# An Autism Case History to Review the Systematic Analysis of Large-Scale Data to Refine the Diagnosis and Treatment of Neuropsychiatric Disorders

Isaac S. Kohane

## ABSTRACT

Analysis of large-scale systems of biomedical data provides a perspective on neuropsychiatric disease that may be otherwise elusive. Described here is an analysis of three large-scale systems of data from autism spectrum disorder (ASD) and of ASD research as an exemplar of what might be achieved from study of such data. First is the biomedical literature that highlights the fact that there are two very successful but quite separate research communities and findings pertaining to genetics and the molecular biology of ASD. There are those studies positing ASD causes that are related to immunological dysregulation and those related to disorders of synaptic function and neuronal connectivity. Second is the emerging use of electronic health record systems and other large clinical databases that allow the data acquired during the course of care to be used to identify distinct subpopulations, clinical trajectories, and pathophysiological substructures of ASD. These systems reveal subsets of patients with distinct clinical trajectories, some of which are immunologically related and others which follow pathologies conventionally thought of as neurological. The third is genome-wide genomic and transcriptomic analyses which show molecular pathways that overlap neurological and immunological mechanisms. The convergence of these three large-scale data perspectives illustrates the scientific leverage that large-scale data analyses can provide in guiding researchers in an approach to the diagnosis of neuropsychiatric disease that is inclusive and comprehensive.

***Keywords:*** Autism, Data Science, Electronic Health Records, Neuroimmunology, Neuropsychiatry, Synaptic Function

http://dx.doi.org/10.1016/j.biopsych.2014.05.024

Perhaps the branch of medicine most successful in achieving a precise diagnosis of disease, one directly linked to its cause, has been that of infectious disease. Only a little over 100 years passed between the identification of microorganisms as the causative agents for multiple diseases and the consequent development of dozens of therapies in immunizations and antibiotics that have had a greater impact on mortality and morbidity than any other medical intervention (1). It is this understanding of the consequences of cause and precise diagnostic capabilities that were the main drivers of the recent National Academy of Sciences report on Precision Medicine: to use multiple comprehensive measurement modalities to identify which subgroup of patients a given patient most resembles and therefore to be able to both assign a diagnostic label and predict a clinical course in response to therapeutic intervention. I review here how a systematic approach to large-scale data can make some preliminary and illuminating strides toward a "precision medicine" of neuropsychiatric disease. I use the autism spectrum disorders (ASDs) as the prismatic example of the larger opportunity by illustrating how this approach reveals two richly productive but largely separate avenues of research in ASD defined by apparently distinct mechanistic hypotheses, that is, ASD as a disorder of neural connectivity and specifically synaptic connectivity regulation (2,3) and ASD as a disorder of immunological signaling (4–6).

First, some framing is required regarding the task being addressed: diagnosis of the disorder. Here, diagnosis of ASD will be defined in the probabilistic framework used in decision making: the probability of a disease, $D$, given the findings $F$ summarized by the notation $p(D|F)$. In ASD, we often attempt to diagnose or rule out a single disease (i.e., autism), even though it is recognized that there are likely to be multiple diseases (i.e., the set of diseases $D$ composed of $\{D_1....D_n\}$ that together constitute ASD). A diagnosis will be more useful to the extent that $p(D|F)$ is high (i.e., close to 1.0) corresponding to the high likelihood of disease or low (i.e., close to 0.0) corresponding to the low likelihood of disease. Further confidence in this likelihood estimate is provided if the error of this estimate is low. The appropriateness of therapy can then be determined by how well it is matched to the disease. This thereby highlights the value of determining which of the diseases that constitute ASD of the set $\{D_1....D_n\}$ have the highest probability as each therapy will have different efficacy for each of them.

## PUBLISHED LITERATURE FOR LARGE-SCALE CHARACTERIZATION OF RESEARCH

In the recently published DSM-5, ASD is defined as including persistent deficits in social communication and social

