Psychological Medicine (2018), **48**, 177–186. © Cambridge University Press 2017
doi:10.1017/S0033291717001726

INVITED REVIEW

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

D. W. Joyce[1]*, D. K. Tracy[1,2] and S. S. Shergill[1]

[1] *Cognition Schizophrenia and Imaging Laboratory, Department of Psychosis Studies, Institute of Psychiatry, Psychology and Neuroscience, King's College London, De Crespigny Park, London SE5 8AF, PO63, UK*
[2] *Oxleas NHS Foundation Trust, London, UK*

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.

Received 1 December 2016; Revised 25 May 2017; Accepted 25 May 2017; First published online 29 June 2017

Key words: Trials, methodology, outcomes, dimensional definitions of disorder.

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 Fig. 1-1, for a given disorder or DP, there will be a disease state (S) that corresponds with that disorder – but not necessarily in a

[*] Address for correspondence: D. W. Joyce, Cognition Schizophrenia and Imaging Laboratory, Department of Psychosis Studies, Institute of Psychiatry, Psychology and Neuroscience, King's College London, De Crespigny Park, London SE5 8AF, PO63, UK.
(Email: danjoyce@doctors.org.uk)

https://doi.org/10.1017/S0033291717001726 Published online by Cambridge University Press

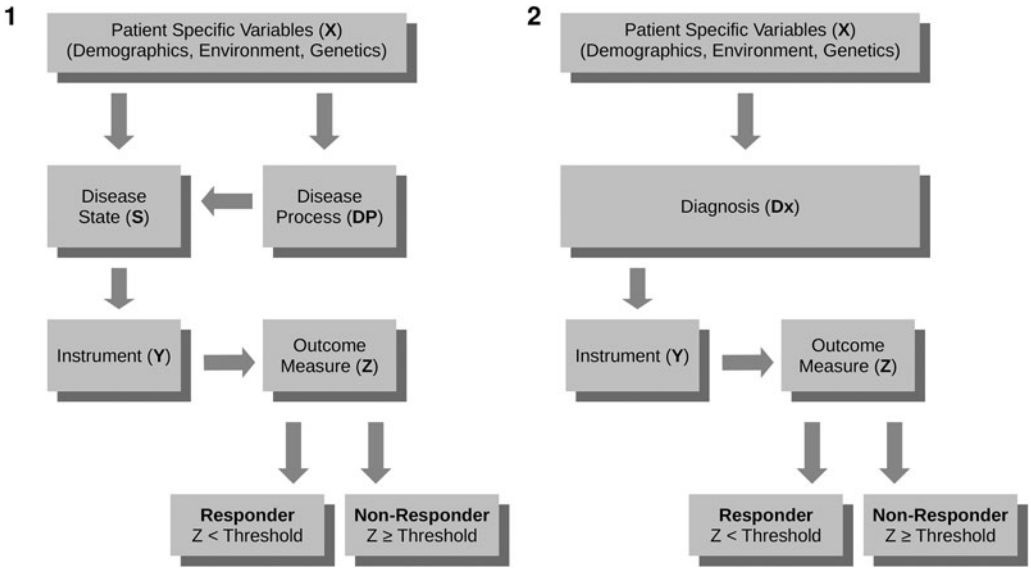

Fig. 1. (left) illustrates the typical model of a clinical trial in medicine. 2 (right) illustrates the typical model as applied to psychiatry, where there is a lack of a clear link between disease state (S) and disease process (DP) and consequently, they are usually replaced by a diagnostic category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

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.

Psychiatry does not benefit from as clear a correspondence between disease state (S) and 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 Fig. 1-2, where a diagnostic category (Dx) such as schizophrenia or bipolar affective disorder replaces 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) (Overall & Gorham, 1962; Kay *et al.* 1987). 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 (Nishizono-Maher *et al.* 1993; Barnow *et al.* 2010; Glaser *et al.* 2010; Schroeder *et al.* 2013). In bipolar disorder and schizophrenia, there are similarities in nonverbal communication (Annen *et al.* 2012), affective symptoms (Keshavan *et al.* 2011) and cognitive deficits (Green, 2006; Jabben *et al.* 2009).

