King's Research Portal

DOI:

[10.1017/S0033291717001726](https://doi.org/10.1017/S0033291717001726)

Document Version

Peer reviewed version

[Link to publication record in King's Research Portal](#)

Citation for published version (APA):

Joyce, D. W., Tracy, D. K., & Shergill, S. S. (2018). Are we failing clinical trials? A case for strong aggregate outcomes. *Psychological Medicine*, 1-10. <https://doi.org/10.1017/S0033291717001726>

Citing this paper

Please note that where the full-text provided on King's Research Portal is the Author Accepted Manuscript or Post-Print version this may differ from the final Published version. If citing, it is advised that you check and use the publisher's definitive version for pagination, volume/issue, and date of publication details. And where the final published version is provided on the Research Portal, if citing you are again advised to check the publisher's website for any subsequent corrections.

General rights

Copyright and moral rights for the publications made accessible in the Research Portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognize and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the Research Portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the Research Portal

Take down policy

If you believe that this document breaches copyright please contact librarypure@kcl.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.

Are we failing clinical trials? A case for strong aggregate outcomes.

Dan W. Joyce¹, Derek K. Tracy^{1,2}, Sukhwinder S. Shergill¹.

Word Count : 4501

Affiliations :

1 - Cognition Schizophrenia and Imaging Laboratory,

Department of Psychosis Studies, PO63

Institute of Psychiatry, Psychology and Neuroscience

King's College London

De Crespigny Park

London SE5 8AF

2 – Oxleas NHS Foundation Trust, London, UK

Corresponding Author :

Dan W. Joyce

danjoyce@doctors.org.uk

Acknowledgements :

The authors gratefully acknowledge data used in the preparation of this article that resides in the NIH-supported NIMH Data Repositories (Clinical Antipsychotic Trials of Intervention Effectiveness).

Conflicts of Interests :

None

Funding :

DWJ is funded by a National Institute of Health Research Integrated Academic Training Clinical Lectureship. S.S.S. is supported by a European Research Council Consolidator Award (grant number 311686) and the National Institute for Health Research (NIHR) Mental Health Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King's College London. The funders had no role in study design, data collection, data analysis, data interpretation, or writing of the report.

Abstract

Clinical trials in psychiatry inherit methods for design and statistical analysis from evidence-based medicine. However, trials in other clinical disciplines benefit from a more specific relationship between instruments that measure disease state (e.g. biomarkers, clinical signs), the underlying pathology and diagnosis such that primary outcomes can be readily defined. Trials in psychiatry use diagnosis (i.e. a categorical label for a syndrome) as a proxy for the underlying disorder, and outcomes are defined, for example, as a percentage change in a univariate *total score* on some clinical instrument. We label this approach to defining outcomes *weak aggregation* of disease state. Univariate measures are necessary, because statistical methodology is both tractable and well-developed for scalar outcomes, but we show that weak aggregate approaches do not capture disease state sufficiently, potentially leading to loss of information about response to intervention. We demonstrate how multivariate disease state can be captured using geometric concepts of spaces defined over routine clinical instruments, and show how clinically meaningful disease states (e.g. representing different profiles of symptoms, recovery or remission) can be defined as prototypes (geometric locations) in these spaces. Then, we show how to derive univariate (scalar) measures which capture patient's relationships to these prototypes and argue these represent *strong aggregates* of disease state that may be a better basis for outcome measures. We demonstrate our proposal using a large publically available dataset. We conclude by discussing the impact of strong aggregates for analyses in traditional and novel trial designs.

1. Introduction

A clinician from any discipline selects a treatment for a patient based on evidence from clinical trials. The clinician applies the evidence based on the assumption that the patient has a given disease and that available treatments produce an outcome – response, remission or failure to respond - for that disease. We will argue that currently, much of the clinical trial evidence in psychiatry relies on the assumption that diagnosis is an adequate proxy for a disease or disorder and this leads us to use an inappropriate model of outcome (Joyce et al. 2017). This results in evidence that informs us only of the *average* response for a group of patients presumed to be homogenous with respect to their categorical diagnosis. This may also explain the limited changes in prescribing practices after the publication of large trials (Berkowitz et al. 2012).

Figure 1-1 describes a model of the relationship between trial outcomes and the underlying disorder; a concrete example being chronic systemic inflammatory disorders such as rheumatoid arthritis, where a disease process (DP: autoimmune-mediated inflammation) is reflected in a disease state (S: pain symptoms, inflammatory changes in joints, biochemical changes) that can be quantified by instruments (Y: pain and activity function scales; serological erythrocyte sedimentation rate, rheumatoid factor and anti-cyclic citrullinated peptide; radiological evidence of joint changes) and for which outcomes can be defined as changes in those instruments (Z: differences in pain and function scales, changes in serological markers and reduction in joint injury). When a patient is treated, disease states (S) change, and if the instruments (Y) are sensitive to these changes, they can be subjected to statistical methods that establish treatment efficacy (response, or failure to respond) by e.g. defining thresholds on Z, and identifying which patient-specific factors, X, mediate response.

