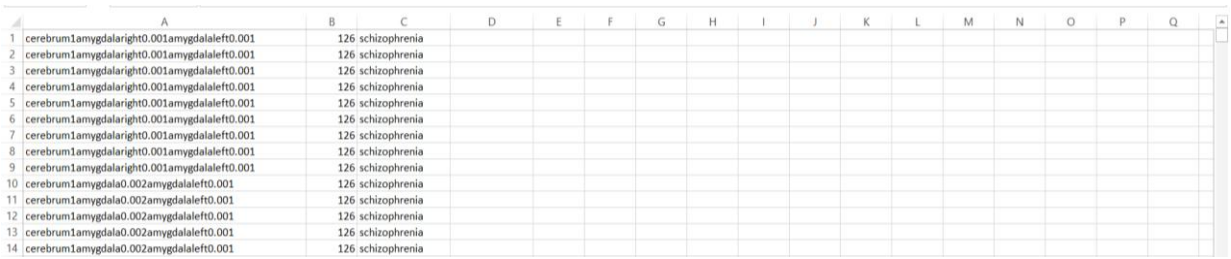
	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	CIT_NU	A	B	C	MATHEMEMATICAL MARKER (AX:BY:CZ)	MATHEMEMATICAL MARKER (AX:BY:CZ)	X	Y	Z	X	Y	Z	X	Y
2	126	cerebrum	amygdalaright	amygdalaleft	cerebrum1amygdalaright0.001amygdalaleft0.001	amygdalaright0.001amygdalaleft0.001cerebrum1amygd				1.26	1	0.00123	0.00116	1
3	126	cerebrum	amygdalaright	amygdalaleft	cerebrum1amygdalaright0.001amygdalaleft0.001	amygdalaright0.001amygdalaleft0.001cerebrum1amygd				1.26	1	0.00123	0.00116	1
4	126	cerebrum	amygdalaright	amygdalaleft	cerebrum1amygdalaright0.001amygdalaleft0.001	amygdalaright0.001amygdalaleft0.001cerebrum1amygd				1.26	1	0.00123	0.00116	1
5	126	cerebrum	amygdalaright	amygdalaleft	cerebrum1amygdalaright0.001amygdalaleft0.001	amygdalaright0.001amygdalaleft0.001cerebrum1amygd				1.26	1	0.00123	0.00116	1
6	126	cerebrum	amygdalaright	amygdalaleft	cerebrum1amygdalaright0.001amygdalaleft0.001	amygdalaright0.001amygdalaleft0.001cerebrum1amygd				1.26	1	0.00123	0.00116	1
7	126	cerebrum	amygdalaright	amygdalaleft	cerebrum1amygdalaright0.001amygdalaleft0.001	amygdalaright0.001amygdalaleft0.001cerebrum1amygd				1.26	1	0.00123	0.00116	1
8	126	cerebrum	amygdalaright	amygdalaleft	cerebrum1amygdalaright0.001amygdalaleft0.001	amygdalaright0.001amygdalaleft0.001cerebrum1amygd				1.26	1	0.00123	0.00116	1
9	126	cerebrum	amygdalaright	amygdalaleft	cerebrum1amygdalaright0.001amygdalaleft0.001	amygdalaright0.001amygdalaleft0.001cerebrum1amygd				1.26	1	0.00123	0.00116	1
10	126	cerebrum	amygdalaright	amygdalaleft	cerebrum1amygdalaright0.001amygdalaleft0.001	amygdalaright0.001amygdalaleft0.001cerebrum1amygd				1.26	1	0.00123	0.00116	1
11	126	cerebrum	amygdala	amygdalaleft	cerebrum1amygdala0.002amygdalaleft0.001	amygdala0.002amygdalaleft0.001cerebrum1amygdala0.				1.26	1	0.00239	0.00116	1
12	126	cerebrum	amygdala	amygdalaleft	cerebrum1amygdala0.002amygdalaleft0.001	amygdala0.002amygdalaleft0.001cerebrum1amygdala0.				1.26	1	0.00239	0.00116	1
13	126	cerebrum	amygdala	amygdalaleft	cerebrum1amygdala0.002amygdalaleft0.001	amygdala0.002amygdalaleft0.001cerebrum1amygdala0.	1084	2.59	1.26	1	0.00239	0.00116	1	0
14	126	cerebrum	amygdala	amygdalaleft	cerebrum1amygdala0.002amygdalaleft0.001	amygdala0.002amygdalaleft0.001cerebrum1amygdala0.	1084	2.59	1.26	1	0.00239	0.00116	1	0
15	126	cerebrum	amygdala	amygdalaleft	cerebrum1amygdala0.002amygdalaleft0.001	amygdala0.002amygdalaleft0.001cerebrum1amygdala0.	1084	2.59	1.26	1	0.00239	0.00116	1	0
16	126	cerebrum	amygdala	amygdalaleft	cerebrum1amygdala0.002amygdalaleft0.001	amygdala0.002amygdalaleft0.001cerebrum1amygdala0.	1084	2.59	1.26	1	0.00239	0.00116	1	0
17	126	cerebrum	amygdala	amygdalaleft	cerebrum1amygdala0.002amygdalaleft0.001	amygdala0.002amygdalaleft0.001cerebrum1amygdala0.	1084	2.59	1.26	1	0.00239	0.00116	1	0
18	126	cerebrum	amygdala	amygdalaleft	cerebrum1amygdala0.002amygdalaleft0.001	amygdala0.002amygdalaleft0.001cerebrum1amygdala0.	1084	2.59	1.26	1	0.00239	0.00116	1	0
19	126	cerebrum	amygdala	amygdalaleft	cerebrum1amygdala0.002amygdalaleft0.001	amygdala0.002amygdalaleft0.001cerebrum1amygdala0.	1084	2.59	1.26	1	0.00239	0.00116	1	0
20	126	cerebrum	hippocampusleft	amygdalaleft	cerebrum1hippocampusleft0.003amygdalaleft0.001	hippocampusleft0.003amygdalaleft0.001cerebrum1hipp	1084	3.58	1.26	1	0.0033	0.00116	1	0
21	126	cerebrum	hippocampusleft	amygdalaleft	cerebrum1hippocampusleft0.003amygdalaleft0.001	hippocampusleft0.003amygdalaleft0.001cerebrum1hipp	1084	3.58	1.26	1	0.0033	0.00116	1	0

Make a backup copy of the workbook. Finally, highlight the columns as shown below, right click, and delete.