There is also consensus, for example, that the diagnosis of schizophrenia is not a single DP, but rather a categorical label for a syndrome with different aetiologies (Walker *et al*. 2002; Jablensky *et al.* 2006; Demjaha *et al.* 2009; Demjaha *et al.* 2012; Ripke *et al.* 2014; Reininghaus *et al.* 2016) and shared genetic risk factors (Craddock *et al.* 2009; Lichtenstein *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 (e.g. Insel *et al.* 2010; Morris & Cuthbert, 2012; Cuthbert & Insel, 2013; Schumann *et al.* 2014).

Currently, clinical trials in psychiatry have to contend with a lack of clear relationship between disease

https://doi.org/10.1017/S0033291717001726 Published online by Cambridge University Press

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 Fig. 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 symptoms 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 Fig. 1-2, adopting the model shown in Fig. 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.

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 Fig. 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 Fig. 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 DP 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

https://doi.org/10.1017/S0033291717001726 Published online by Cambridge University Press

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.

Strong aggregate outcomes

There is an additional problem with weak aggregation in psychiatry illustrated in Fig. 1-2, when DP 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.* 2010; Moore *et al.* 2013) 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 DP.

Clinical example

To illustrate our proposal, consider Fig. 2 (equivalent colour versions can be found in online 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 (Lindenmayer *et al.* 1995; Wallwork *et al.* 2012], 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 Figs. 2-1, 2-2 and 2-3 respectively.

We then define four *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

https://doi.org/10.1017/S0033291717001726 Published online by Cambridge University Press

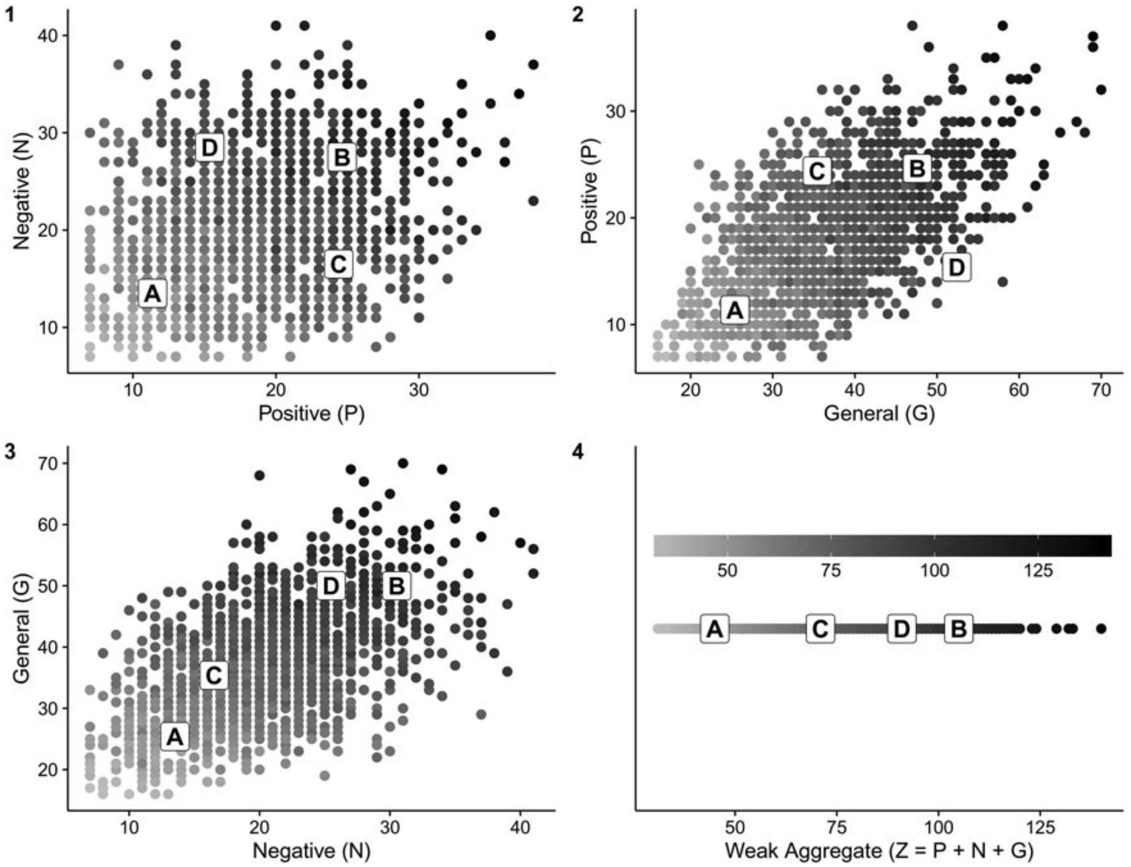

Fig. 2. 1459 patients represented as positive, negative and general psychopathology scores under a weak aggregation scheme.