In the idealised model shown in Figure 1-1, for a given disorder or disease process (DP) there will be a disease state (S) that corresponds with that disorder – but not necessarily in a one-one relationship. The discovery of disease states and instruments that have predictive power to identify diagnoses is the domain of biomarker and psychometric research – see for example, (Marquand et al. 2016) for discussion of statistical methodology. Here, we require that one or more instruments (Y) quantify variables of the disease state (S) at a given time and we take this collection of variables to be a vector identifying a location in a multidimensional space -- discussed below, and see also (Joyce et al. 2017) for a more detailed discussion. It is, however, common practice in clinical trials to *aggregate* the variables in an instrument, Y, to obtain e.g. a total ‘score’ for measuring the severity of the patient’s disease state at any given time (i.e. pre- and post-intervention). Finally, an outcome (Z) must then capture *changes* measured by instruments (Y) that are clinically meaningful. We recognise that the terms ‘disorder’, ‘disease’ and so forth can be contentious in mental health, but they are herein adopted for convention throughout this paper. Further, as the analogy above suggests, we adopt a position consistent with biological realism (Kendler 2016) regarding the nature of psychiatric disorders.

FIGURE 1 HERE

Psychiatry does not benefit from as clear a correspondence between disease state (S) and disease process (DP) nor are there instruments (Y) analogous to erythrocyte sedimentation rate, radiological evidence or joint changes from the example of rheumatoid disease. Consequently, psychiatry is faced with the model shown in Figure 1-2, where a diagnostic category (Dx) such as schizophrenia or bipolar affective disorder replaces disease process (DP).

In psychiatry, the instruments (Y) that measure disease state are multivariate scales that capture the severity of signs and symptoms – for example, in psychotic disorders, the positive and negative

symptoms scale (PANSS) or the brief psychiatric rating scale (BPRS) (Kay et al. 1987; Overall & Gorham 1962) . Psychiatric diagnoses represent constellations of signs and symptoms, but it is possible for these to overlap between diagnoses: for example psychotic features such as auditory verbal hallucinations are common to both schizophrenia, bipolar and major depressive disorder (Toh et al. 2015) and borderline personality disorder (Barnow et al. 2010; Schroeder et al. 2013; Glaser et al. 2010; Nishizono-Maher et al. 1993). In bipolar disorder and schizophrenia, there are similarities in non-verbal communication (Annen et al. 2012) , affective symptoms (Keshavan et al. 2011) and cognitive deficits (Jabben et al. 2009; Green 2006) .

There is also consensus, for example, that the diagnosis of schizophrenia is not a single disease process, but rather a categorical label for a syndrome with different aetiologies (Jablensky et al. 2006; Ripke et al. 2014; Reininghaus et al. 2016; Walker et al. 2002; Demjaha et al. 2009; Demjaha et al. 2012) and shared genetic risk factors (Lichtenstein et al. 2009; Craddock et al. 2009; Purcell et al. 2009) . There is progress in trying to parse diagnostic categories along phenotypes, endophenotypes, biomarkers and underlying cellular and molecular aetiologies, for example (Schumann et al. 2014; Morris & Cuthbert 2012; Insel et al. 2010; Cuthbert & Insel 2013).

Currently, clinical trials in psychiatry have to contend with a lack of clear relationship between disease state and process. Patients are therefore recruited into trials on the basis of their diagnostic category (Dx) and treatment efficacy is established based on usually dichotomous outcomes (Z), defined as threshold changes in aggregates of instruments (Y). For example, in mood disorders, remission of symptoms can be defined as the summed (total) Hamilton depression scale score of ≤ 7 for at least 2 months (Frank et al. 1991), and similarly for schizophrenia, a 50% reduction in baseline PANSS or BPRS score (Leucht et al. 2009; Jakubovski et al. 2015). In the language of

statistics, disease state and process (S and DP) are latent (or hidden, unmeasured) variables that are largely ignored.

We will show that the model exemplified in Figure 1-2 results in patients being assigned the *same* outcome (Z) if we define aggregates on instruments (Y) without attending to differences in disease states (S). For example, using the PANSS instrument, patients with high positive and low negative symptom severity risk being equated with patients who have low positive but high negative symptom severity. Given the uncertainty in relationships between disease states and processes, ignoring how disease states differ between patients means we are effectively failing to identify groups of patients that may benefit from an intervention. Just as importantly, we may also be subjecting patients to treatments and side-effects that are not effective for their specific manifestation of disorder.

Our proposal is that instead of assuming the model of Figure 1-2, adopting the model shown in Figure 1-1 allows the derivation of outcomes (Z) by directly attending to the concept of differing disease states as they are measured and represented by instruments (Y). For example, at a given time, the PANSS instrument measures 30 individual symptoms – a measure of disease state – which can change over time, for example, in response to intervention. We note that a single instrument may not be sufficient to capture disease state with enough fidelity to have analytical utility, for example, we may augment PANSS with some measure of affective symptoms or instruments measuring social and occupational functioning. This has the potential to expose treatments that are effective for *some* patients (e.g. those with a certain profile of positive, negative and general symptoms) and avoids measuring efficacy as the ‘average’ response for the homogenous diagnostic category.

