Clinically meaningful biomarkers for psychosis: A systematic and quantitative review

Diana Prata*, Andrea Mechelli, Shitij Kapur

Department of Psychosis Studies, Institute of Psychiatry, King’s College London, King’s Health Partners, 16 De Crespigny Park, SE5 8AF, UK

A B S T R A C T

Despite five decades of search for clinically meaningful ‘biomarkers’ in schizophrenia there are still no common tests to inform diagnosis or treatment. Our aim was to understand why it has been so difficult to convert biological findings into clinical tests. We categorized all PubMed-indexed articles investigating psychosis-related biomarkers to date (over 3200). Studies showed an evident publication bias, a confusing array of terminology, and few systematic efforts at longitudinal evaluation or external validation. Fewer than 200 studies investigated biomarkers, longitudinally, for prediction of illness course and treatment response. These biomarkers were then evaluated in terms of their statistical reliability and clinical effect size. Only one passed our a priori threshold for clinical applicability. This is a modest record. In order to promote real progress, the field needs: (a) consistent use of terminology so that studies can be compared; (b) a system of standardized universal reporting to overcome the existing publication bias; and (c) practical criteria [a prototype is suggested here] for assessing the clinical applicability of the findings.

© 2014 Elsevier Ltd. All rights reserved.

1. Introduction ................................................................. 135
2. Methods ................................................................. 136
  2.1. Search and categorisation of studies .......................... 136
  2.2. Evaluation of the specific biomarkers ...................... 136
3. Results ................................................................. 137
  3.1. Search, categorisation and classification .................. 137
  3.2. Evaluation of specific biomarkers ............................ 138
4. Discussion .............................................................. 138
  4.1. A diverse, much discussed, but biased literature ........ 138
  4.2. Evaluation of the clinical applicability ..................... 138
  4.3. From potential clinical applicability to real-world clinical utility .................................................. 139
  4.4. Some recommendations for the future ..................... 140
    4.4.1. Clarity of terminology .................................. 140
    4.4.2. Registration of all biomarker trials .................. 140
    4.4.3. A standard framework for reporting and evaluating biomarkers .................................. 140

Acknowledgments ................................................................ 141
Appendix A. Supplementary data ........................................ 141
References ......................................................................... 141

* Corresponding author at: Department of Psychosis Studies, Institute of Psychiatry, 16 De Crespigny Park, London SE5 8AF, UK. Tel.: +44 2078480810.
E-mail address: diana.prata@kcl.ac.uk (D. Prata).

http://dx.doi.org/10.1016/j.neubiorev.2014.05.010
0149-7634/© 2014 Elsevier Ltd. All rights reserved.
1. Introduction

A search of the scientific literature for 'biomarkers' in psychosis brings up a few thousand articles spanning over half a century. Despite this large body of research, the use of biomarkers in drug development or clinical practice is still extremely limited. In theory, biomarker research should enhance the biological understanding of the illness, which should lead to better mechanism-driven biological therapeutics. However, the main purpose is the finding of biomarkers that can serve as clinical 'tests' that diagnose the disorder or predict outcome (be it prognosis or treatment response or monitoring). This review is focused on that aim: the use of biomarkers to develop meaningful clinical tests, in the context of psychosis. This focus is further narrowed to biomarkers that predict outcome, which we hereby designate by 'outcome' markers.