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.

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 (Fig. 2-1) as well as C and D being distinguished along both general and positive (Fig. 2-2) negative and general dimensions (Fig. 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 Figs. 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 Fig. 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 (Fig. 2-4).

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

https://doi.org/10.1017/S0033291717001726 Published online by Cambridge University Press

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 Fig. 2) such that clinically relevant regions of that space (prototypes in Fig. 2) are mapped onto *sufficiently different* values in the univariate aggregate *Z*. After applying SVD to the same patients and prototypes in Fig. 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 online Supplementary Information).

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 with the weak aggregate shown in Fig. 2). In Figs. 3-1, 3-2 and 3-3 (colour versions are reproduced in online Supplementary Information), the patients in Fig. 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 Fig. 2, emphasising the score of patients varying to their proximity along a line dividing C and D.

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 DP 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.

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 and G5) and another captures negative symptoms (*N*, the sum of N1, N4 and N6). This similarly forms a three-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. online Supplementary Information, Fig. 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 Fig. 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.

https://doi.org/10.1017/S0033291717001726 Published online by Cambridge University Press

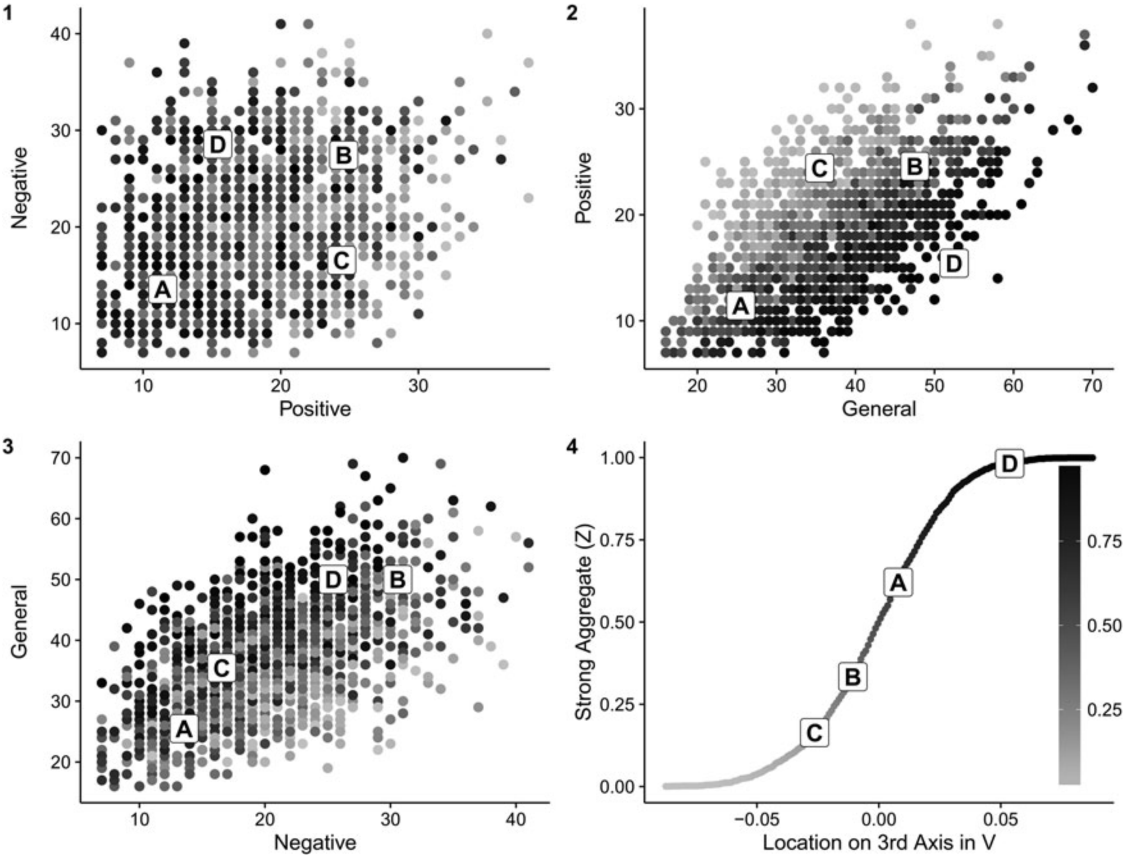

Fig. 3. 1459 patients represented as positive, negative and general psychopathology scores under a strong aggregation scheme.