2. Weak Aggregate Outcomes

To illustrate the inherent problems with current definitions of outcomes, consider the PANSS scale which measures 30 individual variables that if used to measure outcome are analytically intractable because clinical trial statistical methodology requires a *univariate* (one-dimensional or scalar) measure of change e.g. in response to treatment with respect to multiple predictors (i.e. patient specific factors, X, in Figure 1). Tractability is obtained by ‘collapsing’ these 30 variables into a single aggregate by summation and then the outcome measure, Z, summarises clinically meaningful change (e.g. response or remission) as a *threshold* change in this sum. We refer to this approach as *weak aggregation*. For example, in mood disorders, remission of symptoms can be defined as the summed (total) Hamilton depression scale score of ≤ 7 for at least 2 months (Frank et al. 1991) . In schizophrenia, there are proposals for ways of aggregating variables in a more structured way (Andreasen et al. 2005) and these represent thresholds on *selected combinations* of variables in scales that are believed to be clinically meaningful. However, clinical trials require a single primary outcome across the whole participant group, and secondary outcomes are used to measure subtle variation of response in sub-groups of the study population. Using a number of secondary outcomes carries the cost of making analysis vulnerable to criticisms of false positives, or Type I error. The most common approach to statistical analysis of this scalar outcome Z, is to use some variant of generalised linear modelling (GLM) against a set of predictors X (McCullagh & Nelder 1989) .

One consequence of the relationships described above, and in Figure 1, is that defining a primary outcome (Z) by thresholds on a *weak aggregates* of multiple variables (Y) ‘collapses’ information about patients’ disease state (S) into a single univariate, scalar value that may obscure important discriminating information that (optimistically) speaks to the disease process being treated by an intervention (as in the example given above of patients with opposing patterns of positive and

negative symptoms). This is especially problematic for psychiatry, where the correspondence of disease states to processes stands in a many-to-many relationship and we have traditionally used diagnostic category (Dx) as a proxy for both.

As a more detailed illustrative example, consider disease state measured by the three domains in the PANSS instrument: the total positive (P) and negative (N) symptoms scores range from 7 (no symptoms) to 49 (severe) and the general symptoms domain (G) ranges from 16 to 122. To derive an outcome Z, we consider the weak aggregate that is the sum of the total positive and negative symptoms, $Z = P + N$. It is obvious that there are many combinations of P and N (with each combination representing a discrete disease state) that could yield the same outcome value for Z. For example, a patient with $P = 23$ and $N = 44$ has $Z = 67$ whereas another patient with $P = 38$ and $N = 29$ has the same aggregate outcome $Z = 67$, despite these measurements representing quite different disease states; the first patient having high *negative* (but low positive) symptom severity and the second patient having the opposite pattern. Ignoring these differences – as the weak aggregate sum Z does – results in an outcome measure that cannot differentiate between disease states that may be clinically distinct and meaningful. Using this example, there are 32 combinations of values for P and N that yield the sum 67 and therefore, 32 disease states which would be assigned the *same* outcome for the weak aggregate $Z = P + N$. This problem becomes exponentially larger over three variables: defining the aggregate outcome measure over the positive, negative *and* general scales ($Z = P + N + G$) results in 741 discrete combinations of values of P, N and G that yield a total score of $Z = 67$. Although some combinations assigned $Z = 67$ will be clinically meaningful - for example, patients with $(P,N,G) = (18, 25, 24)$ and $(18, 26, 23)$ are suitably alike – in general, many will not. It is clear that *weak aggregation* blindly collapses variables from instruments such as PANSS to a single scalar variable, ignoring clear differences in disease state.

3. Strong Aggregate Outcomes

There is an additional problem with weak aggregation in psychiatry illustrated in Figure 1-2, when disease process and disease state are left un-modelled. In standard analyses using GLMs, by definition, it is only the mean change in the aggregate outcome (Z) that is modelled as a function of the predictors (X). All patients in the GLM analysis will be assumed to have effectively the *same* disorder that responds according to a unimodal, average response over the whole trial population: we know that response to an intervention is rarely uniform across patients with psychiatric disorders. A relevant analogy in fibromyalgia is given by (Moore et al. 2013; Moore et al. 2010) where response to treatment with pregabalin demonstrates a *bimodal* response; some patients have clinically significant reduction in pain but others show little or no response at all.

Our proposal for *strong aggregates* is as follows: firstly, to retain the strengths and tractability inherent in current statistical methodology (e.g. GLMs) we must find univariate measures – but avoid simply ‘collapsing’ measurements of disease state into a single scalar which may not expose or reflect clinically meaningful differences (e.g. to avoid the problem exposed in the example above for positive and negative symptoms). This requires a way of representing difference in disease state, as measured by instruments Y , so that the outcome Z will reflect clinically relevant differences in the inevitable variation in response between patients i.e. different ‘modes’ of response. An additional (but less essential) requirement would be that the outcome can also be used to *index* patients who share similar disease states. This might, for example, be suitable for use in N-of-1 trial designs (Schork 2015) to collect naturalistic evidence for treatments in the absence of complete understanding of disease process.