Our emphasis on outcome markers derives from them potentially being more useful to a clinician, more cost-effective to a health system, and more impactful in the patient’s wellbeing, than diagnostic ones, in psychosis. There is great and unpredictable variability in psychosis patients’ response to the same treatment, with devastating consequences, from persistence of symptoms, even after several drug treatment courses, to irreversible and/or life-threatening side effects, such as hyperlipidaemia, weight gain, diabetes, movement disorders, tardive dyskinesia, agranulocytosis and hyperprolactinemia, and, not surprisingly, high treatment discontinuation rates (Lieberman, 2007). Among the outcome biomarkers, 'predictive' biomarkers predict a response to a specific therapy, be it psychological or pharmacological, to help determine the optimal treatment in a stratified or personalized manner before it is commenced. This has the much-anticipated potential to reduce incidence of side effects and the often hit-and-miss efficacy of psychiatric treatments (an example in breast cancer has been the BRCA1/2 genetic type). ‘Predictive' biomarkers predict the natural course of the disease (Oldenhuis et al., 2008), ideally without any intervention (e.g. in cancer, it is tumour size or degree of metastasis). Thirdly, 'monitoring' markers, rather than measuring a particular endpoint, tag the current disease state, which is useful to monitor side effects and efficacy of ongoing treatments and infer expected progression (e.g. HbA1C in diabetes, or CD4 cell counts in chronic HIV). ‘Diagnostic' biomarkers have another purpose – they are biological tests used to ascertain the nature or presence of an illness. However, using biomarkers as 'diagnostic' tests poses particular challenges in psychiatry. The gold standard for a psychiatric diagnosis remains the DSM or ICD set of clinical signs and symptoms, but these criteria neither hypothesize a precise biological cause nor require a biological measure. When one combines this with the relatively modest inter-rater reliability of most diagnostic criteria in clinical practice (Goodman et al., 1984; Kitamura et al., 1989; Grove et al., 1981), it is not surprising that finding a one-to-one correspondence between a biological abnormality and a psychiatric diagnosis has been difficult. The literature is then replete with studies where a diagnostic biomarker shows a statistical difference between a group of patients with the ICD/DSM illness and some well-characterized normal controls – but such a differentiation is of little clinical utility. Insofar as the traditional clinically-defined diagnostic systems are used to identify and validate biological markers, it is unlikely that these could improve the existing diagnostic classification. Nevertheless, as discussed elsewhere (Kapur et al., 2012), it is plausible that diagnostic biomarkers themselves could be used to identify meaningful clinical sub-phenotypes. Such enhanced precision, if sustained on biological measures, could then make the search for outcome markers easier, with a given diagnostic sub-phenotype directly corresponding to a particular treatment outcome or prognosis. A successful example again in oncology has been in breast cancer: lumps were categorized based on different symptoms, until histopathological differentiation and molecular markers turned them into distinct illnesses subtypes (Luminal A, Luminal B, Triple negative/basal-like and the HER2+) (Arteaga et al., 2012). Now, besides of guiding diagnosis and untreated prognosis, this distinction also guides treatment decisions. There are indeed new exciting attempts, such as the NIMH Research Domain Criteria (RDoC) (Simmons and Quinn, 2013), at designing new multimodal dimensional biomarkers for mental illnesses, stemming from basic behavioural neuroscience research and disregarding current ICD/DSM categorization.

The search for biomarkers in psychosis has involved several areas of expertise, as shows a review by Lawrie et al. (2011). As diagnostic markers, the genetic ones, such as copy number variations (e.g. 22q11 deletion), chromosomal translocations (e.g. DISC1, COMT or NRC) and SNPs (e.g. ZNF804A) are objective, cheap, reliable and some have been replicated, but their applicability is low, given the small individual effect sizes and/or low prevalence. Brain imaging has been the most promising tool, identifying high effect sizes and replicability for the volume of hippocampus, ventricles and other areas, and white matter integrity and hypofrontality. The use of machine learning algorithms on this data has provided high prediction accuracies but their generalizability is still to be established. Electroencephalography of mismatch negativity has been found to have both high sensitivity and specificity, and is relatively inexpensive but not sufficiently replicated. In terms of treatment response, structural imaging however does not seem to help predict response to treatment as well, although functional imaging, such as reduced basal ganglia metabolism and increased striatal D2 receptor occupancy, has been repeatedly shown – however, clinical usefulness is still to be evaluated. As for metabolic markers, higher antipsychotic drug plasma levels and raised homovanillic acid (HVA) and other peripheral markers in plasma (and CSF) have been repeatedly related to treatment response, but replicability and accuracy is still unclear. Genetic markers in COMT, the 5-HT2A receptor, the DRD3 or the DRD2 gene have been implicated but only the latter has been consistent and, still, of small effect size. These qualitative reviews give a good feel for a breadth of investigations, with several potential leads – but do not provide a good sense of potential effect sizes, strength of association and clinical applicability.