We note that in principle, disease states and prototypes need not be restricted to measures in three-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.

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

https://doi.org/10.1017/S0033291717001726 Published online by Cambridge University Press

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 towards recruiting for the 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.

Supplementary material

The supplementary material for this article can be found at https://doi.org/10.1017/S0033291717001726

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). D.W.J. 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.

Declaration of Interests

None.

References

Andreasen NC, Carpenter WT, Kane JM, Lasser RA, Marder SR and Weinberger DR (2005) 'Remission in Schizophrenia: Proposed Criteria and Rationale for Consensus', *Am J Psychiatry*, **162**(3), pp. 441–449.

Annen S, Roser P and 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, Arens EA, Sieswerda S, Dinu-Biringer R, Spitzer C and Lang S (2010) 'Borderline personality disorder and psychosis: A review', *Current Psychiatry Reports*, **12**, pp. 186–195.

Berkowitz RL, Patel U, Ni Q, Parks JJ and Docherty JP (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.

Craddock N, O'Donovan MC and Owen MJ (2009) 'Psychosis genetics: Modeling the relationship between schizophrenia, bipolar disorder and mixed (or "schizoaffective") psychoses', *Schizophrenia Bulletin*, **35**(3), pp. 482–490.

Cuthbert BN and Insel TR (2013) 'Toward the future of psychiatric diagnosis: the seven pillars of RDoC.', *BMC medicine*, **11**, p. 126.

Demjaha A, Morgan K, Morgan C, Landau S, Dean K, Reichenberg A, Sham P, Fearon P, Hutchinson G, Jones PB, Murray RM and Dazzan P (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 JH and Murray RM (2012) 'How genes and environmental factors determine the different neurodevelopmental trajectories of schizophrenia and bipolar disorder', *Schizophrenia Bulletin*, **38**, pp. 209–214.

Frank E, Prien RF, Jarrett RB, Keller MB, Kupfer DJ, Lavori PW, Rush AJ and Weissman MM (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.

https://doi.org/10.1017/S0033291717001726 Published online by Cambridge University Press

Glaser JP, Van Os J, Thewissen V and Myin-Germeys I (2010) 'Psychotic reactivity in borderline personality disorder', *Acta Psychiatrica Scandinavica*, **121**, pp. 125–134.

Green MF (2006) 'Cognitive impairment and functional outcome in schizophrenia and bipolar disorder.', *The Journal of clinical psychiatry*, **67**, pp. 3–8.

Insel T, Cuthbert B, Garvey M, Heinssen R, Pine DS, Quinn K, Sanislow C and Wang P (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, Arts B, Krabbendam L and Van Os J (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 (2006) 'Subtyping schizophrenia: implications for genetic research.', *Molecular psychiatry*, **11**(9), pp. 815–36.

Jakubovski E, Carlson JP and Bloch MH (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 DW, Kehagia AA, Tracy DK, Proctor J and Shergill SS (2017) 'Realising stratified psychiatry using multidimensional signatures and trajectories', *Journal of Translational Medicine. BioMed Central*, **15**(1), p. 15.

Kay SR, Fiszbein A and Opler L (1987) 'The positive and negative syndrome scale (PANSS) for schizophrenia.', *Schizophrenia bulletin*, **13**, pp. 261–76.

Kendler KS (2016) 'The nature of psychiatric disorders.', *World psychiatry. World Psychiatric Association*, **15**(1), pp. 5–12.

Keshavan MS, Morris DW, Sweeney JA, Pearlson G, Thaker G, Seidman LJ, Eack SM and Tamminga C (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.

Krzanowski WJ (2000) Principles of Multivariate Analysis. A User's Perspective, Oxford Statistical Science Series No. 23.