4. Clinical Example

To illustrate our proposal, consider Figure 2 (equivalent colour versions can be found in Supplementary Information), that shows 1459 individual patient's disease states as measured by the PANSS domain scores (positive, negative and general symptoms) from the baseline assessment of the Clinical Antipsychotic Trials of Intervention Effectiveness (CATIE) trial (Stroup et al. 2003) . We use three variables purely to ease visualisation and exposition of the key concepts, acknowledging that we have summed the items in the PANSS to obtain three domains, but the principles will generalise and do not require summed domains. Indeed, in practice, it would be prudent to use factor-analytic decompositions of these domains, see for example, (Wallwork et al. 2012; Lindenmayer et al. 1995), but to keep our explanation simple and easy to visualise, we restrict ourselves to the 3 original groupings of signs and symptoms in PANSS. In our example, the PANSS domains form a three-dimensional space, where each patient is represented by a point located on orthogonal (perpendicular) axes, representing low-to-high severity on positive (P), negative (N) and general (G) domains. Visualising this space is difficult, so following standard practice in multivariate analysis, we present three 'views' obtained by plotting each combination of P, N and G in two-dimensional planes: $P \times N$, $G \times P$ and $N \times G$ shown in Figure 2-1, 2-2 and 2-3 respectively.

FIGURE 2 HERE

We then define 4 *prototypes* proposed to represent hypothesised disease states with clinically meaningful or interesting locations in this space: for example, prototype A represents disease states that are low severity across positive, negative and general symptoms (i.e. a relatively well patient). Prototype B represents the opposite extreme – a patient that is globally unwell with high symptom severity across positive, negative and general symptoms. Prototype C represents a patient who has relatively high positive but low negative and relatively low general symptom severity. Prototype D

represents a disease state where a patient has relatively low positive but high negative and general symptom severity. Note that in defining prototypes, we are specifying structure with respect to the *actual* study population that can be exploited to define an outcome that *preserves* (rather than collapses) information that has clinical relevance, as required by our definition of strong aggregates. Patients with disease states near prototype A (relatively well) are clearly very different to those near B (globally unwell), and similarly for C and D where these patients are far from ‘well’ but whose disease states reflect different patterns of disease state.

4.1 Measures Derived from Weak Aggregates

Figure 2-4 shows the resulting univariate (i.e. one-dimensional) scale of the summed *weak* aggregate (where $Z = P + N + G$), with the location of each patient and the four prototypes along this scalar measure by their respective scores Z . Of note, relatively well (A) and globally unwell (B) prototypes are well demarcated (i.e. distant) at each end of the univariate scale, but C and D less so. This stands in contrast to the relative positions in the original space where C, D and B are well separated along the positive and negative symptoms dimension (Figure 2-1) as well as C and D being distinguished along both general and positive (Figure 2-2) negative and general dimensions (Figure 2-3). The critical point is that a weak aggregate can only discriminate between *well* and *unwell* patients. This is emphasised in the original three-dimensional space where each point (a single patient) in Figures 2-1, 2-2 and 2-3 are shaded light-to-dark according to their weak (summed) aggregate score Z (e.g. the greyscale of the points and scale bar shown in Figure 2-4). Notice how the gradient from light-to-dark (reflecting low to high Z) is broadly uniform in direction (bottom left to top right) over each of the three views, enabling distinction between A and B, but less so for C, and poorly for B and D. This demonstrates graphically how weak aggregates collapse and obscure meaningful distinctions between potential different disease states unless they remain well separated on the univariate scale of Z (Figure 2-4).

4.2 Measures Derived from Strong Aggregates

We now consider how the structure displayed in the prototypes can be captured in a way that enables a univariate outcome measure to be derived, but *preserving* distinctions between them in a meaningful way i.e. *strong aggregation*. The method we used was singular value decomposition (SVD) which is similar to principle components analysis (Strang 2004), embedding a high-dimensional space into a lower dimensional representation and in this case, exposes properties of interest (e.g. the separation of clinically relevant prototypes). We note that SVD is one of many possibilities for this embedding transformation; the key requirements of any chosen method are dimensionality reduction with distance preservation (isometry) and other approaches include, for example, multidimensional scaling (Krzanowski 2000), isomap embedding (Tenenbaum et al. 2000), locally linear embedding (Roweis & Saul 2000) and self-organising maps (Kohonen 1995) . Essentially, rather than ‘blindly’ summing, we use a method of combining or mapping each variable from Y (equivalently, the axes in Figure 2) such that clinically relevant regions of that space (prototypes in Figure 2) are mapped onto *sufficiently different* values in the univariate aggregate Z. After applying SVD to the same patients and prototypes in Figure 2, we are able to find a new, univariate strong aggregate that preserves the proposed clinically relevant difference in disease states exemplified by the prototypes (details are given in Supplementary Information).

FIGURE 3 HERE

Figure 3-4 shows the shape of the new strong univariate aggregate Z (using a soft rather than hard threshold scheme) with the new values of the prototypes illustrated. This aggregate crucially separates the clinically relevant prototypes C and D (compared to the weak aggregate shown in

Figure 2). In Figures 3-1, 3-2 and 3-3 (colour versions are reproduced in Supplementary Information), the patients in Figure 2 are assigned new values according to the strong aggregate Z (light grey = low Z, dark grey = high Z). Note the difference in how the gradient of greyscales compares with Figure 2, emphasising the score of patients varying to their proximity along a line dividing C and D.