A review of this field throws up a complex array of terminology: besides the precise above-defined biomarker types, the general term ‘biomarker’ is often but inconsistently interchanged with ‘intermediate phenotype’, ‘endophenotype’ and ‘surrogate endpoint’. We here provide their formal definitions. A ‘biomarker’ is a biologic characteristic objectively measured and evaluated as an indicator of normal or pathogenic processes; or of response to a treatment or challenge (Group, 2001). It can be identified at the molecular, cellular, organ or system levels. In a psychiatric biomarker definition, the ‘processes’ would be psycho-pathogenic, and the treatment generally either psychotropic medication or psychotherapy. (Specifically, in psychosis, pathogenic processes are still under scrutiny but there is the common stance that a common final pathway of different causes leads to an increased striatal dopamine tone. In terms of treatment, the one usually applicable to psychosis is antipsychotics administration, all consisting of, at least, D2 receptor blockade.) An ‘intermediate phenotype’ tends to be a systems-level biomarker and is termed an ‘endophenotype’ if it shows 5 well-defined pre-requisites that give indications of a strong genetic basis (Gottesman and Gould, 2003): (1) it is statistically associated with the illness, (2) it is heritable, (3) it is primarily state-independent, manifesting in an individual whether or not illness is active, (4) it segregates with the illness within families and (5) it is found in non-affected family members at a higher rate than in the general population. A ‘surrogate endpoint’ is an outcome that can substitute for (because it is so highly correlated with) the usual
clinical endpoint, which is how a patient feels, functions or survives. Since this clinical endpoint can be hard to collect, noisy or less sensitive to change, one can use a ‘surrogate endpoint’ which is more sensitive, more objective or simply easier, cheaper or more ethical to ascertain (Aronson, 2005). In this sense, surrogate endpoints are biomarkers that reach a high level of specificity and sensitivity. A good recent example in psychiatry is the use of b-amyloid plaque density which when measured with PET imaging can be used as a surrogate endpoint for monitoring progression of Alzheimer’s Disease (Doraiswamy et al., 2012).

While defining markers is easy – validating them and making them clinically useful is complex. A biomarker is validated by showing that its relationship to a certain clinical endpoint is reliable (statistically significant), plausible (causal or mechanistically understandable), accurate (sensitive and specific), and reproducible across clinically relevant settings. However, even valid, applicable, biomarkers are not useful until shown to provide a meaningful advantage when incorporated into decision-making or clinical care (Group, 2001). There is no simple algorithm to judge the validity of biomarkers, and judging their usefulness is even more contextual.

To assess the current status of clinically meaningful biomarkers in psychiatry, we first identified all relevant reports using standardized search methods. We then categorized all articles and subsequently classified the original data ones in diagnosis, prodrome or outcome studies. We further sub-categorized these according to their aims and according to their biological or non-biological nature. The outcome studies (predictive, prognostic or monitoring) were further distinguished according to their technical subtypes and their biomarker findings evaluated for quality of evidence (study design; statistical strength and reproducibility) and predictive power (effect size) using a two-dimensional scale adapted from Lassare (Lassere, 2008).

2. Methods

2.1. Search and categorisation of studies

We searched the PubMed database for extant literature published online up to January 2012, using the following search syntax: “(marker OR biomarker) AND (psychosis OR schizophrenia)”. This identified studies containing the term ‘marker’ or ‘biomarker’ in conjunction with ‘psychosis’ or ‘schizophrenia’ or related MeSH (Medical Subject Headings) terms such as ‘biological markers’ or ‘psychotic disorders’ in any of the fields searchable in PubMed.

The abstract for each article, and the full-text when necessary, was reviewed. Based on this, articles were categorized as reviews, meta-analyses, original data studies (‘prodrome’ studies and those regarding ‘diagnosis’, which are typically cross-sectional, and ‘outcome’ studies) or as ‘other’. The ‘other’ included all other studies that did not fit into any of the above categories of interest, e.g. investigating other illnesses or main effects of treatment, case-studies, methodological and animal studies or those not reported in English. Within the ‘outcome’ category which was our focus for discussion (typically longitudinal studies, i.e. prospective or retrospective), we distinguished between those for markers of: (i) prognosis of illness (i.e. having no consideration of specific treatments or time frames), (ii) prediction of treatment response before treatment initiation and (iii) monitoring of the treatment process. All studies in the ‘outcome’ category reporting biological markers were reviewed in order to more accurately distinguish them and further divide them according to the biological system/methodology used: genetic (e.g. DNA sequence variants, etc.), metabolic (e.g. blood/brain metabolites, protein and mRNA levels, etc.), imaging (e.g. MRI, PET, SPECT, EEG or eye-tracking) and demographic (e.g. sex or age). Although we appreciate that the borders between psychological, behavioural and neurological, or genetic and metabolic, are not clear-cut, we had to adopt a simple convention to classify the large number of possible biomarkers. We do not emphasize any differences or conclusions between these different classes of markers. These studies were then further reviewed in order for their biomarkers to be rated following criteria in the Evaluation section.