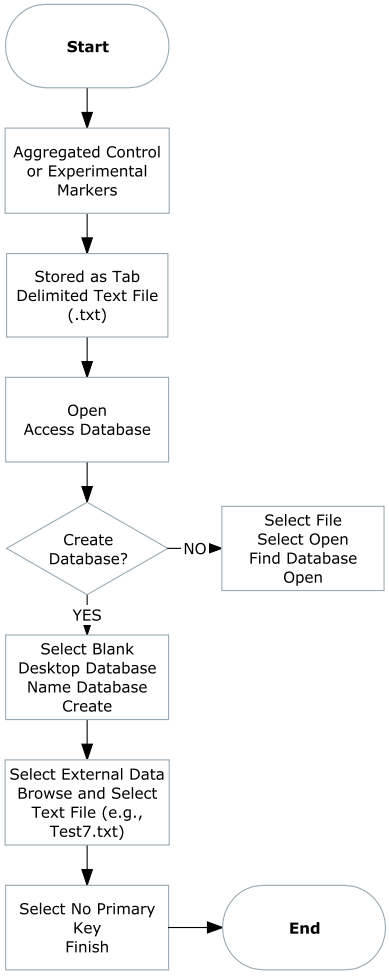
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	CIT_NU	A	B	C	MATHEMEMATICAL MARKER (AX:BY:CZ)	MATHEMEMATICAL MARKER (AX:BY:CZ)	X	Y	Z	X	Y	Z	X	Y	Z	DISEASE	ATTRIBUTE
2	126	cerebrum	amygdalaright	amygdalaleft	cerebrum1amygdalaright0.001amygdalaleft0.001	cerebrum1amygdalaright0.001amygdalaleft0.001	1084	1.33	1.26	1	0.00127	0.00116	1	0.001	0.001	schizophrenia	
3	126	cerebrum	amygdalaright	amygdalaleft	cerebrum1amygdalaright0.001amygdalaleft0.001	cerebrum1amygdalaright0.001amygdalaleft0.001	1084	1.33	1.26	1	0.00127	0.00116	1	0.001	0.001	schizophrenia	
4	126	cerebrum	amygdalaright	amygdalaleft	cerebrum1amygdalaright0.001amygdalaleft0.001	cerebrum1amygdalaright0.001amygdalaleft0.001	1084	1.33	1.26	1	0.00127	0.00116	1	0.001	0.001	schizophrenia	
5	126	cerebrum	amygdalaright	amygdalaleft	cerebrum1amygdalaright0.001amygdalaleft0.001	cerebrum1amygdalaright0.001amygdalaleft0.001	1084	1.33	1.26	1	0.00127	0.00116	1	0.001	0.001	schizophrenia	
6	126	cerebrum	amygdalaright	amygdalaleft	cerebrum1amygdalaright0.001amygdalaleft0.001	cerebrum1amygdalaright0.001amygdalaleft0.001	1084	1.33	1.26	1	0.00127	0.00116	1	0.001	0.001	schizophrenia	
7	126	cerebrum	amygdalaright	amygdalaleft	cerebrum1amygdalaright0.001amygdalaleft0.001	cerebrum1amygdalaright0.001amygdalaleft0.001	1084	1.33	1.26	1	0.00127	0.00116	1	0.001	0.001	schizophrenia	
8	126	cerebrum	amygdalaright	amygdalaleft	cerebrum1amygdalaright0.001amygdalaleft0.001	cerebrum1amygdalaright0.001amygdalaleft0.001	1084	1.33	1.26	1	0.00127	0.00116	1	0.001	0.001	schizophrenia	
9	126	cerebrum	amygdalaright	amygdalaleft	cerebrum1amygdalaright0.001amygdalaleft0.001	cerebrum1amygdalaright0.001amygdalaleft0.001	1084	1.33	1.26	1	0.00127	0.00116	1	0.001	0.001	schizophrenia	
10	126	cerebrum	amygdalaright	amygdalaleft	cerebrum1amygdalaright0.001amygdalaleft0.001	cerebrum1amygdalaright0.001amygdalaleft0.001	1084	1.33	1.26	1	0.00127	0.00116	1	0.001	0.001	schizophrenia	
11	126	cerebrum	amygdala	amygdalaleft	cerebrum1amygdala0.002amygdalaleft0.001	cerebrum1amygdala0.002amygdalaleft0.001	1084	2.59	1.26	1	0.00239	0.00116	1	0.002	0.001	schizophrenia	
12	126	cerebrum	amygdala	amygdalaleft	cerebrum1amygdala0.002amygdalaleft0.001	cerebrum1amygdala0.002amygdalaleft0.001	1084	2.59	1.26	1	0.00239	0.00116	1	0.002	0.001	schizophrenia	
13	126	cerebrum	amygdala	amygdalaleft	cerebrum1amygdala0.002amygdalaleft0.001	cerebrum1amygdala0.002amygdalaleft0.001	1084	2.59	1.26	1	0.00239	0.00116	1	0.002	0.001	schizophrenia	
14	126	cerebrum	amygdala	amygdalaleft	cerebrum1amygdala0.002amygdalaleft0.001	cerebrum1amygdala0.002amygdalaleft0.001	1084	2.59	1.26	1	0.00239	0.00116	1	0.002	0.001	schizophrenia	
15	126	cerebrum	amygdala	amygdalaleft	cerebrum1amygdala0.002amygdalaleft0.001	cerebrum1amygdala0.002amygdalaleft0.001	1084	2.59	1.26	1	0.00239	0.00116	1	0.002	0.001	schizophrenia	
16	126	cerebrum	amygdala	amygdalaleft	cerebrum1amygdala0.002amygdalaleft0.001	cerebrum1amygdala0.002amygdalaleft0.001	1084	2.59	1.26	1	0.00239	0.00116	1	0.002	0.001	schizophrenia	
17	126	cerebrum	amygdala	amygdalaleft	cerebrum1amygdala0.002amygdalaleft0.001	cerebrum1amygdala0.002amygdalaleft0.001	1084	2.59	1.26	1	0.00239	0.00116	1	0.002	0.001	schizophrenia	
18	126	cerebrum	amygdala	amygdalaleft	cerebrum1amygdala0.002amygdalaleft0.001	cerebrum1amygdala0.002amygdalaleft0.001	1084	2.59	1.26	1	0.00239	0.00116	1	0.002	0.001	schizophrenia	
19	126	cerebrum	amygdala	amygdalaleft	cerebrum1amygdala0.002amygdalaleft0.001	cerebrum1amygdala0.002amygdalaleft0.001	1084	2.59	1.26	1	0.00239	0.00116	1	0.002	0.001	schizophrenia	
20	126	cerebrum	hippocampusleft	amygdalaleft	cerebrum1hippocampusleft0.003amygdalaleft0.001	cerebrum1hippocampusleft0.003amygdalaleft0.001	1084	3.58	1.26	1	0.00303	0.00116	1	0.003	0.001	schizophrenia	
21	126	cerebrum	hippocampusleft	amygdalaleft	cerebrum1hippocampusleft0.003amygdalaleft0.001	cerebrum1hippocampusleft0.003amygdalaleft0.001	1084	3.58	1.26	1	0.00303	0.00116	1	0.003	0.001	schizophrenia	

Remove the heading, highlight the page and select no border, arrange the order of the columns as shown, add a term – if absent - in column D (e.g., hold) to identify the existence of the attributes column, and store the worksheet as a tab delimited text file (.txt); name it TEST-TRIPLET-DATABASE.txt.