5. Discussion and Conclusions

In this paper, we have discussed how clinical trials in psychiatry have to cope with uncertain relationships between the treatment, disease state and process, and how this has potentially hindered clinical trial research. Further, relationships among patients can be anchored to prototypes of clinical interest, in disease states as measured by clinical instruments. This may provide more useful information when defining outcomes by essentially exploiting geometric structure. Our example used the PANSS as an exemplar instrument chosen for its general clinical familiarity, but the principles extend to any other instrument for other disorder areas (e.g. affective disorders and the Hamilton Depression Scale).

We suggest increased use of strong aggregates because they capture important structure *between regions of the measured disease state* that weak aggregates ignore by blindly summing (or averaging) and that we have shown cause patients with different disease states to be mapped onto the same aggregate value. We used one specific method (SVD) to define a strong aggregate, but any similar method that captures clinically relevant structure over regions of the space of disease states and then assigns a single univariate (scalar) variable that preserves the distance between these regions would be suitable. Importantly, we defined prototypes only with respect to the actual patient population – remaining agnostic to the *actual* unknown disease process or classical

diagnostic category, but specifying tentative clinically relevant disease states. The resulting strong aggregate is univariate and therefore compatible with current statistical methodology. We now consider specific implications for trial methodology, design and analysis.

5.1 Prototype and Response Definition

The crux of our proposal is that univariate strong aggregates expose differences between *relevant* prototypes. In the example provided earlier, we chose four prototypes in the 3-dimensional space of positive, negative and general symptoms representing globally well (A), globally unwell (B), as well as dominantly positive (C) and negative symptoms (D). As a further example of *a priori* prototypes, the remission criteria defined in (Andreasen et al. 2005) can be interpreted in our framework as follows; one dimension measures reality distortion (R , the sum of items P1, G9 and P3), another captures disorganisation (D , sum of P2, G5) and another captures negative symptoms (N , the sum of N1, N4 and N6). This similarly forms a 3-dimensional space with axes R , D and N that can be visualised similarly to the examples presented earlier. There are then two relevant prototypes capturing the two extremes of *full* (best) and *no* (worst) remission. The participants and prototypes are then transformed into a space V (e.g. Supplementary Information, Figure S3; by singular-valued decomposition) which can then be visualised to find the single *univariate* dimension, V^* , that *best* exposes the gradient between these two extreme prototypes. Prior to the trial intervention, each participant is then assigned a value (the strong aggregate) defined as their location along this dimension V^* (as in Figure 3-4 above) that defines the participant's pre-intervention remission state. At the end of the trial, each participant's *post-treatment* values of R , D , and N are transformed into V , and their positions along the dimension V^* are 'read off' resulting in the participant's strong aggregate measure of remission state after intervention. Alternatively, a 'hard' threshold over V^* can be defined – for example, if seeking to use the strong aggregate in a binary logistic/probit GLM analysis. We note that in principle, disease states and prototypes need not be restricted to measures in 3-dimensional space and is used here only for ease of exposition.

The key is that one identifies the single dimension (in the transformation, by e.g. SVD) that exposes the differences between prototypes.

5.2 Power and Recruitment

In calculating *a priori* the sample size required for an adequately powered trial, the distribution of the data, the expected effect size (means) or response rate, and measures of variance are required. Therefore, just as for weak aggregates used as outcomes, if data is available from previous studies or pilot data, then the distribution, means/response rates and variance assumptions should be justified by applying the proposed strong aggregate definition on available pilot data. Our proposal for strong aggregates is motivated by the idea that we may be failing to capture meaningful differences using weak aggregates; for this reason, there is potential to increase the power of a given study design.

Recruitment to a prospective trial need not differ when using strong aggregates – however, they offer a potential advantage because participants are assigned a continuous univariate score based on their relationship to prototypes (see Figure 3). The prototypes can define not only proposed or desired endpoints (e.g. defining two prototypes at extremes of positive symptoms) but also, ‘landmarks’ of interest (for example, as we did for prototypes A, B, C and D in the above examples). Then, participants could be stratified to treatment by their score in relation to prototypes – for example, those closer to prototype A may be assigned one treatment, those closer to prototype B another. Further, we note a further application in *N-of-1* trials, where diagnostic uncertainty is likely to be even more problematic and uncontrolled; in such instances, using SVD, it is straightforward to index patients by their similarity to each other and their similarity to prototypes. In practice, a new patient attending a clinic with a certain disease state (e.g. measured by PANSS) can be easily compared to patients and prototypes using an SVD model (see Supplementary

Information), assigned a predicted outcome and stratified to treatment if, historically, there were treatments that worked for patients similarly proximal to certain prototypes.

5.3 Trial Data Re-analysis

One compelling reason to use strong aggregates is that it mitigates against multiple secondary analyses by i) requiring prototypes to be *a priori* defined to capture the proposed disease states (e.g. globally well, or dominant negative symptoms) relevant to the intervention and ii) providing a univariate measure over these disease states that reflects a given participant's symptoms (or response to a treatment). We suggest this prevents the scenario where, after failing to find a desired result using a weak aggregate primary outcome, secondary analyses are then required on 'subsets' of participants. Necessarily, defining prototypes forces us to consider the clinically meaningful states – rather than looking for a global change in e.g. 'total' PANSS scores – and provides a way to define a *single* measure that captures participants' disease state relative to these. There is significant potential for many re-analyses of existing data using this paradigm. To illustrate, we recently conducted a systematic review of trials for treating the cognitive symptoms of schizophrenia registered on ClinicalTrials.org in the period 2004-2015 (Joyce et al. 2017). We identified a total of 114 studies, but when we specifically examined definitions of primary outcome and available results, only 18 were eligible for inclusion. We explored the definition of primary outcomes on instruments (e.g. PANSS), finding only 4 of the 18 studies considered specific combinations of variables or domain scores (a necessary first step to define prototypes for strong aggregates) instead of using weak aggregates. Unsurprisingly, secondary outcomes were often used to understand the multi-dimensional (rather than univariate) measurement of disease state.

