Assessment of Response to Lithium Maintenance Treatment in Bipolar Disorder: A Consortium on Lithium Genetics (ConLiGen) Report

Mirko Manchia[1], Mazda Adli[2], Nirmala Akula[3], Raffaella Ardau[4], Jean-Michel Aubry[5], Lena Backlund[6], Claudio E.M. Banzato[7], Bernhard T. Baune[8], Frank Bellivier[9], Susanne Bengesser[10], Joanna M. Biernacka[11], Clara Brichant-Petitjean[9], Elise Bui[3], Cynthia V. Calkin[1], Andrew Tai Ann Cheng[12], Caterina Chillotti[4], Sven Cichon[13], Scott Clark[8], Piotr M. Czerski[14], Clarissa Dantas[7], Maria Del Zompo[4,15], J. Raymond DePaulo[16], Sevilla D. Detera-Wadleigh[3], Bruno Etain[17], Peter Falkai[18], Louise Frisén[6], Mark A. Frye[11], Jan Fullerton[19], Sébastien Gard[20], Julie Garnham[1], Fernando S. Goes[16], Paul Grof[21,22], Oliver Gruber[23], Ryota Hashimoto[24], Joanna Hauser[25], Urs Heilbronner[23], Rebecca Hoban[26,27], Liping Hou[3], Stéphane Jamain[17], Jean-Pierre Kahn[28], Layla Kassem[3], Tadafumi Kato[29], John R. Kelsoe[26,27], Sarah Kittel-Schneider[30], Sebastian Kliwicki[31], Po-Hsiu Kuo[32], Ichiro Kusumi[29,33], Gonzalo Laje[3], Catharina Lavebratt[34], Marion Leboyer[17], Susan G. Leckband[26,35,36], Carlos A. López Jaramillo[37], Mario Maj[38], Alain Malafosse[5], Lina Martinsson[6], Takuya Masui[33], Philip B. Mitchell[39], Frank Mondimore[16], Palmiero Monteleone[38], Audrey Nallet[5], Maria Neuner[30], Tomás Novák[40], Claire O'Donovan[1], Urban Ösby[34], Norio Ozaki[41,42], Roy H. Perlis[43], Andrea Pfennig[44], James B. Potash[16,45], Daniela Reich-Erkelenz[23], Andreas Reif[30], Eva Reininghaus[10], Sara Richardson[3], Guy A. Rouleau[46], Janusz K. Rybakowski[31], Martin Schalling[34], Peter R. Schofield[19], Oliver K. Schubert[8], Barbara Schweizer[16], Florian Seemüller[18], Maria Grigoroiu-Serbanescu[47], Giovanni Severino[15], Lisa R. Seymour[11], Claire Slaney[1], Jordan W. Smoller[43], Alessio Squassina[15], Thomas Stamm[2], Jo Steele[3], Pavla Stopkova[40], Sarah K. Tighe[16], Alfonso Tortorella[38], Gustavo Turecki[48], Naomi R. Wray[49], Adam Wright[39], Peter P. Zandi[50], David Zilles[23], Michael Bauer[44], Marcella Rietschel[51], Francis J. McMahon[3], Thomas G. Schulze[3,16,23,51]*¶, Martin Alda[1]*¶