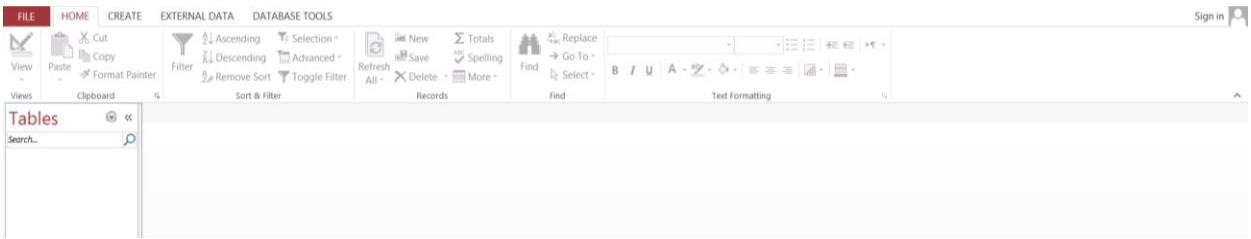
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	cerebrum1amygdalaright0.001amygdalaleft0.001	126	schizophrenia														
2	cerebrum1amygdalaright0.001amygdalaleft0.001	126	schizophrenia														
3	cerebrum1amygdalaright0.001amygdalaleft0.001	126	schizophrenia														
4	cerebrum1amygdalaright0.001amygdalaleft0.001	126	schizophrenia														
5	cerebrum1amygdalaright0.001amygdalaleft0.001	126	schizophrenia														
6	cerebrum1amygdalaright0.001amygdalaleft0.001	126	schizophrenia														
7	cerebrum1amygdalaright0.001amygdalaleft0.001	126	schizophrenia														
8	cerebrum1amygdalaright0.001amygdalaleft0.001	126	schizophrenia														
9	cerebrum1amygdalaright0.001amygdalaleft0.001	126	schizophrenia														
10	cerebrum1amygdala0.002amygdalaleft0.001	126	schizophrenia														
11	cerebrum1amygdala0.002amygdalaleft0.001	126	schizophrenia														
12	cerebrum1amygdala0.002amygdalaleft0.001	126	schizophrenia														
13	cerebrum1amygdala0.002amygdalaleft0.001	126	schizophrenia														
14	cerebrum1amygdala0.002amygdalaleft0.001	126	schizophrenia														

Making a Database

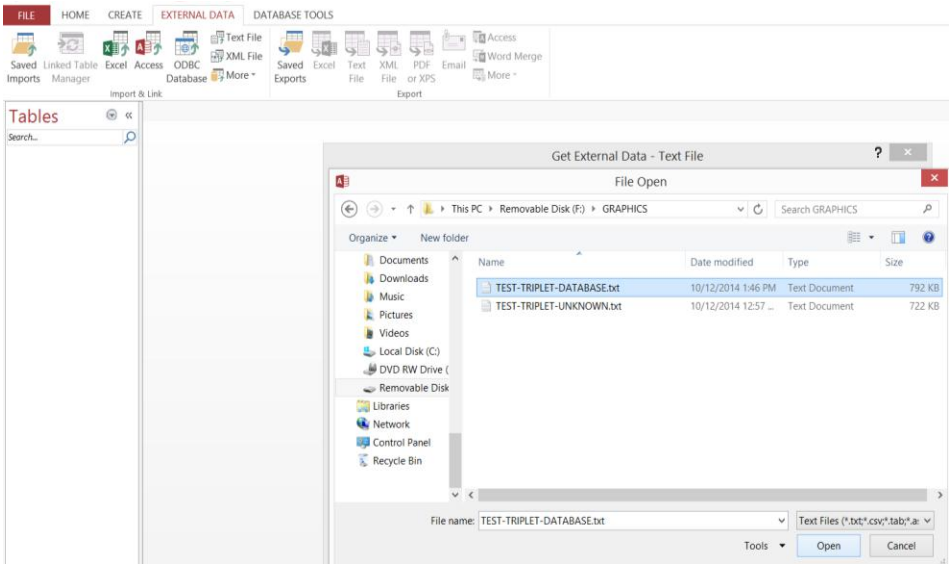


A text file created in Excel (TEST-TRIPLET-DATABASE.txt) becomes a database when imported into Access. To illustrate the process of creating a database and using it to diagnose an unknown, we will use this text file (cit nu=126) both for the database of disorders (schizophrenia) and for the unknown data (we will change schizophrenia to unknown). To do this, copy the text file (TEST-TRIPLET-DATABASE.txt) as TEST-TRIPLET-UNKNOWN.txt, open it, and replace schizophrenia with normal.

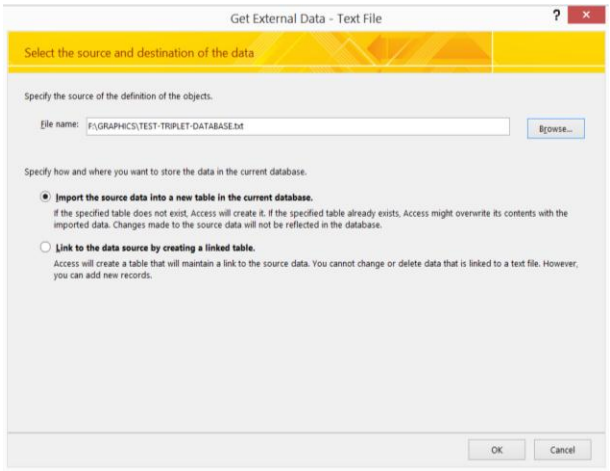
6First, we make the database. Run the Access database, select Blank desktop Database, name it **Test-Triplet-Database**, and click on Create. The following screen appears with an open, but empty table. Close the table.