In summary, the implications of our proposal are two-fold; first, given our arguments for the importance of disease state, we should reorient clinical trials toward recruiting for specific symptoms rather than diagnoses. Second, clinical trial analyses should explore and then exploit the

heterogeneity of disease states (measured by familiar clinical instruments) and seek strong aggregates that focus outcomes on specific, relevant definitions of treatment response or failure rather than assuming homogeneity, weak aggregation and settling for the average response for a diagnostic category.

References

- Andreasen, N.C. et al., 2005. Remission in Schizophrenia: Proposed Criteria and Rationale for Consensus. *Am J Psychiatry*, 162(3), pp.441–449. Available at: <http://ajp.psychiatryonline.org> [Accessed November 10, 2016].
- Annen, S., Roser, P. & Brüne, M., 2012. Nonverbal behavior during clinical interviews: similarities and dissimilarities among schizophrenia, mania, and depression. *The Journal of nervous and mental disease*, 200, pp.26–32.
- Barnow, S. et al., 2010. Borderline personality disorder and psychosis: A review. *Current Psychiatry Reports*, 12, pp.186–195.
- Berkowitz, R.L. et al., 2012. The impact of the clinical antipsychotic trials of intervention effectiveness (CATIE) on prescribing practices: An analysis of data from a large midwestern state. *Journal of Clinical Psychiatry*, 73, pp.498–503.
- Jakubovski, E., Carlson, J.P. & Bloch, M.H., 2015. Prognostic Subgroups for Remission, Response, and Treatment Continuation in the Clinical Antipsychotic Trials of Intervention Effectiveness (CATIE) Trial. *The Journal of Clinical Psychiatry*, 76(11), pp.1535–1545.
- Joyce, D.W. et al., 2017. Realising stratified psychiatry using multidimensional signatures and trajectories. *Journal of Translational Medicine*, 15(1), p.15.
- Kendler, K.S., 2016. The nature of psychiatric disorders. *World psychiatry*, 15(1), pp.5–12.
- Krzanowski, W.J., 2000. *Principles of Multivariate Analysis. A User's Perspective*,
- Leucht, S. et al., 2009. Definitions of response and remission in schizophrenia: recommendations for their use and their presentation. *Acta Psychiatrica Scandinavica*, 119(438), pp.7–14.
- Lindenmayer, J.P. et al., 1995. Five factor model of schizophrenia: replication across samples. *Schizophrenia research*, 14(3), pp.229–34.
- Marquand, A.F. et al., 2016. Beyond Lumping and Splitting: A Review of Computational Approaches for Stratifying Psychiatric Disorders. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, 1(5), pp.433–447.
- Strang, G., 2004. *Linear Algebra and Its Applications* 4th ed., WB Saunders.
- Toh, W.L., Thomas, N. & Rossell, S.L., 2015. Auditory verbal hallucinations in bipolar disorder (BD) and major depressive disorder (MDD): A systematic review. *Journal of Affective Disorders*, 184, pp.18–28.
- Wallwork, R.S. et al., 2012. Searching for a consensus five-factor model of the Positive and Negative Syndrome Scale for schizophrenia. *Schizophrenia research*, 137(1–3), pp.246–50.
- Craddock, N., O'Donovan, M.C. & Owen, M.J., 2009. Psychosis genetics: Modeling the relationship between schizophrenia, bipolar disorder and mixed (or “schizoaffective”) psychoses. *Schizophrenia Bulletin*, 35(3), pp.482–490.
- Cuthbert, B.N. & Insel, T.R., 2013. Toward the future of psychiatric diagnosis: the seven pillars of RDoC. *BMC medicine*, 11, p.126. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3653747&tool=pmcentrez&rendertype=abstract>.