For quality control, sorting was verified independently by a second rater (AM) for 40 randomly chosen articles. There was no disagreement with the first rater (DP). Non-biological markers included those environmental, social or behavioural (including neurological soft signs), etc. In the rare cases where studies had both types, we treated them as biomarker studies, since these are the main focus of the review.

The articles (except reviews because they generally had a broad scope) were then further sub-divided. The ‘prodrome’ studies were divided into whether they focused on distinguishing the prodromal/at-risk state from illness and health (with a cross-sectional design) or on predicting the conversion from an at-risk state to a psychotic illness (with a longitudinal design). Within the general ‘diagnosis’ category, we differentiated the classic case–control diagnostic marker studies from those where the main focus was to detect intermediate phenotypes or subtypes within the patient population or to detect psychosis-related intermediate phenotypes or endophenotypes within the healthy population. Categorization of articles that reported more than one type of analysis was based on the main goal or main finding of the paper, as defined in the abstract.

2.2. Evaluation of the specific biomarkers

Lassere (Lassere, 2008) has proposed a sophisticated scheme to evaluate biomarkers, whereby validity is rated in terms of ‘study design’ (e.g. strength and quantity of evidence from in vitro, animal and trial studies), ‘target outcome’ (e.g. the reversibility or fatality of the outcome) and ‘statistical evaluation’ (e.g. the effect size among other issues), each in a 1–5 scale. There is also a ‘penalties’ domain, whereby points are discounted to permit considerations on biological plausibility, risk-benefit and generalizability. By summing each domain’s scores and any penalties, an overall 0–15 score is reached. Notwithstanding the rigour and completeness of Lassere’s system, it is, as the author acknowledges, iterative and cumbersome, and its implementation in current practice may not be easily attainable.

We used a much simpler version, fashioned after Lassere’s rating system and criteria (Lassere, 2008) which is also better adapted to psychiatric illnesses: a two-dimensional scale assessing quality of evidence (scored 1–4) and effect size (scored 1–4); criteria are delineated in Table 1. After a controlled status was established, the quality of evidence increased (from 1 to 4) as the hypotheses were more specific, trials prospectively designed, appropriately powered, and independently replicated.

There is no simple framework to decide how much of an effect size a biomarker in psychosis should have to be considered applicable. A recent report and online-tool by Uher et al. (Uher et al., 2012) has grappled with this issue in depression and proposed that if a biomarker can explain at least half-as much variance as the treatment itself – it would be of clinical interest (we come back to this issue, the number-needed-to-assess, NNA, in the Discussion). We know that the acute treatment of schizophrenia with antipsychotics has a standardized mean difference of 0.5 (Leucht et al., 2012). Using this as a guide and adapting suggestions by Bedard et al. (Bedard et al., 2007), we ranked biomarkers in terms of their odds ratios (O.R.), provided they scored at least 2 for quality of evidence, and anchored these to other effect size measures (Cohen, 1988) (Table 2). Finally, we declared biomarkers reaching...
3. Results

3.1. Search, categorisation and classification

The PubMed search resulted in an output of 3221 articles, all categorized in Fig. 1 Supplemental Table 1, the first one being published in 1965. Of these, 531 were a mixture of categories labelled as ‘other’, 260 studies related mainly to non-biological markers (further categorized in Supplemental Table 1) and 2430 to biomarkers. Of the latter, 608 were reviews, 22 meta-analyses,