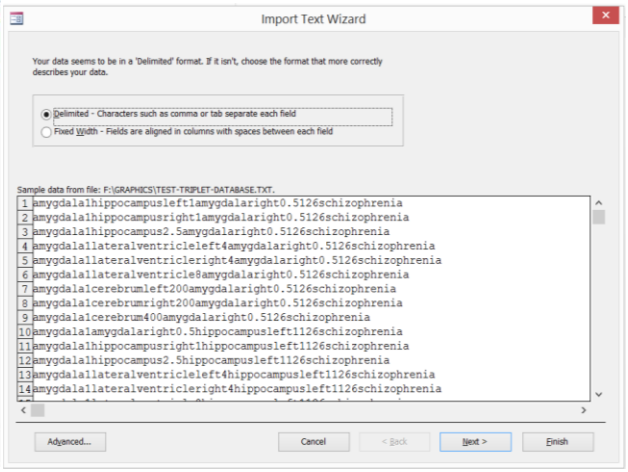


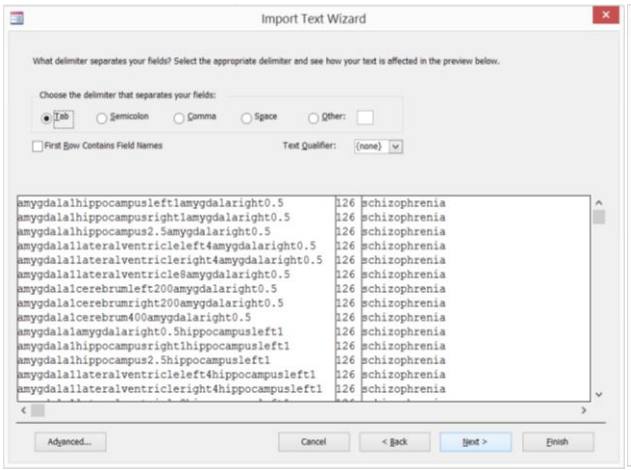
Click on EXTERNAL DATA. Select the text file for the database – TEST-TRIPLET-DATABASE.txt and click on Open.

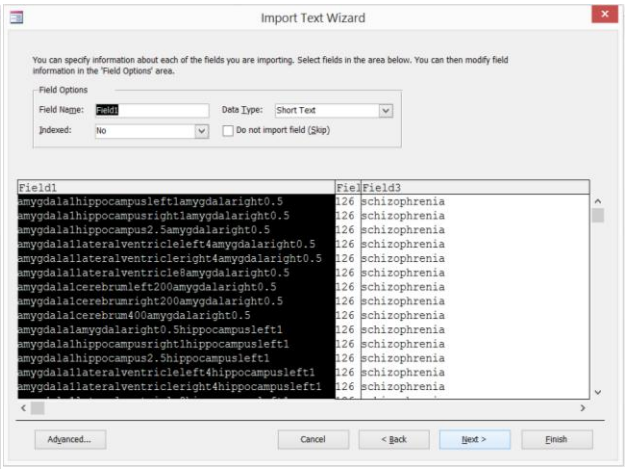


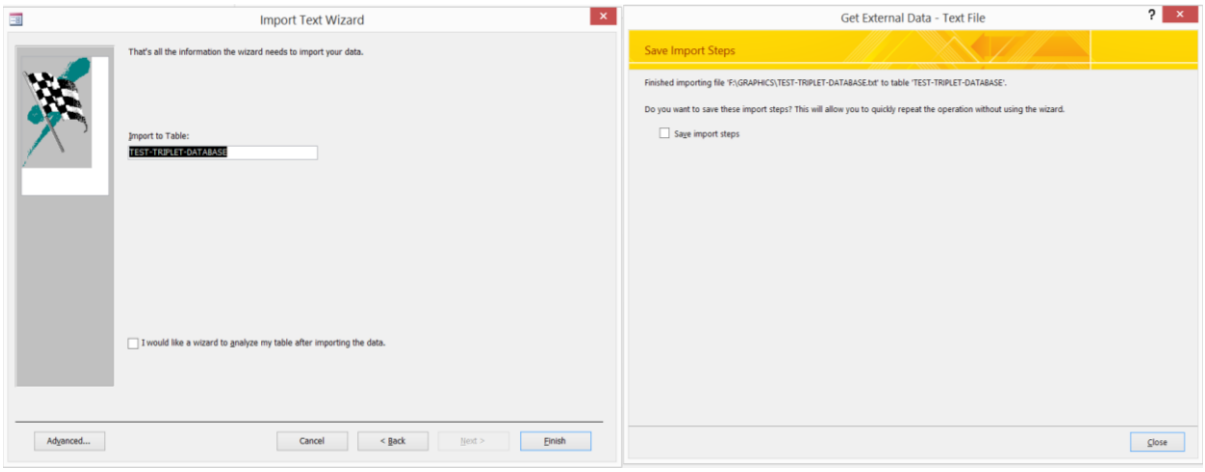
Select Import the source data into a new table in the current database. Duplicate the following screens.