Leucht S, Davis JM, Engel RR, Kissling W and Kane JM (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, Yip BH, Björk C, Pawitan Y, Cannon TD, Sullivan PF and Hultman CM (2009) 'Common genetic determinants of schizophrenia and bipolar disorder in Swedish families: a population-based study.', *Lancet*, **373**, pp. 234–239.

Lindenmayer JP, Grochowski S, Hyman RB, Powchik P and Davidson M (1995) 'Five factor model of schizophrenia: replication across samples.', *Schizophrenia research. Netherlands, Amsterdam*, **14**(3), pp. 229–34.

Marquand AF, Wolfers T, Mennes M, Buitelaar J and Beckmann CF (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.

McCullagh P and Nelder JA (1989) *Generalized Linear Models*. Second Edi. Chapman & Hall/CRC.

Moore A, Derry S, Eccleston C and Kalso E (2013) 'Expect analgesic failure; pursue analgesic success', *British Medical Journal*, **346**(May), pp. 2690–2690.

Moore R, Derry S, McQuay HJ, Straube S, Aldington D, Wiffen P, Bell RF, Kalso E and Rowbotham MC (2010) 'Clinical effectiveness: An approach to clinical trial design more relevant to clinical practice, acknowledging the importance of individual differences', *Pain. International Association for the Study of Pain*, **149**(2), pp. 173–176.

Morris SE and Cuthbert BN (2012) 'Research domain criteria: Cognitive systems, neural circuits, and dimensions of behavior', *Dialogues in Clinical Neuroscience*, **14**, pp. 29–37.

Nishizono-Maher A, Ikuta N, Ogiso Y, Moriya N, Miyake Y and Minakawa K (1993) 'Psychotic symptoms in depression and borderline personality disorder', *Journal of Affective Disorders*, **28**, pp. 279–285.

Overall JE and Gorham DR (1962) 'THE BRIEF PSYCHIATRIC RATING SCALE', Psychological Reports. *Ammons Scientific*, **10**(3), pp. 799–812.

Purcell SM, Wray NR, Stone JL, Visscher PM, O'Donovan MC, Sullivan PF and Sklar P (2009) 'Common polygenic variation contributes to risk of schizophrenia and bipolar disorder.', *Nature*, **460**, pp. 748–752.

Reininghaus U, Böhnke JR, Hosang G, Farmer A, Burns T, McGuffin P, Bentall RP, Insel TR, Adam D, First MB, Reed GM, Hyman SE, Saxena S, Bebbington P, Caspi AHR, Belsky DW, Goldman-Mellor SJ, Harrington H, Israel S, Meier MH, Carpenter WT, Bustillo JR, Thaker GK, Os J van, Krueger RF, Green MJ, Craddock N, Owen MJ, Cardno AG, Rijsdijk FV, Sham PC, Murray RM, McGuffin P, Lichtenstein P, Yip BH, Bjork C, Pawitan Y, Cannon TD, Sullivan PF, Brown AS, Os J van, Driessens C, Hoek HW, Susser ES, Clarke MC, Harley M, Cannon M, Matheson SL, Shepherd AM, Pinchbeck RM, Laurens KR, Carr VJ, Henquet C, Krabbendam L, Graaf R, Have M, Os J van, Henquet C, Murray R, Linszen D, Os J van, Heinz A, Deserno L, Reininghaus U, Kirkbride JB, Errazuriz A, Croudace TJ, Morgan C, Jackson D, Boydell J, Os J van, Kapur S, Reininghaus U, Priebe S, Bentall RP, Russo M, Levine SZ, Demjaha A, Forti MD, Bonaccorso S, Fearon P, McGuffin P, Farmer A, Harvey I, Burns T, Creed F, Fahy T, Thompson S, Tyrer P, White I, Cohen-Woods S, Craig I, Gaysina D, Gray J, Gunasinghe C, Craddock N, Spitzer RL, Endicott J, Robins E, Rucker J, Newman S, Gray J, Gunasinghe C, Broadbent M, Brittain P, Chalmers RP, Kass RW, Wasserman L, Peterson LE, Coleman MA, Kohavi R, Efron BG, Gong G, Böhnke JR, Croudace TJ, Cardno AG, Jones LA, Murphy KC, Murphy KC, Asherson P, Scott LC, Markon KE, Reininghaus U, Morgan C, Simpson J, Dazzan P, Morgan K, Doody G, Morgan C, Reininghaus U, Fearon P, Hutchinson G, Morgan K, Dazzan P, Reininghaus U, Craig T, Fisher H, Hutchinson G, Fearon P, Morgan K, Tamminga CA, Ivleva EI, Keshavan MS, Pearlson GD, Clementz BA and Witte B (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.