1 Department of Psychiatry, Dalhousie University, Halifax, Nova Scotia, Canada, **2** Department of Psychiatry and Psychotherapy, Charité Universitätsmedizin, Berlin, Germany, **3** Human Genetics Branch, Division of Intramural Research Programs, National Institute of Mental Health (NIMH), National Institutes of Health (NIH), Bethesda, Maryland, United States of America, **4** Unit of Clinical Pharmacology, University-Hospital of Cagliari, Cagliari, Italy, **5** Hôpitaux Universitaires de Genève, Department of Mental Health and Psychiatry, Geneva, Switzerland, **6** Department of Clinical Neuroscience, Karolinska Institutet, Stockholm, Sweden, **7** Department of Psychiatry, University of Campinas, Campinas, Brazil, **8** Department of Psychiatry, The University of Adelaide, Adelaide, Australia, **9** Assistance publique - Hôpitaux de Paris, Groupe Hospitalier Lariboisière-F. Widal, Pôle de Psychiatrie, Paris, France, **10** Department of Psychiatry, Medical University of Graz, Graz, Austria, **11** Department of Psychiatry, Mayo Clinic, Rochester, Minnesota, United States of America, **12** Division of Epidemiology and Genetics, Academia Sinica, Institute of Biomedical Sciences, Taipei, Taiwan, **13** Department of Genomics, Life and Brain Center and Institute of Human Genetics, Bonn University, Bonn, Germany, **14** Psychiatric Genetic Unit, Poznan University of Medical Sciences, Poznan, Poland, **15** Section of Neuroscience and Clinical Pharmacology, Department of Biomedical Science, University of Cagliari, Cagliari, Italy, **16** Department of Psychiatry and Behavioral Sciences, Johns Hopkins University School of Medicine, Baltimore, Maryland, United States of America, **17** Institut National de la Santé et de la Recherche Médicale, Unité 955, Institut Mondor de Recherche Biomédicale, Equipe 15, Faculté de médecine, Créteil, France, **18** Department of Psychiatry and Psychotherapy, Ludwig Maximilian University, Munich, Germany, **19** Neuroscience Research Australia - Genetics of Mental Illness and Brain Function, Sydney, Australia, **20** Service de psychiatrie, Hôpital Charles Perrens, Bordeaux, France, **21** Mood Disorders Center of Ottawa, Ottawa, Canada, **22** Department of Psychiatry, University of Toronto, Toronto, Canada, **23** Department of Psychiatry and Psychotherapy, Georg-August-Universität, Göttingen, Germany, **24** Osaka University Graduate School of Medicine, Osaka, Japan, **25** Psychiatric Genetic Unit, Poznan University of Medical Sciences, Poznan, Poland, **26** Department of Psychiatry, University of California San Diego, San Diego, California, United States of America, **27** Department of Psychiatry, Veterans Affairs San Diego Healthcare System, San Diego, California, United States of America, **28** Service de Psychiatrie et Psychologie Clinique, Centre Hospitalier Universitaire de Nancy, Nancy, France, **29** Laboratory for Molecular Dynamics of Mental Disorders, RIKEN Brain Science Institute, Saitama, Japan, **30** Department of Psychiatry, Psychosomatics, and Psychotherapy, University of Würzburg, Würzburg, Germany, **31** Department of Adult Psychiatry, Poznan University of Medical Sciences, Poznan, Poland, **32** Institute of Epidemiology and Preventive Medicine, National Taiwan University, Taipei, Taiwan, **33** Department of Psychiatry, Hokkaido University Graduate School of Medicine, Sapporo, Japan, **34** Department of Molecular Medicine and Surgery, Karolinska Institutet and Center for Molecular Medicine, Karolinska University Hospital, Stockholm, Sweden, **35** Department of Pharmacy, Veterans Affairs San Diego Healthcare System, San Diego, California, United States of America, **36** Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California San Diego, San Diego, California, United States of America, **37** Department of Psychiatry, University of Antioquia, Medellin, Colombia, **38** Department of Psychiatry, University of Napoli, Napoli, Italy, **39** School of Psychiatry, University of New South Wales, and Black Dog Institute, Sydney, Australia, **40** Prague Psychiatric Center, University of Prague, Prague, Czech Republic, **41** Department of Psychiatry, Fujita Health University School of Medicine, Toyoake, Japan, **42** Department of Psychiatry, Nagoya University Graduate School of Medicine, Nagoya, Japan, **43** Department of Psychiatry, Massachusetts General Hospital and Harvard Medical School, Boston, Massachusetts, United States of America, **44** Department of Psychiatry and Psychotherapy, Technische Universität Dresden, Germany, **45** Department of Psychiatry, University of Iowa, Iowa City, Iowa, United States of America, **46** Centre of Excellence in Neuroscience of Université de Montréal, Centre Hospitalier de l'Université de Montréal and Department of Medicine, Université de Montréal, Montréal, Canada, **47** Alexandru Obregia Psychiatric Hospital, Biometric Psychiatric Genetics Research Unit, Bucharest, Romania, **48** McGill Group for Suicide Studies, Douglas Mental Health University Institute, Montréal, Canada, **49** The University of Queensland, Queensland Brain Institute, Brisbane, Australia, **50** Johns Hopkins Bloomberg School of Public Health, Department of Mental Health, Baltimore, Maryland, United States of America, **51** Department of Genetic Epidemiology in Psychiatry, Central Institute of Mental Health Mannheim, University Medical Center Mannheim, University of Heidelberg, Mannheim, Germany