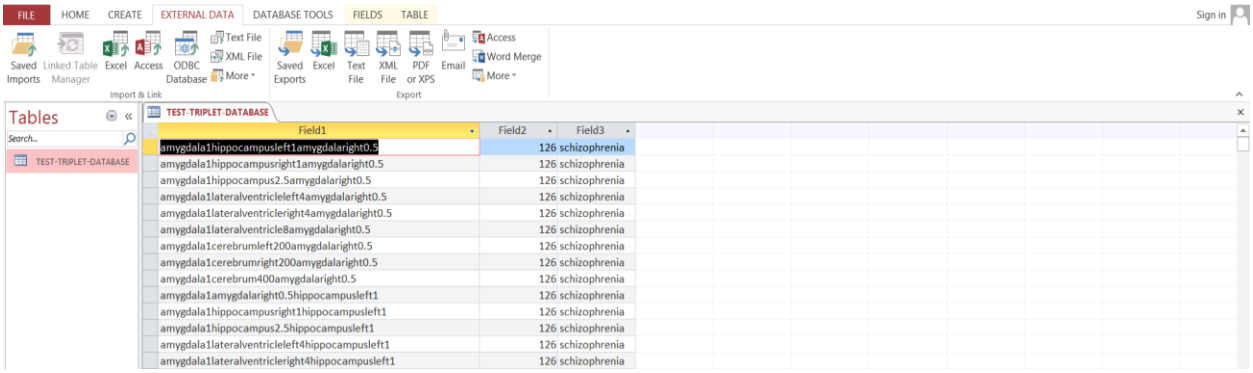




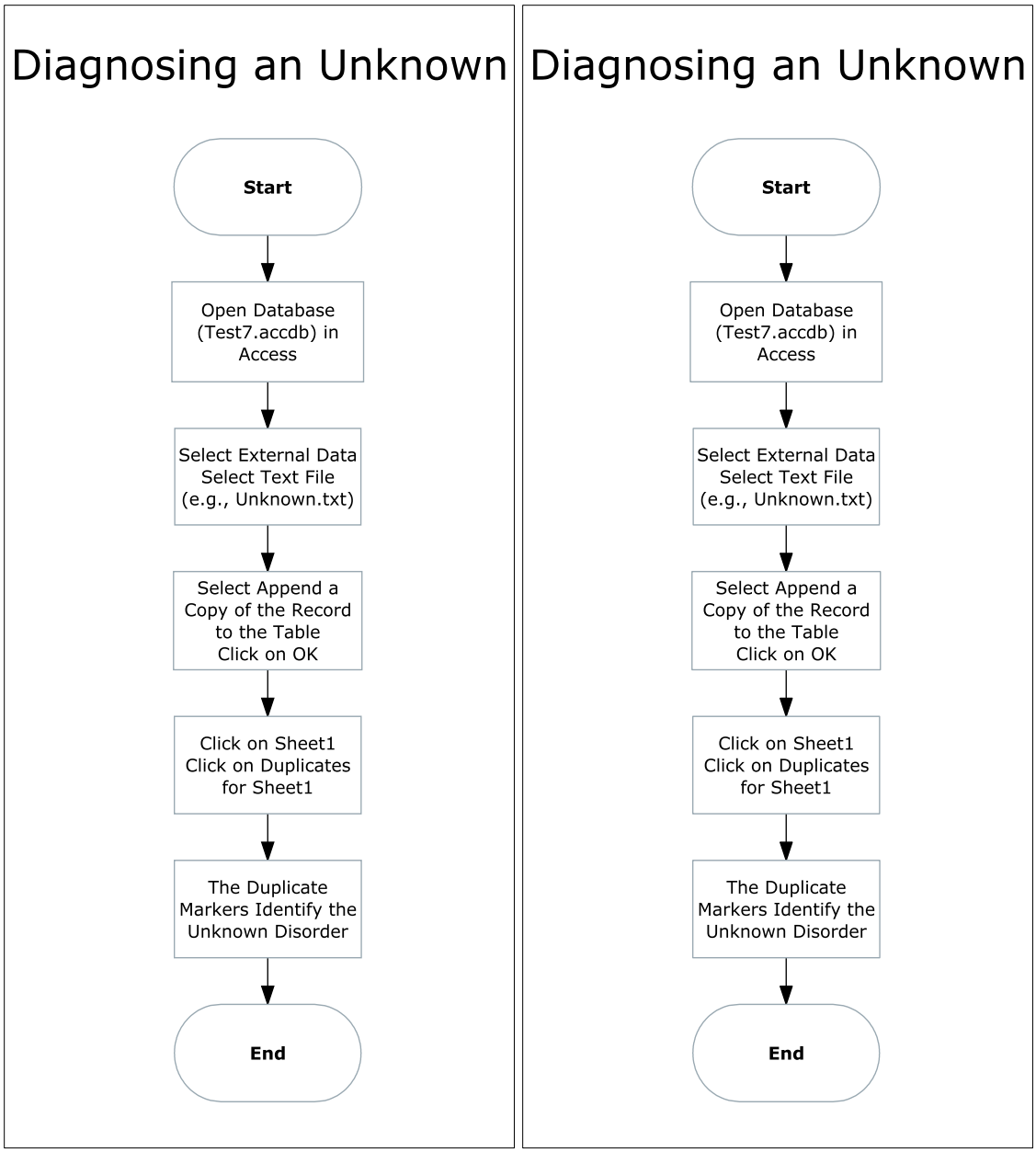




The result is a new Access database containing a file of known markers coming from publication 126 of the IBVD.

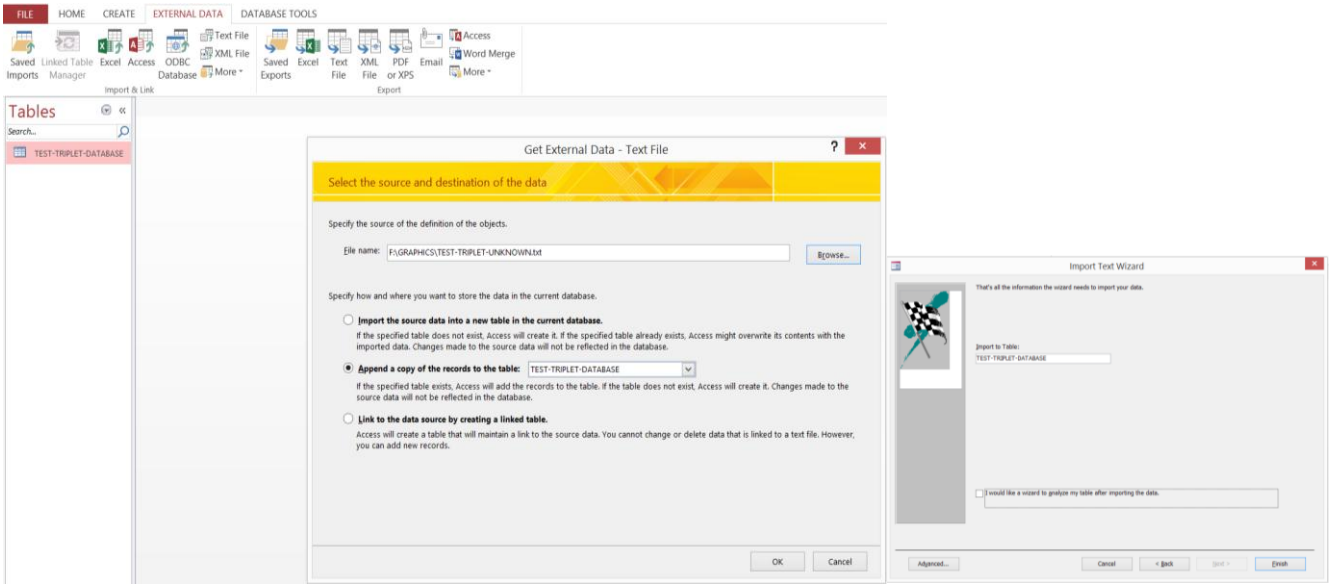


In the next example (Diagnosing an Unknown), we will import the unknown file (TEST-TRIPLET-UNKNOWN.txt), and run it against the TEST TRIPLET DATABASE.