- Demjaha, A. et al., 2009. Combining dimensional and categorical representation of psychosis: the way forward for DSM-V and ICD-11? *Psychological medicine*, 39, pp.1943–1955.
- Demjaha, A., MacCabe, J.H. & Murray, R.M., 2012. How genes and environmental factors determine the different neurodevelopmental trajectories of schizophrenia and bipolar disorder. *Schizophrenia Bulletin*, 38, pp.209–214.
- Frank, E. et al., 1991. Conceptualization and rationale for consensus definitions of terms in major depressive disorder. Remission, recovery, relapse, and recurrence. *Archives of general psychiatry*, 48(9), pp.851–5. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/1929776> [Accessed November 18, 2016].
- Glaser, J.P. et al., 2010. Psychotic reactivity in borderline personality disorder. *Acta Psychiatrica Scandinavica*, 121, pp.125–134.
- Green, M.F., 2006. Cognitive impairment and functional outcome in schizophrenia and bipolar disorder. *The Journal of clinical psychiatry*, 67, pp.3–8.
- Insel, T. et al., 2010. Research Domain Criteria (RDoC): Toward a new classification framework for research on mental disorders. *American Journal of Psychiatry*, 167, pp.748–751.
- Jabben, N. et al., 2009. Investigating the association between neurocognition and psychosis in bipolar disorder: Further evidence for the overlap with schizophrenia. *Bipolar Disorders*, 11, pp.166–177.
- Jablensky, A., Jablensky, a & Review, F., 2006. Subtyping schizophrenia: implications for genetic research. *Molecular psychiatry*, 11(9), pp.815–36. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/16801952> [Accessed October 6, 2016].
- Jakubovski, E., Carlson, J.P. & Bloch, M.H., 2015. Prognostic Subgroups for Remission, Response, and Treatment Continuation in the Clinical Antipsychotic Trials of Intervention Effectiveness (CATIE) Trial. *The Journal of Clinical Psychiatry*, 76(11), pp.1535–1545.
- Kay, S.R., Fiszbein, A. & Opler, L. a, 1987. The positive and negative syndrome scale (PANSS) for schizophrenia. *Schizophrenia bulletin*, 13, pp.261–76. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/23434198>.
- Keshavan, M.S. et al., 2011. A dimensional approach to the psychosis spectrum between bipolar disorder and schizophrenia: The Schizo-Bipolar Scale. *Schizophrenia Research*, 133, pp.250–254.
- Kohonen, T., 1995. *Self-Organizing Maps*, Berlin: Springer.
- Leucht, S. et al., 2009. Definitions of response and remission in schizophrenia: recommendations for their use and their presentation. *Acta Psychiatrica Scandinavica*, 119(438), pp.7–14.
- Lichtenstein, P. et al., 2009. Common genetic determinants of schizophrenia and bipolar disorder in Swedish families: a population-based study. *Lancet*, 373, pp.234–239.
- McCullagh, P. & Nelder, J.A., 1989. *Generalized Linear Models* Second Edi., Chapman & Hall/CRC. Available at: <http://www.amazon.com/dp/0412317605>.

- Moore, A. et al., 2013. Expect analgesic failure; pursue analgesic success. *British Medical Journal*, 346(May), pp.2690–2690. Available at: <http://www.bmj.com/cgi/doi/10.1136/bmj.f2690>.
- Moore, R. a. et al., 2010. Clinical effectiveness: An approach to clinical trial design more relevant to clinical practice, acknowledging the importance of individual differences. *Pain*, 149(2), pp.173–176. Available at: <http://dx.doi.org/10.1016/j.pain.2009.08.007>.
- Morris, S.E. & Cuthbert, B.N., 2012. Research domain criteria: Cognitive systems, neural circuits, and dimensions of behavior. *Dialogues in Clinical Neuroscience*, 14, pp.29–37.
- Nishizono-Maher, A. et al., 1993. Psychotic symptoms in depression and borderline personality disorder. *Journal of Affective Disorders*, 28, pp.279–285.
- Overall, J.E. & Gorham, D.R., 1962. THE BRIEF PSYCHIATRIC RATING SCALE. *Psychological Reports*, 10(3), pp.799–812. Available at: <http://dx.doi.org/10.2466/pr0.1962.10.3.799>.
- Purcell, S.M. et al., 2009. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature*, 460, pp.748–752.
- Reininghaus, U. et al., 2016. Evaluation of the validity and utility of a transdiagnostic psychosis dimension encompassing schizophrenia and bipolar disorder. *The British journal of psychiatry : the journal of mental science*, 209(2), pp.107–13. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/26989099> [Accessed August 15, 2016].
- Ripke, S. et al., 2014. Biological insights from 108 schizophrenia-associated genetic loci. *Nature*, 511, pp.421–427. Available at: <http://www.nature.com/doi/10.1038/nature13595>.
- Roweis, S.T. & Saul, L.K., 2000. Nonlinear Dimensionality Reduction by Locally Linear Embedding. *Science*, 290(5500).
- Schork, N.J., 2015. Time for one-person trials. *Nature*, 520(7549), pp.609–611. Available at: <http://www.nature.com/doi/10.1038/520609a>.
- Schroeder, K., Fisher, H.L. & Schäfer, I., 2013. Psychotic symptoms in patients with borderline personality disorder: prevalence and clinical management. *Current opinion in psychiatry*, 26, pp.113–9.
- Schumann, G. et al., 2014. Stratified medicine for mental disorders. *European Neuropsychopharmacology*, 24(1), pp.5–50.
- Strang, G., 2004. *Linear Algebra and Its Applications* 4th ed., WB Saunders. Available at: <http://www.maa.org/publications/maa-reviews/introduction-to-linear-algebra-0> [Accessed November 18, 2016].
- Stroup, T.S. et al., 2003. The National Institute of Mental Health Clinical Antipsychotic Trials of Intervention Effectiveness (CATIE) project: schizophrenia trial design and protocol development. *Schizophrenia bulletin*, 29(11). Stroup TS, Mcevoy JP, Swartz MS, et The National Institute of Mental Health Clinical Antipsychotic Trials of Intervention Effectiveness (CATIE) project: schizophrenia trial design and protocol development. *Schizophr Bull*. 2003;29(1):15-31. <http://>, pp.15–31.