<table>
<thead>
<tr>
<th>Quality of evidence</th>
<th>Approximate equivalence to drug effects:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grade</td>
<td>An observation of a positive result (p-value &lt; 0.05, corrected for multiple comparisons) in:</td>
</tr>
<tr>
<td>−</td>
<td>An uncontrolled study.</td>
</tr>
<tr>
<td>1</td>
<td>A study controlled for relevant extraneous variables (confounding, nuisance or effect modifiers), i.e. matched, restricted or adjusted for treatment, age, gender and – for genetic studies – ethnicity.</td>
</tr>
<tr>
<td>2</td>
<td>A study as above (grade 1), but with an explicit a priori intent to discover a precisely defined biomarker, i.e. with a given measure/modalities, cut off and direction of effect of both biomarker and response.</td>
</tr>
<tr>
<td>3</td>
<td>A study as above (grade 2), but designed with adequate power informed by previous positive studies of the same biomarker, i.e. replication in a larger cohort.</td>
</tr>
<tr>
<td>4</td>
<td>At least 2 studies as above (grade 3).</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Effect size</th>
<th>Approximate equivalence to drug effects:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grade</td>
<td>An observation of a positive result (p-value &lt; 0.05, corrected for multiple comparisons) with a:</td>
</tr>
<tr>
<td>−</td>
<td>Estimate from studies with quality of evidence &lt; 1.</td>
</tr>
<tr>
<td>1</td>
<td>Marginal effect (OR &lt; 1.3, SMD &lt; 0.2 or r &lt; 0.1).</td>
</tr>
<tr>
<td>2</td>
<td>Small effect size (OR 1.3–1.5, SMD 0.2–0.5 or r 0.1–0.3).</td>
</tr>
<tr>
<td>3</td>
<td>Medium effect size (possibly rivaling the drug effect; OR 1.5–2.0, SMD 0.5–0.8 or r 0.3–0.5).</td>
</tr>
<tr>
<td>4</td>
<td>Large effect size (possibly exceeding the usual drug effect; OR &gt; 2.0, SMD &gt;0.8 or r &gt;0.5).</td>
</tr>
</tbody>
</table>

OR = odds ratio; SMD = standard mean difference; r = correlation coefficient.

* In genetic studies, this only needs to be applied in relation to the outcome variable, provided that genotypes are unknown (and thus naturally randomized) before subjects’ inclusion, except ethnicity which affects linkage disequilibrium patterns.

a sum score of 6 (out of 8) as clinically ‘applicable’, i.e. particularly worthy of clinical consideration. Although the final score increases with effect size, the choice of 6 was not completely arbitrary: this threshold assures that biomarkers arising from studies which were uncontrolled or did not use an a priori defined measure, cut-off or direction, are never considered clinically applicable, regardless of their effect size. All biomarkers’ evaluation was independently verified by 2 raters (DP and AM), blind to each other’s rating, and the 3.8% discrepancy was agreed and corrected upon discussion.
31 were prodromal studies, 1601 focused on diagnostic biomarkers or related intermediate phenotypes, and 168 on prediction of clinical outcome in psychosis patients (further details in Supplemental Table 1).

3.2. Evaluation of specific biomarkers

All the 168 ‘outcome’ biomarker studies’ categorization and bibliography are available as Supplemental Results. We identified 53 studies on ‘prognostic’ biomarkers of a psychotic illness with imaging (N = 28) or metabolic (N = 25) biomarkers. Given that none of these studies was controlled for treatment (i.e. what medication was used, if at all and for how long), the specific biomarkers could not be scored according to our scale; they did not have enough specificity for clinical application.

The 70 studies of ‘predictive’ biomarkers (for specific treatments) were on genetic (N = 52, 43 of which positive, i.e. reporting at least one statistically significant finding, at p < 0.05), metabolic (N = 11, 7 of which positive), imaging (N = 4, all positive) or demographic (N = 3, all positive) biomarkers. The 45 studies on ‘monitoring’ biomarkers (for specific treatments) were mostly metabolic (N = 38, all positive except one) and a few, imaging (N = 7, 4 of which positive); rating results are in Table 2.

The only biomarker with a final score above 6 from the total of 362 predictive & monitoring biomarkers in the 114 studies was a pharmacogenetic biomarker that scored 7: the C allele of the 6672G>C single nucleotide polymorphism (SNP) in the HLA-DQB1 region (Athanasiou et al., 2011) predicted risk for clozapine-induced agranulocytosis with an O.R. 16.8, was defined a priori and its effect replicated in an independent sample.