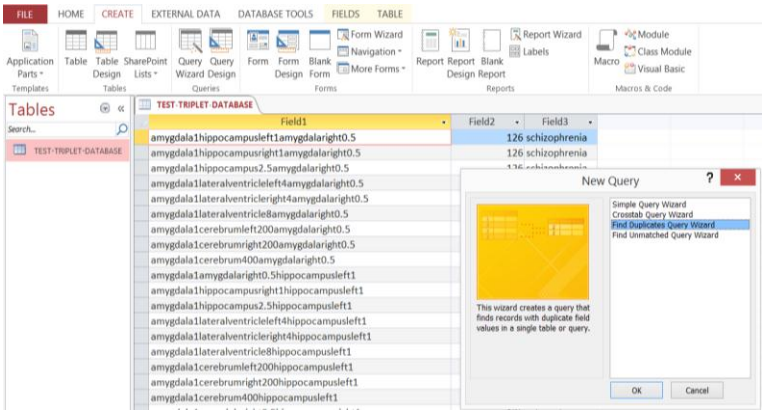


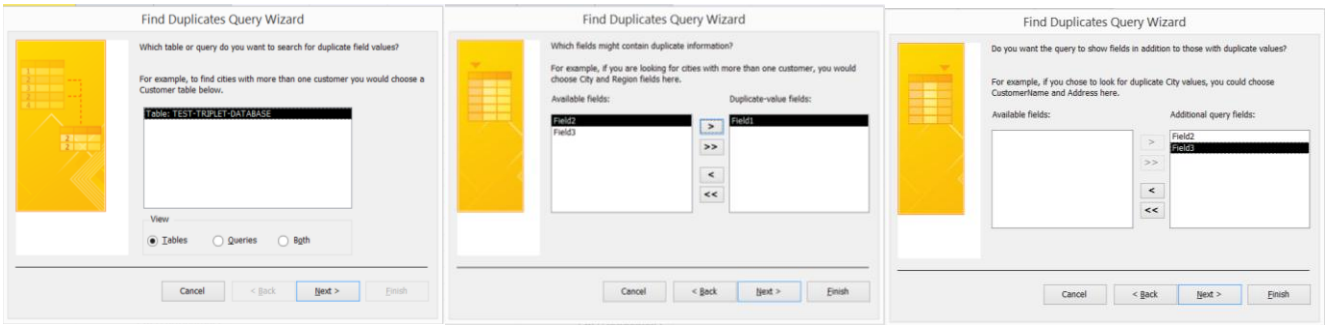
The algorithm on the left describes the ongoing example, whereas the one on the right uses the complete diagnosis database. In the software package, the Test7 database becomes MRI_T-DIAG-100.accdb.

With the TEST-TRIPLETS-DATABASE.accdb open and all the tables colosed, select EXTERNAL DATA and then Text File. When the screen below appears, click on the **Browse…** button, find the file < TEST-TRIPLETS-UNKNOWN.txt> and select **Append a copy of the records to the table:** TEST-TRIPLETS-DATABASE. Continue clicking on the **Next** button until the Import Wizard is finished.

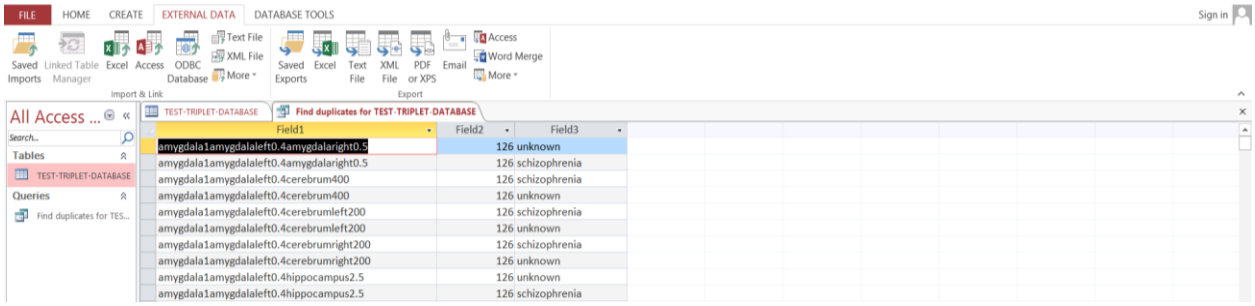


At this point, the database contains markers for both the knowns and unknowns. To find out the disorder associated with the unknows markers, select CREATE and click on Query Wizard. When the New Query screen appears, select **Find Duplicates Query Wizard**. Duplicate the screens as shown below and click **Finish**.



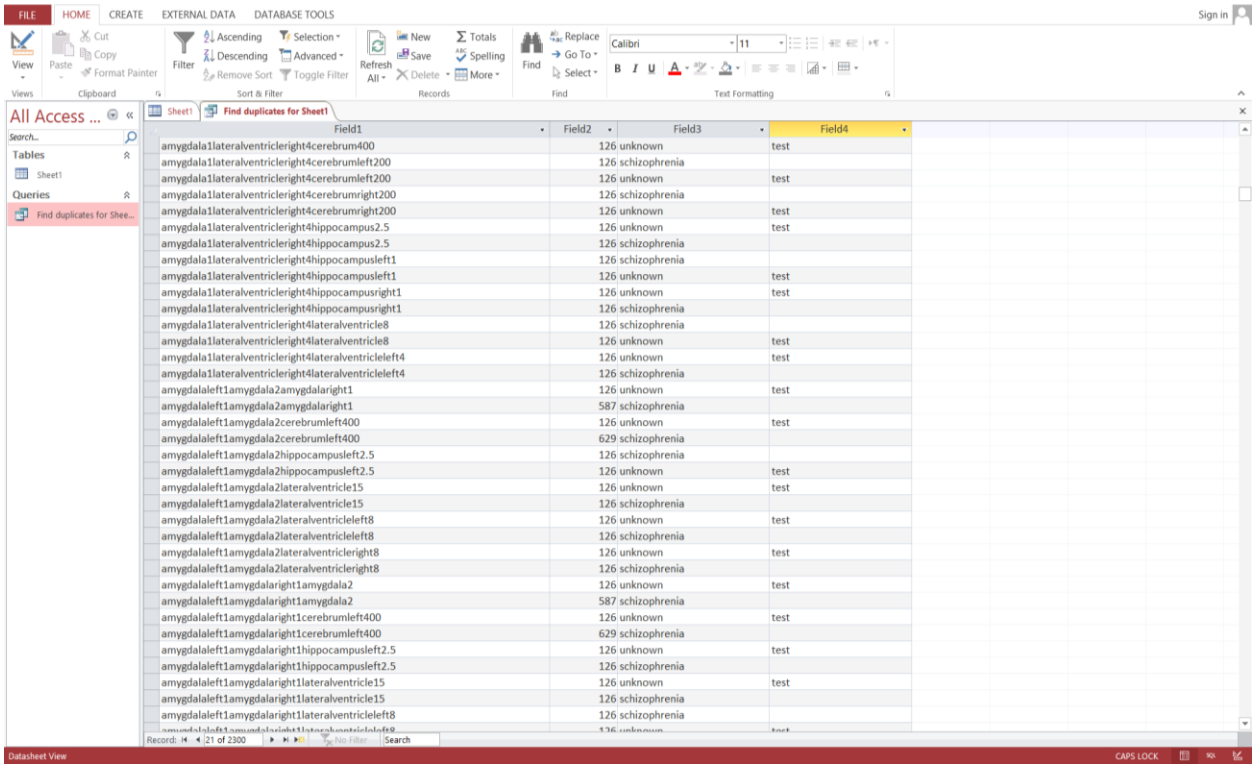


The duplicates reveal that the unknown markers come from patients with schizophrenia – the correct diagnosis.