SEE COMMENTARY ON PAGE 3

ISSN: 0006-3223

interaction and restricted, repetitive patterns of behavior, interests, or activities. This new single disorder replaces several previously defined disorders including autistic disorder, Asperger's disorder, and pervasive developmental disorder not otherwise specified. This redefinition will surely lead to a change in diagnosis for many individuals and possibly a change in funding of support services. The controversy that emerged prior to and after this publication illustrates the challenge posed by the diagnostic and prognostic tasks when applied to a disease complex that many recognize to be a constellation of heterogeneous pathophysiologies (5,7–12), some of which have genetic causes and some environmental or a combination thereof. A multidimensional characterization of the patient population of interest, which measures the multiple genetic, molecular, clinical, and environmental exposure features of each patient to derive the overall landscape of the constellation of heterogeneous diseases that distinguish that population, provides the most comprehensive and systematic viewpoint (13). Of course, such integrative data sets are currently far and few between, with the Simon's Simplex Collection (14) constituting a notable example of what such integration can yield (and the effort and investments required to bring it together). With the steady accretion of clinical and research data sets, we can anticipate such multidimensional assessment to grow. Therefore, it will become essential to determine which of the set of diseases comprising ASD in $\{D_1,...D_n\}$ are being diagnostically evaluated. Merely making this determination of which diseases are being considered as part of the ASD set is challenging. This challenge is best illustrated by a large-scale database available to all ASD researchers: that of the published literature. If we focus on those recent publications that were supported by the U.S. National Institutes of Health (NIH) and therefore deposited in the PubMed Central Open Access Subset repository (15), then, as illustrated in Figure 1, not only is the primary literature balkanized, but even the citations made by the authors of this literature largely address disparate domains of biology. If we label the autism and genetics literature as pertaining nonexclusively to four sets: neuronal synaptic function (*N*) and immunological function/disorders (*I*), with *N cit*. and *I cit*. denoting the literature cited by these first two sets, then as shown in Figure 1, the overlap is remarkably slight. For example, of 290 publications in *N*, only 18 are also in *I*, and of the 12,391 cited by the publications in *N*, only 1551 are cited by *I*. At best, this suggests that either the set of findings or the set of diseases considered in developing a precision diagnosis of the ASDs is incomplete, depending on which research community is addressed. This raises the question of what population studies can reveal regarding this apparent dichotomy. By way of example, large-scale population genomics have revealed previously poorly defined or unsuspected subtypes of disease within breast cancer (16), non-small-cell lung carcinomas (17), and leukemia (18). However preceding the advent of genomics by more than a century, physician-scientists have used observational studies to define disease subtypes. Jean Martin Charcot, for example, systematically and comprehensively studied the patients in a large neurological hospital in Paris and was thereby able to define new and lasting disease entities out of a pool of previously monolithic and broad neurological diagnoses (19). A century and a half after Charcot, can we undertake large-scale observational studies of patients enabled by the recent acceleration in electronic health record systems deployment to augment our ability to generate an integrated view of $p(D|F)$ for ASD?

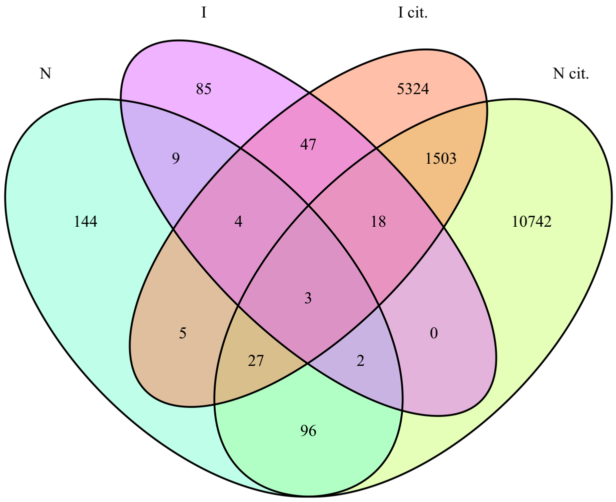

**Figure 1.** Illustration of the incomplete overlap in research of ASD genetics based on investigations of synapses and research in ASD genetics based on investigations of the immune system. Values indicate number of publications in that category. Four ellipses are shown corresponding to four corpora all selected from PubMed Central. ASD, autism spectrum disorder; I, publications focused on genetics and immune system; I cit., publications cited in I; N, publications focused on genetics and synapses; N cit., publications cited in N (the intersection between N and I accounts for only 4% of the combined publications, and the intersection between N cit. and I cit. accounts for only 8% of the combined citations).

## ELECTRONIC HEALTH RECORDS FOR LARGE-SCALE CHARACTERIZATIONS

Acceleration of the adoption of electronic health records (EHR) in clinical care through the HITECH Act of 2009 (20) may or may not increase the productivity or safety of healthcare delivery, but it certainly has provided a large source of detailed clinical documentation of patients. This enables researchers adept in the "secondary use" of EHR data to identify patients with the clinical phenotype of interest and then use the samples acquired in subsequent visits for clinical diagnostics for the purposes of genotyping and resequencing and even epigenetic characterization [reviewed in (21,22)]. In addition to structured or codified data (e.g., laboratory test, medications, and diagnostic and procedure billing codes), the development of "natural language processing" (NLP) techniques (23–27) enables the narrative text of clinical notes to be mined for a far more accurate phenotypic assessment of the patients than from codified data. Given that codified billing data are well known to be biased for reimbursement and insufficiently fine grained, this is not surprising. However, when codified data are combined with NLP-derived data, the phenotyping accuracy is higher than with either clinical source alone (22). Furthermore, this automated phenotyping has been shown to be generalizable, portable, and reproducible across healthcare systems (28,29). These very encouraging early studies