Ripke S, Neale BM, Corvin A, Walters JTR, Farh K-H, Holmans Pa, Lee P, Bulik-Sullivan B, Collier Da, Huang

https://doi.org/10.1017/S0033291717001726 Published online by Cambridge University Press

H, Pers TH, Agartz I, Agerbo E, Albus M, Alexander M, Amin F, Bacanu Sa, Begemann M, Belliveau Jr Ra, Bene J, Bergen SE, Bevilacqua E, Bigdeli TB, Black DW, Bruggeman R, Buccola NG, Buckner RL, Byerley W, Cahn W, Cai G, Campion D, Cantor RM, Carr VJ, Carrera N, Catts SV, Chambert KD, Chan RCK, Chen RYL, Chen EYH, Cheng W, Cheung EFC, Ann Chong S, Robert Cloninger C, Cohen D, Cohen N, Cormican P, Craddock N, Crowley JJ, Curtis D, Davidson M, Davis KL, Degenhardt F, Del Favero J, Demontis D, Dikeos D, Dinan T, Djurovic S, Donohoe G, Drapeau E, Duan J, Dudbridge F, Durmishi N, Eichhammer P, Eriksson J, Escott-Price V, Essioux L, Fanous AH, Farrell MS, Frank J, Franke L, Freedman R, Freimer NB, Friedl M, Friedman JI, Fromer M, Genovese G, Georgieva L, Giegling I, Giusti-Rodríguez P, Godard S, Goldstein JI, Golimbet V, Gopal S, Gratten J, de Haan L, Hammer C, Hamshere ML, Hansen M, Hansen T, Haroutunian V, Hartmann AM, Henskens Fa, Herms S, Hirschhorn JN, Hoffmann P, Hofman A, Hollegaard MV, Hougaard DM, Ikeda M, Joa I, Julià A, Kahn RS, Kalaydjieva L, Karachanak-Yankova S, Karjalainen J, Kavanagh D, Keller MC, Kennedy JL, Khrunin A, Kim Y, Klovins J, Knowles Ja, Konte B, Kucinskas V, Ausrele Kucinskiene Z, Kuzelova-Ptackova H, Kähler AK, Laurent C, Lee Chee Keong J, Hong Lee S, Legge SE, Lerer B, Li M, Li T, Liang K-Y, Lieberman J, Limborska S, Loughland CM, Lubinski J, Lönnqvist J, Macek Jr M, Magnusson PKE, Maher BS, Maier W, Mallet J, Marsal S, Mattheisen M, Mattingsdal M, McCarley RW, McDonald C, McIntosh AM, Meier S, Meijer CJ, Melegh B, Melle I, Mesholam-Gately RI, Metspalu A, Michie PT, Milani L, Milanova V, Mokrab Y, Morris DW, Mors O, Murphy KC, Murray RM, Myin-Germeys I, Müller-Myhsok B, Nelis M, Nenadic I, Nertney Da, Nestadt G, Nicodemus KK, Nikitina-Zake L, Nisenbaum L, Nordin A, O'Callaghan E, O'Dushlaine C, O'Neill FA, Oh S-Y, Olincy A, Olsen L, Van Os J, Endophenotypes International Consortium P, Pantelis C, Papadimitriou GN, Papiol S, Parkhomenko E, Pato MT, Paunio T, Pejovic-Milovancevic M, Perkins DO, Pietiläinen O, Pimm J, Pocklington AJ, Powell J, Price A, Pulver AE, Purcell SM, Quested D, Rasmussen HB, Reichenberg A, Reimers Ma, Richards AL, Roffman JL, Roussos P, Ruderfer DM, Salomaa V, Sanders AR, Schall U, Schubert CR, Schulze TG, Schwab SG, Scolnick EM, Scott RJ, Seidman LJ, Shi J, Sigurdsson E, Silagadze T, Silverman JM, Sim K, Slominsky P, Smoller JW, So H-C, Spencer CCA, Stahl EA, Stefansson H, Steinberg S, Stogmann E, Straub RE, Strengman E, Strohmaier J, Scott Stroup T, Subramaniam M, Suvisaari J, Svrakic DM, Szatkiewicz JP, Söderman E, Thirumalai S, Toncheva D, Tosato S, Veijola J, Waddington J, Walsh D, Wang D, Wang Q, Webb BT, Weiser M, Wildenauer DB, Williams NM, Williams S, Witt SH, Wolen AR, Wong EHM, Wormley BK, Simon Xi H, Zai CC, Zheng X, Zimprich F, Wray NR, Stefansson K, Visscher PM, Trust Case-Control Consortium W, Adolfsson R, Andreassen Oa, Blackwood DHR, Bramon E, Buxbaum JD, Børglum AD, Cichon S, Darvasi A, Domenici E, Ehrenreich H, Esko T, Gejman PV, Gill M, Gurling H, Hultman CM, Iwata N, Jablensky AV, Jönsson EG, Kendler KS, Kirov G, Knight J, Lencz T, Levinson DF, Li QS, Liu J, Malhotra AK, McCarroll Sa, McQuillin A, Moran JL, Mortensen PB, Mowry BJ, Nöthen MM, Ophoff Ra, Owen MJ, Palotie A, Pato CN, Petryshen TL, Posthuma D, Rietschel M, Riley BP, Rujescu D, Sham PC, Sklar P, St Clair D, Weinberger DR, Wendland JR, Werge T, Daly MJ, Sullivan PF and O'Donovan MC (2014) 'Biological insights from 108 schizophrenia-associated genetic loci', *Nature*, **511**, pp. 421–427.