Abstract

Objective: The assessment of response to lithium maintenance treatment in bipolar disorder (BD) is complicated by variable length of treatment, unpredictable clinical course, and often inconsistent compliance. Prospective and retrospective methods of assessment of lithium response have been proposed in the literature. In this study we report the key phenotypic measures of the ''Retrospective Criteria of Long-Term Treatment Response in Research Subjects with Bipolar Disorder'' scale currently used in the Consortium on Lithium Genetics (ConLiGen) study.

Materials and Methods: Twenty-nine ConLiGen sites took part in a two-stage case-vignette rating procedure to examine inter-rater agreement [Kappa (κ)] and reliability [intra-class correlation coefficient (ICC)] of lithium response. Annotated first-round vignettes and rating guidelines were circulated to expert research clinicians for training purposes between the two stages. Further, we analyzed the distributional properties of the treatment response scores available for 1,308 patients using mixture modeling.

Results: Substantial and moderate agreement was shown across sites in the first and second sets of vignettes ($\kappa = 0.66$ and $\kappa = 0.54$, respectively), without significant improvement from training. However, definition of response using the A score as a quantitative trait and selecting cases with B criteria of 4 or less showed an improvement between the two stages ($ICC_1 = 0.71$ and $ICC_2 = 0.75$, respectively). Mixture modeling of score distribution indicated three subpopulations (full responders, partial responders, non responders).

Conclusions: We identified two definitions of lithium response, one dichotomous and the other continuous, with moderate to substantial inter-rater agreement and reliability. Accurate phenotypic measurement of lithium response is crucial for the ongoing ConLiGen pharmacogenomic study.

Citation: Manchia M, Adli M, Akula N, Ardau R, Aubry J-M, et al. (2013) Assessment of Response to Lithium Maintenance Treatment in Bipolar Disorder: A Consortium on Lithium Genetics (ConLiGen) Report. PLoS ONE 8(6): e65636. doi:10.1371/journal.pone.0065636

Editor: Kazutaka Ikeda, Tokyo Metropolitan Institute of Medical Science, Japan

Received January 24, 2013; **Accepted** April 26, 2013; **Published** June 19, 2013

Funding: The work on assessment of lithium response has been supported by a grant from Canadian Institutes of Health Research #64410 to MA. MG-S was supported by Unitatea Executiva pentru Finantarea Invatamantului Superior, a Cercetarii, Dezvoltarii si Inovarii (UEFISCDI), Bucharest, Romania (grant no. 89/2012). JMA and AN were supported by a grant from the Swiss National Foundation (#32003B_125469/1 to JM Aubry). ConLiGen is in part supported by funds from the Intramural Research Program of the National Institute of Mental Health (NIMH) at the National Institutes of Health (NIH), Department of Health and Human Services, United States Government. It is further supported by a grant from the Deutsche Forschungsgemeinschaft to MR, MB, and TGS (RI 908/7-1). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: Co-authors James B. Potash, Andreas Reif, and Bernard T. Baune are PLOS ONE Editorial members. This does not alter the authors' adherence to all the PLOS ONE policies on sharing data and materials.

* E-mail: thomas.schulze@med.uni-goettingen.de (TS); malda@dal.ca (M. Alda)

◔ These authors contributed equally to this work.

Introduction

Bipolar disorder (BD) is a lifelong and severe psychiatric illness characterized by recurrences of episodes of depression and hypomania/mania [1]. Lithium is among the first-line maintenance treatments for BD [2,3], preventing relapses and recurrences of opposite polarity. In addition, lithium decreases the risk of suicidal behaviour and all-cause mortality in mood disorders [4–6].

Naturalistic analyses show that approximately one third of BD patients achieve complete remission on lithium [7–14]. Lithium-responsive BD patients have distinct clinical features, such as episodicity of clinical course [15], absence of rapid cycling [16], and a family history of BD [17], corresponding to the BD ''core phenotype'' [18].