4. Discussion

First we wish to highlight this review’s limitations. Of necessity we restricted our search to studies using the keyword ‘marker’ or ‘biomarker’ because that was our initial intent. However, as there is a very diverse use of terminology, it would not be surprising that articles with relevant findings may not have used the word ‘biomarker’ or ‘marker’ and thus may have been missed. Therefore, we recommend our quantitative, systematic and term-driven search be complemented by expert reviews, such as Lawrie et al. (2011) which provides a summarized appreciation of the most relevant diagnostic, prognostic and predictive (bio)markers findings in schizophrenia – though it has to be said that none reaches our 6/8 applicability threshold. We also acknowledge that since this is a rapidly developing field, with obvious commercial implications, there could be a number of other findings relevant to this issue that are not published for commercial reasons. We can only speculate about their existence as a number of companies are making claims of such biomarkers (e.g. Veripsych®, AssureRxHealth®, SureGene® and Genomind®). Finally, we observed indications of a high publication bias in the literature – which raises serious concerns we discuss below.

4.1. A diverse, much discussed, but biased literature

A review of the publications in the field provides some very interesting observations: 20% of all articles we retrieved were reviews (meta-analyses were 0.8%), whereas only 62% were original data, with the rest belonging to the ‘other’ category. This means that there is a review for only every 3 original articles published. In psychosis, more generally, the ratio of reviews to primary articles is usually lower, 1:10 (6407 reviews for 55315 articles as per PubMed).

Within the original data-driven articles on markers for psychosis (prodrome studies aside), the vast majority (90%) was concerned with finding markers to cross-sectionally distinguish patients with schizophrenia/related disorders from matched controls. These have improved our understanding of the genetic, metabolic, neuronal and environmental basis of psychosis, but do not allow us to find outcome biomarkers, which by definition require a longitudinal design.

Unfortunately, there is reason to suspect that outcome biomarkers suffer from publication bias for positive results (Andre et al., 2011). For example, only 16 in the 115 monitoring and predictive marker studies combined reported only negative results. That biological psychiatry may be particularly subject to publication bias has been raised before (Jennings and Van Horn, 2012) and this fosters significant problems. Just relying on the published literature prevents us factoring in non-replications of biomarkers in a grading scale like the one proposed and impairs an accurate and contextualized interpretation of the overall results (Ioannidis, 2011).

4.2. Evaluation of the clinical applicability

Only 4% of the original data-driven studies were focussed on predictive biomarkers, and 2–3% on monitoring and prognostic ones. The predictive ones are those with the greatest translational potential as they can inform treatment choice before treatment initiation and thus could lead to the development of stratified (Trusheim et al., 2007) or precision medicine (Committee NRCU, 2011). Only 1 out of the 257 predictive biomarkers (the vast majority genetic (91%) passed our threshold of clinical applicability (6/8) (Table 2). A third of them had not been tested in a controlled design (see Table 1) and thus were not given any score. A common occurrence was the matching of healthy volunteers (when used) and patients, for extraneous variables while, which is more important, no account of these variables’ relationship with either the biomarker or the clinical endpoint was made. The rest of the biomarkers were discovered in controlled studies (67%), but, except one, were not explicitly defined a priori in terms of their direction or specific cut-off, and thus only got a quality of evidence score of 1. For example, there was often no a priori prediction of the specific allele, or the direction of a change in a drug metabolite, which would be associated with an improved response. Or, when the response was measured with a continuous scale, there was no predefined cut-off point for response. Thus, these could only be considered as exploratory. The biomarker which passed our quality and effect size threshold was the HLA-DQB1 6672G>C SNP (Athanasiou et al., 2011) as a predictor of clozapine induced agranulocytosis. The contribution offered by genome-wide association (GWA) studies for the development of outcome biomarkers, although promising, is still limited. The 4 studies we found focused on the association of individual polymorphisms with treatment response rather than any combination. (That said, there has since been an effort, still in its early days, to build polygenic risk scores in GWA studies for psychosis and its markers; McIntosh et al., 2013) That was either not hypothesis-based (no predefined gene or allele) or reported very small effect sizes, and thus also did not survive our threshold of clinical applicability.

The HLA-DQB1 6672G>C SNP biomarker (Athanasiou et al., 2011) predicts risk of agranulocytosis, a serious and potentially fatal side-effect which results from clozapine administration and necessitates a precautionary weekly/biweekly blood monitoring in all patients (Athanasiou et al., 2011). Psychiatric patients find this blood-draw particularly onerous, and many refuse to try clozapine for the fear of blood draws. According to the published (and replicated) data, the CC individuals have a greatly increased risk of agranulocytosis (O.R. 16.8) and conversely the GG individuals have a 99.7% chance of being agranulocytosis-free thus raising the possibility of a much lighter monitoring in these individuals (Athanasiou
Furthermore, this finding was consistent with previous studies implicating HLA variants (Dettling et al., 2007).