Roweis ST and Saul LK (2000) 'Nonlinear Dimensionality Reduction by Locally Linear Embedding', *Science*, **290** (5500), pp. 2323–2326.

Schork NJ (2015) 'Time for one-person trials', *Nature*, **520** (7549), pp. 609–611.

Schroeder K, Fisher HL and 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, Binder EB, Holte A, de Kloet E.R, Oedegaard KJ, Robbins TW, Walker-Tilley TR, Bitter I, Brown VJ, Buitelaar J, Ciccocioppo R, Cools R, Escera C, Fleischhacker W, Flor H, Frith CD, Heinz A, Johnsen E, Kirschbaum C, Klingberg T, Lesch KP, Lewis S, Maier W, Mann K, Martinot JL, Meyer-Lindenberg A, Müller CP, Müller WE, Nutt DJ, Persico A, Perugi G, Pessiglione M, Preuss UW, Roiser JP, Rossini PM, Rybakowski JK, Sandi C, Stephan KE, Undurraga J, Vieta E, van der Wee N, Wykes T, Haro JM and Wittchen HU (2014) 'Stratified medicine for mental disorders', *European Neuropsychopharmacology*, **24**(1), pp. 5–50.

Strang G (2004) *Linear Algebra and Its Applications*. 4th edn. WB Saunders. doi: 10.2307/2003783.

Stroup T.S, Mcevoy JP, Swartz MS, Byerly MJ, Qlick ID, Canive JM, Mcqee MF, Simpson QM, Stevens MC and Lieberman JA (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.

Tenenbaum JB, de Silva V and Langford JC (2000) 'A global geometric framework for nonlinear dimensionality reduction.', *Science*, **290**(5500), pp. 2319–23.

Toh WL, Thomas N and Rossell SL (2015) 'Auditory verbal hallucinations in bipolar disorder (BD) and major depressive disorder (MDD): A systematic review', *Journal of Affective Disorders*, **184**, pp. 18–28.

Walker J, Curtis V and Murray RM (2002) 'Schizophrenia and bipolar disorder: similarities in pathogenic mechanisms but differences in neurodevelopment.', *International clinical psychopharmacology*, **17** Suppl 3, pp. S11–S19.

Wallwork RS, Fortgang R, Hashimoto R, Weinberger DR and Dickinson D (2012) 'Searching for a consensus five-factor model of the Positive and Negative Syndrome Scale for schizophrenia.', Schizophrenia research. NIH Public Access, **137**(1–3), pp. 246–50.

https://doi.org/10.1017/S0033291717001726 Published online by Cambridge University Press