Field1	Field2	Field3
amygdala1amygdalaleft0.4amygdalaright0.5	126	unknown
amygdala1amygdalaleft0.4amygdalaright0.5	126	schizophrenia
amygdala1amygdalaleft0.4cerebrum400	126	schizophrenia
amygdala1amygdalaleft0.4cerebrum400	126	unknown
amygdala1amygdalaleft0.4cerebrumleft200	126	schizophrenia
amygdala1amygdalaleft0.4cerebrumleft200	126	unknown
amygdala1amygdalaleft0.4cerebrumright200	126	schizophrenia
amygdala1amygdalaleft0.4cerebrumright200	126	unknown
amygdala1amygdalaleft0.4hippocampus2.5	126	unknown
amygdala1amygdalaleft0.4hippocampus2.5	126	schizophrenia

If instead, we run the unknown markers against the full diagnosis database, we also get the correct diagnosis of schizophrenia. Notice that the unknowns are also being detected correctly with markers coming from other papers (e,g., 587 and 629).



Field1	Field2	Field3	Field4
amygdala1lateralventricleright4cerebrum400	126	unknown	test
amygdala1lateralventricleright4cerebrumleft200	126	schizophrenia	
amygdala1lateralventricleright4cerebrumleft200	126	unknown	test
amygdala1lateralventricleright4cerebrumright200	126	schizophrenia	
amygdala1lateralventricleright4cerebrumright200	126	unknown	test
amygdala1lateralventricleright4hippocampus2.5	126	unknown	test
amygdala1lateralventricleright4hippocampus2.5	126	schizophrenia	
amygdala1lateralventricleright4hippocampusleft1	126	schizophrenia	
amygdala1lateralventricleright4hippocampusleft1	126	unknown	test
amygdala1lateralventricleright4hippocampusright1	126	unknown	test
amygdala1lateralventricleright4hippocampusright1	126	schizophrenia	
amygdala1lateralventricleright4lateralventricle8	126	schizophrenia	
amygdala1lateralventricleright4lateralventricle8	126	unknown	test
amygdala1lateralventricleright4lateralventricleleft4	126	unknown	test
amygdala1lateralventricleright4lateralventricleleft4	126	schizophrenia	
amygdalaleft1amygdala2amygdalaright1	126	unknown	test
amygdalaleft1amygdala2amygdalaright1	587	schizophrenia	
amygdalaleft1amygdala2cerebrumleft400	126	unknown	test
amygdalaleft1amygdala2cerebrumleft400	629	schizophrenia	
amygdalaleft1amygdala2hippocampusleft2.5	126	schizophrenia	
amygdalaleft1amygdala2hippocampusleft2.5	126	unknown	test
amygdalaleft1amygdala2lateralventricle15	126	unknown	test
amygdalaleft1amygdala2lateralventricle15	126	schizophrenia	
amygdalaleft1amygdala2lateralventricleleft8	126	unknown	test
amygdalaleft1amygdala2lateralventricleleft8	126	schizophrenia	
amygdalaleft1amygdala2lateralventricleright8	126	unknown	test
amygdalaleft1amygdala2lateralventricleright8	126	schizophrenia	
amygdalaleft1amygdalaright1amygdala2	126	unknown	test
amygdalaleft1amygdalaright1amygdala2	587	schizophrenia	
amygdalaleft1amygdalaright1cerebrumleft400	126	unknown	test
amygdalaleft1amygdalaright1cerebrumleft400	629	schizophrenia	
amygdalaleft1amygdalaright1hippocampusleft2.5	126	unknown	test
amygdalaleft1amygdalaright1hippocampusleft2.5	126	schizophrenia	
amygdalaleft1amygdalaright1lateralventricle15	126	unknown	test
amygdalaleft1amygdalaright1lateralventricle15	126	schizophrenia	
amygdalaleft1amygdalaright1lateralventricleleft8	126	schizophrenia	

APPENDIX III

Theory of Biological Complexity

The overarching principle of the new theory is that it takes a complexity to solve a complexity. This means that to test the theory empirically we need to construct a parallel complexity as close to the original as possible, relying exclusively on the rules that exist first in biology and then mirrored in our complexity. The sampling methods of stereology play an essential role in this building process by providing access to unbiased data and by supplying equations that can estimate and connect the data of a parallel complexity.

Complexity is an unfamiliar place. New rules apply, our perceptions change, and we get to ask and answer questions differently. The first order of business is to learn the rules of the game, which in science consists of developing a new theory structure. This represents an ongoing process wherein the theory evolves in step with the discovery process.

Recall that the fundamental building blocks of a biological complexity include parts and connections. Volumes, surfaces, lengths, or numbers define the parts quantitatively and ratios derived therefrom the connections. From this simple beginning, the complexity of an organism grows as the parts and connections cascade throughout the hierarchical levels of an organism. Since everything consists of the same basic building blocks and all the blocks are connected, our parallel complexity begins to resemble the original biology – at least on a limited scale. Testing the theory consists of looking for persistent patterns - locally and globally – and then using these patterns to define the rules of the game.

A collection of working lists, including Goals, Requirements, Basic Principles and Definitions, Derivatives, and Rationale summarize recent progress in constructing this new theory structure.

Theory of Biological Complexity: In its simplest form, the theory states that it takes a complexity to solve a complexity. We can define a biological complexity mathematically as a distinct set of elements (parts and connections) that combine to form patterns (e.g., mathematical markers) capable of scaling at both local and global levels. Typically, biology displays its complexity as a stoichiometry based on the ratios of it parts. Biology uses this simple rule to create both order and disorder. We define a rule as a mathematical pattern that persists at both local and global levels.

Theory Structure: The accompanying theory structure includes a current set of guidelines for exploring biology as a complexity. **Items highlighted in red identify recent additions, whereas items highlighted in green identify areas of notable progress.**

Goals

- **Generalize the data of the biology literature.**
- **Define and assemble a data-driven approach to the basic and clinical sciences.**
- **Identify mathematical patterns in biology.**
- **Explore biology as a rule-based system.**
- **Use published data to create a parallel complexity using rules intrinsic to biology.**
- **Remove postmortem distortions by harmonizing pre and postmortem data.**
- **Offer alternatives to the misguided practice of comparing concentrations in a biological setting.**
- **Demonstrate the effectiveness of a new approach to problem solving based on empirical data and guided by the rules of biology.**
- **Develop software that can accelerate productivity by transforming biological data into problem-solving tools.**
- **Capture biological phenotypes mathematically and use them to diagnose and predict outcomes.**
- **Evaluate current methods in the basic and clinical sciences.**
- **Assemble a diagnostic platform from the biology literature that can provide the correct diagnosis 100% of the time.**
- **Advance the technology surrounding mathematical markers from small to big data.**

- **Figure out how to extract meaningful patterns from large data sets.**
- Identify algorithms that biology uses to create disorders of the brain.
- Connect phenotypes to genotypes.
- **Optimize outcomes.**

Requirements

- Collect data with unbiased sampling methods.
- Express data as volumes, surfaces, length, or numbers. Concentration data formed from these and other parameters are subject to specific rules and limitations (See earlier reports).
- Assemble data as connected sets.
- Integrate data within and across hierarchical levels.
- Use a common format to organize and generalize data.
- Configure data to accommodate local and global patterns simultaneously.
- Operate within the bounds of a complexity parallel to the one of biology.
- Correct the volume distortions associated with postmortem data.
- Reconfigure data sets to enhance diagnostic and predictive properties.
- **Define the outputs of a database by applying filters.**
- Store and distribute data in digital form.
- Encourage open access to data.