4.3. From potential clinical applicability to real-world clinical utility

While we screened biomarkers for potential clinical applicability using the criteria in Table 1, this by itself does not assure real world clinical utility. Effect sizes are certainly a better proxy of clinical efficacy than statistical significance (p-values) (Perlis, 2011) but their clinical usefulness depends on the clinical context (Cohen, 1988). A biomarker test with a small effect size that would prevent a fatal side effect is more meaningful than one with a large effect size that would prevent a transitory rash. Our consideration here was based on it being statistically reliable and clinically applicable. Nevertheless, the real-world usefulness of a biomarker arises from its ability to generate a ‘divergent prediction’ against the standard of care. In this regard, the HLA-DQB1 SNP is a particularly interesting case as an O.R. of 16.8 for predicting a serious side effect and detecting a genotype group with an 1175% increased risk, appears very promising at first. However, when this is seen in the light that the base rate of this side-effect is rather low (present in only 0.4% of clozapine users), the test shows only a limited sensitivity of 21.5% (i.e. it will pick up only a fifth of all the cases) and a limited positive predictive value of 5% (i.e. of all those positive, only 5% will actually develop agranulocytosis). Also, while the HLA-DQB1 genetic marker has a high negative predictive value (99.7%), when compared to the negative base rate of 99.6% – it is not a sufficient improvement to obviate regular blood monitoring. So, here then is a case of a finding that is well replicated, with a remarkable O.R., and yet its use is unlikely to lead to a divergent course of action, i.e. of being clinically useful. In conclusion, after a marker has passed our preliminary test of clinical applicability one has to articulate the divergent prediction that the use of the test would make and how it would be superior to the standard practice.

If, for example, the HLA test above had an even higher negative predictive value, say 99.95%, one could envisage foregoing regular blood monitoring in GG individuals thus creating a new, more desirable, clinical pathway. The efficacy of such a choice can then be encapsulated by a simple index – the number-needed-to-assess (NNA) to result in one more positive outcome along the new (biomarker) pathway compared to the standard pathway. The NNA is analogous to the index number-needed-to-treat (NNT) which is well established as a measure of the positive effects of a new treatment (Uher et al., 2012). The NNA and the clinical advantage of the alternative action needs to be balanced against the cost of administration, potential risks/side-effects, burden and delays associated with biomarker testing (Perlis, 2011; Perlis et al., 2009; Simon and Perlis, 2010). Obviously, no simple or universal formulas can be developed for this – therefore we provide a broad conceptual
4.4. Some recommendations for the future

The lack of clinically applicable biomarker findings may be partially due to the methodological complexity of the neuroscience field. It poses design difficulties (large sample sizes needed to assess the multifactorial nature of the causes that explain the variance in outcome) and technological challenges (insufficient systems-level understanding of the underlying physiology and computational power), which have been widely discussed in the field (Kapur et al., 2012; Perlis, 2011). Nevertheless, in addition to our proposed clinical applicability rating system, there are a number of improvements at the reporting level that can optimize the delivery of biomarker research findings.

4.4.1. Clarity of terminology

Categorizing the studies we retrieved showed a wide variation in the use of the term “predictive” biomarker or marker. Also, studies measuring the main effect of medication on phenotypes or surrogate endpoints were often misleadingly referred to as biomarker studies. In conclusion, to contribute to easy, accurate and comparable searches of the literature, authors should make clear in the title or abstract whether the biomarker they report is for diagnosing an illness (though we pointed out the circularity of this), for predicting treatment response (i.e. predictive biomarkers); for predicting general prognosis (i.e. prognostic biomarkers) or for monitoring an outcome’s progression (i.e. monitoring biomarker). Other more ambiguous designations such as intermediate outcomes, intermediate endpoints, leading indicators, surrogate intermediate endpoints and early markers of response may be best avoided, as suggested elsewhere (Lassere, 2008). A consistent terminology does not produce better findings, but will ensure that the field discusses them more precisely and completely.