- Tenenbaum, J.B., de Silva, V. & Langford, J.C., 2000. A global geometric framework for nonlinear dimensionality reduction. *Science (New York, N.Y.)*, 290(5500), pp.2319–23. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/11125149> [Accessed November 18, 2016].
- Walker, J., Curtis, V. & Murray, R.M., 2002. Schizophrenia and bipolar disorder: similarities in pathogenic mechanisms but differences in neurodevelopment. *International clinical psychopharmacology*, 17 Suppl 3, pp.S11–S19.

Figure Captions :

Figure 1 : 1 (left) illustrates the typical model of a clinical trial in medicine. 2 (right) illustrates the typical model as applied to psychiatry, where there is a lack of a clear link between disease state (S) and disease process (DP) and consequently, they are usually replaced by a diagnostic category (Dx). Arrows indicate dependence relations between variables – for example, disease state (S) depends on the disease process (DP).

Figure 1

Supplementary Information :

Figure S1 – Colour Version of Weak Aggregate (Main Text – Figure 2)

Figure S2 – Colour Version of Strong Aggregate (Main Text – Figure 3)

Application of Singular Valued Decomposition (SVD) to Derive Strong Aggregate

Suppose that, as in the main text, the distinction between disease states in the neighbourhood of prototypes C and D are clinically relevant. Using the positive, negative and general symptom domain scores for each patient, a 3-dimensional space is formed, shown in Figure 2 and 3 of the main text, and above, in colour form.

We first define the $M \times N$ matrix, $\mathbf{Y}$, where rows, M , are variables measuring clinical state e.g. PANSS domains (positive, negative and general respectively) and the columns are values of each variable for the 1459 patients. We then append to $\mathbf{Y}$ the prototypes A, B, C and D resulting in $N = 1463$.

An SVD (Strang, 2004) decomposes $\mathbf{Y}$ into three matrices:

$$\mathbf{Y} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$$

Where $\mathbf{U}$ and $\mathbf{V}$ are square matrices (the so-called left and right singular vectors) whose columns are the orthogonal eigenvectors of the matrix $\mathbf{Y}\mathbf{Y}^T$ and $\mathbf{Y}^T\mathbf{Y}$ respectively. Elements (i,j) of $\mathbf{Y}^T\mathbf{Y}$ represent a measure of similarity between patients i and j based on co-occurrence of their PANSS domains scores (positive, negative and general). $\mathbf{\Sigma}$ is the diagonal matrix of singular values of $\mathbf{Y}$.

We can now examine the pattern of similarity between patients and prototypes by using only the columns of $\mathbf{V}$ representing the 3 largest singular values - resulting in a matrix of size 1463 x 3, which can be visualised as shown in Figure S3.

Figure S3

By inspection, we find that the prototypes of clinical relevance that we wish to emphasise (see main text), C and D, are most separated along the axis V3 (corresponding to the 3rd singular value). We label this the *best* or *optimal* dimension V^* . To form a univariate score, we simply retrieve the position of each of the 1459 points (for each patient) on V^* . To define our strong aggregate (in the traditional sense of a univariate outcome), we can then use either a hard or soft threshold. In the former, we chose a fixed point along V^* , and all patients higher/lower are assigned a binary outcome (which could represent for example, response or non-response). Alternatively, we used a ‘soft-threshold’, assigning each patients’ strong aggregate value according to a smooth function over their values along V^* (Figure 3 in the main text, shown in colour above as Figure S2).

Using SVD For Matching / Retrieving Patients

In the Discussions and Conclusions of the main text, we suggested that beyond defining strong aggregates, the formulation of SVD has utility for ‘matching’ patients. In this application - as it is in information retrieval and latent semantic indexing (Manning, Raghavan and Shutze, 2008) – given an SVD of some set of patients represented by $\mathbf{Y}$ defined as before (e.g. rows correspond to variables measuring clinical state via an instrument such as PANSS items or domains, with columns again representing individual patients), a new patient becomes a ‘query’ for which we seek similar patients and measure the proximity of the new patient to prototypes.

Assume an SVD has been constructed in exactly the same way as before, but now, we retain some number k of the singular values (i.e. rather than 3, we might opt for larger k). This results in the reduced-rank representation of $\mathbf{Y}$ as follows:

$$\mathbf{Y}_k = \mathbf{U}_k \mathbf{\Sigma}_k \mathbf{V}_k^T$$

$\mathbf{Y}_k$ is now a lower-dimensional approximation of the original $\mathbf{Y}$ that, by virtue of retaining only a subset of the largest k singular-values, reduces redundancy and noise, emphasising the dominant similarities between patients’ clinical states. Any new patient y' represented by their state in the original space $\mathbf{Y}$ can be queried for similarity to other patients and prototypes by first computing their projection into the lower dimensional space $\mathbf{Y}_k$:

$$y'_k = \mathbf{\Sigma}_k^{-1} \mathbf{U}_k^T y'$$

Then, comparison to other patients (and prototypes) proceeds by measuring the cosine similarity between y'_k and columns of $\mathbf{V}_k^T$ – see (Manning, Raghavan and Shutze, 2008) for further details and demonstration.