Basic Principles and Definitions

- A biological complexity consists of parts and connections distributed hierarchically.
- Complexities can be both local and global.
- A biological complexity can unfold into smaller patterns or fold into larger ones.
- Parts and connections define the organizational framework of biology as distinct patterns. As such, they represent a rule-based management system.
- A parallel complexity represents a data-driven construct designed specifically to capture biological complexity quantitatively.
- Ratios and derivatives thereof (i.e., mathematical markers) serve as the basic units of information in a parallel complexity.
- Mathematical markers include parts (names) and connections (ratios).
- A second complexity exists in the postmortem data of biological stereology, produced by the methods of specimen preparation and data collection.
- Parts display quantitative (volume, surface, length, number) and qualitative properties (names, locations).
- All parts are connected or connectable by forming ratios.
- A ratio defines the relationship of one part to another. Moreover, ratios define nested and modular sets of connections within and across hierarchical levels.
- Parts and connections form patterns that scale in size, beginning with a ratio of two parts and ending with a ratio of n parts - where n would represent an entire organism.
- Patterns captured as mathematical markers increase their specificity as the number of parts and connections in the marker increase.
- In living subjects, mathematical markers routinely detect the same patterns (e.g., markers) locally and globally.
- In postmortem subjects, mathematical markers can detect the same local and global patterns, but only when correction factors for volume distortions are applied.
- **Prediction in complex living systems requires interactions with parallel complexities capable of producing a correct diagnosis 100% of the time.**
- Valances describe the ability of the same set of parts to display different numerical ratios (connections). They reflect biological rules of stoichiometry.

Derivatives

A derivative includes - as a minimum - the names of two parts and their corresponding values formed into a ratio. In forming a ratio, the original published values may be used directly (repertoire value) or

converted to a decimal step (decimal repertoire value). Data pair ratios take the form X:Y, data triplets X:Y:Z, and data quadruplets X:Y:Z:Q. Mathematical markers add the names of the parts (A, B, C, D) to the ratio: AX:BY, AX:BY:CZ, and AX:BY:CZ:DQ.

Data Pairs

- A data pair consist of two parts (names) and two connections (ratios) expressed as repertoire and a decimal repertoire values. Data pairs can be formed by inspection or by taking all possible permutations of the names of the two parts – to which numerical values are assigned.
  - Data pair values – expressed as a decimal step (decimal repertoire value) – combine with names to form mathematical markers.
  - A data pair can use data before or after corrections are applied for the volume distortions of postmortem material.
  - Data pairs display valences in that the same two parts can occur in different proportions.

Data Triplets

- A data triplet consists of three parts and three connections with the ratios expressed as repertoire and decimal repertoire values. Triplets are formed by inspection or by taking all possible permutations of the three names of the parts – to which numerical values are assigned. Mathematical markers use decimal repertoire values.
  - A data triplet can use data before or after corrections are applied for the volume distortions of postmortem material.
  - Triplets display valences in that the same three parts can occur in different proportions.

Data Quadruplets

- A data quadruplet consists of four parts and four connections with the ratios expressed as repertoire and decimal repertoire values. Quadruplets are formed by inspection or by taking all possible permutations of the four names of the parts – to which numerical values are assigned. Mathematical markers use decimal repertoire values.
  - A data quadruplet can use data before or after corrections are applied for the volume distortions of postmortem material.
  - Quadruplets display valences in that the same four parts can occur in different proportions.

Properties of Data Pairs, Triplets, and Quadruplets

- Data pairs, triplets, and quadruplets form both general and diagnostic patterns that can be unique or shared.
- Conservation of patterns occurs within and across animal species.
- All patterns and their antecedents can be stored in a single database table.
- Mathematical markers – as a universal data set - offer a general solution to the problem of biological complexity.
- Mathematical markers can detect the distorted volumes of postmortem brains.
- **The sensitivity of mathematical markers increases by adding variables.**

Rationale

- Complexity theory represents a long overdue response to the limitations of our current theory structure based on reductionism.
- Reductionist theory takes biology apart, studies parts in isolation, and applies statistical tests to detect changes. It purports to simplify biology, but instead adds a second complexity, often making reliable interpretations difficult to impossible. This second complexity includes a wide range of distortions caused by death and by the methods of specimen preparation and data collection. Concentrations, which are the most common form of biological data, often fail to detect biological changes accurately because they ignore complexity. In a biological setting, comparing concentrations involves four variables not two – a fact largely unknown to biologists. Hierarchy equations, which are used to convert concentrations into absolute values, can be expected to fail when the variables used to evaluate the equations carry volume distortions.

- The methods of reductionist theory minimize the effectiveness of published data, obscure biological patterns, and substitute reproducibility and significant differences for accuracy. By corrupting biological data, such methods actively impede learning, discovery, and innovation.
- Quantifying biology in the absence of a theory structure consistent with biological complexity will not turn biology into a quantitative science.
- Complexity theory addresses many of the limitations imposed by reductionism, while adding a host of new capabilities. A principal argument for studying biology within the framework of complexity theory is that it simplifies everything and provides a tent large enough to accommodate all parts of the biology enterprise.
    - Absolute values can be estimated without hierarchy equations.
    - Mathematical markers transform old forms of biological data into new patterns consistent with complexity.
    - All mathematical markers can be stored in a single database table, searched for patterns, and used directly for problem solving.
    - By defining phenotypes robustly, mathematical markers support diagnosis and prediction.
    - Quantitative phenotypes can provide mathematical pathways to and from the genome.
    - **Mathematical markers can detect the algorithms biology uses to define itself in health and disease.**
    - Biological patterns exist both locally and globally.
    - Global patterns lead to generalizations and rules.
    - The biology literature can supply the large, integrated data sets fundamental to complexity theory.
    - Forming data ratios (data pairs, triplets, and quadruplets) helps to minimize bias.
    - Outcomes can be subjected to rigorous testing.
    - New data formats capture the complexity of biology as patterns.
    - Patterns can provide multiple solutions to the same problem.
    - Data distortions can be identified and corrected.
    - **Prediction in biology relies importantly on diagnosis.**
    - **Parallel complexities consisting of unique markers can diagnose outcomes correctly 100% of the time – in well-defined settings.**