4.4.2. Registration of all biomarker trials

Just as one needs all the pharmacological clinical data in the public domain in order to draw further cost-effective studies and unbiased conclusions regarding the efficacy of drugs – so it should be for biomarkers. The publication bias noted is quite likely to have given us a rosier-than-real picture of the evidence; and unfortunately may lead to more researchers devoting time to explore biomarkers which are already known to be a dead-end (but not yet published). Even if it may not be possible to make all biomarker data published, it should be possible to have all serious efforts declared. One may take a page from the developments in the area of clinical trial databases and, as such, grant agencies and editors should make pre-registration of trials a condition of their funding and publication consideration, respectively. Where possible, authors should also be encouraged to deposit a summary of results upon completion (Reveiz et al., 2007).

4.4.3. A standard framework for reporting and evaluating biomarkers

One thing that became evident in reviewing hundreds of the articles was the wide variation in study quality, but more troubling was the inconsistent reporting. Often it was difficult to know what controls were used, if certain extraneous variables were taken into account, precisely what the biomarker was, whether it was a priori or exploratory, and there were few summary statistics about effect size, sensitivity, specificity and positive/negative predictive values. This is wasted effort. The CONSORT (CONsolidation of Standards for Reporting Trials) guidelines have had a salutary effect on the reporting, and subsequent evaluation of clinical trials (Novack, 2004). These were developed by researchers and journal editors to help authors improve the standardization and clarity of clinical trial reports. This is done by following a checklist of essential information in each section of the reports and including a flow diagram with number of participants in the different phases of the trial. A similar checklist-and-flowchart system has also been proposed for the reporting of diagnostic biomarkers: a number of journals have adopted the STARD (Standards for Reporting of Diagnostic Accuracy) framework (Bossuyt et al., 2003). It is an important effort to improve the accuracy of these reports to allow for a better estimation of internal and external validity. The latter is increasingly being adapted for cardiology, oncology and radiology. It should be possible to articulate a psychiatric version of the STARD or some other alternative, so that – as the literature in psychiatric biomarkers proliferates – it does so in an orderly way that is more amenable to review and analysis.

Systems to evaluate and rank biomarkers are in their infancy. While Lassare provides a good starting point, it has not been extensively used in psychiatry, to our knowledge. Our adaptation of that method to psychosis is a first such effort. In theory, the framework we have developed here could be of value in other psychiatric disorders, particularly in those in whom treatment outcomes are heterogeneous (e.g. depression). We do not expect that what we have proposed here is the final answer, as we acknowledge in our limitations, but we do hope the lessons learnt herein will inform
future biomarker research and will spur efforts to develop systems to assess them systematically.

In summary, the literature on psychosis biomarkers is diverse, full of reviews, with relatively limited primary data and biased. Only one of hundreds of outcome prediction biomarkers demonstrated clinical applicability, and even in this case there is little data on real-world clinical utility. To make faster progress, the field should adopt a consistent terminology, require the registration of all biomarker trials, and adopt a consistent approach to reporting the results. We hereby also propose a checklist system for rating reported biomarker findings in terms of their clinical applicability. These measures do not produce biomedical breakthroughs per se but make it much more likely that we will recognize a breakthrough when we have one.

Acknowledgments

This research was supported by a National Institute for Health Research (NIHR) fellowship (NIHR-PDF-2010-03-047) to DP and a grant (99859) from the Medical Research Council (MRC) to AM and SK. SK receives also funding from the NIHR Mental Health Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King’s College London. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health. SK has served as ad-hoc consultant and/or speaker for AstraZeneca, Bioline, Bristol Meyers Squibb, Eli Lilly, Enviio, Janssen – Johnson and Johnson, NeuroSearch, Otsuka, Pfizer, Roche, Sanofi-Aventis, Schering-Plough, Servier, Solvay Wyeth and serves on the Scientific Advisory Boards for Lundbeck and Roche. DP and AM report no competing or financial relationships with commercial interests during the past 3 years. DP had full access to all the data and takes responsibility for its integrity and the accuracy of the data analysis. All authors contributed significantly either to conception and design (SK, DP, AM) or analysis (DP, AM) and interpretation of data (SK, DP, AM), and to drafting (DP) or revising (SK, DP, AM) and final approval (SK, DP, AM) of the article. DP had access to all reported and unreported data from the study as well as complete freedom to direct its analysis and reporting, without influence or censorship from sponsors. Three files of supplemental data as Supplemental material are provided.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.neubiorev.2014.05.010.

References