Despite a significant genetic component for lithium-responsive BD [12,19], pharmacogenetic studies have not produced replicated results [20,21]. One possible explanation for the lack of conclusive pharmacogenetic findings is the varying definition of lithium response across the studies. Indeed, the assessment of lithium maintenance treatment response, and consequently the definition of the phenotype under study, is complicated by factors inherent to the natural history of BD. The irregular clinical course of BD [22] as well as variable treatment adherence [23] are only few of the factors that contribute to the complexity in assessing the response to lithium maintenance treatment.

To reduce the impact of the clinical heterogeneity of BD in pharmacogenetics (and possibly to define genetically more homogeneous subgroups of BD patients), researchers have proposed to select prospectively followed patients on lithium monotherapy with unequivocal clinical response [24,25]. However, this may not be practical if large patient samples are needed. In such cases, we need to rely on retrospective evaluation of treatment response. Several such methods have been described in the literature including the Affective Morbidity Index (AMI) [26] and the Illness Severity Index [27]. The AMI takes into account the duration and the severity of an episode, the latter scored on a 4-point scale (0 = no conspicuous affective disturbance, 1 = mild depression or mania, 2 = moderate depression or mania, 3 = severe depression or mania). The area under the curve can be calculated from these two variables and compared between defined treatment periods. Similarly, the Illness Severity Index measures the efficacy of lithium treatment in controlling mood episodes. It is defined as the frequency of affective episodes prior to

Table 1. Number of raters from the Consortium on Lithium Genetics (ConLiGen) centres participating in the two-stage case-vignette rating procedure for inter-rater reliability and agreement.

ConLiGen centres	First stage	Second stage
University of Adelaide, Adelaide (Australia)	1	1
University of Sydney, Sydney (Australia)	1	0
University of Graz, Graz (Austria)	3	3
University of Campinas, São Paulo (Brasil)	3	3
Dalhousie University, Halifax (Canada)	9	2
University of Medellin, Medellin (Colombia)	4	4
Charles University, Prague (Czech Republic)	1	2
Institut national de la santé et de la recherche médicale, Paris (France)	1	1
University of Würzburg, Würzburg (Germany)	2	1
University of Göttingen, Göttingen (Germany)	2	0
Charité - Universitätsmedizin, Berlin (Germany)	1	2
Technische Universität Dresden,Dresden (Germany)	2	2
University of Cagliari, Sardinia (Italy)	3	3
University of Naples SUN, Naples (Italy)	1	2
The Japanese Collaborative Group on the Genetics of Lithium Response in Bipolar Disorder (Japan)*	4	4
University of Medical Sciences, Poznań (Poland)	2	2
Obregia Psychiatric Hospital, Medical University, Bucharest (Romania)	2	2
Karolinska Institutet, Stockholm (Sweden)	1	1
University of Geneva, Geneva (Switzerland)	3	2
Academia Sinica, Taipei (Taiwan)	1	1
National Taiwan University, Taipei (Taiwan)	2	2
National Institute of Mental Health (USA)	4	2
The Johns Hopkins University, Baltimore (USA)	7	5
Mayo Clinic, Rochester (USA)	6	1
Massachusetts General Hospital, Boston (USA)	2	0
University of California, San Diego (USA)	2	0
Total number of raters	70	48

ConLiGen: Consortium on Lithium Genetics.
*Hokkaido, Osaka, Tokio, Riken Brain Science Institute.
doi:10.1371/journal.pone.0065636.t001

starting lithium adjusted for age at the time lithium was started [27]. However, changes of affective morbidity might be not only a result of the treatment, but could be due to other factors. In the Consortium on Lithium Genetics (ConLiGen, www.ConLiGen.org) study [28], we adopted the "Retrospective Criteria of Long-Term Treatment Response in Research Subjects with Bipolar Disorder" as the principal method of evaluation of the response to lithium [12,13]. In addition to measuring the degree of clinical improvement, this scale weighs clinical factors considered relevant in determining whether the observed clinical change is in fact due to the lithium treatment.

Since ConLiGen is an international multi-centre collaboration, it has been crucial to assess the key phenotypic measures and the response to long-term lithium treatment reliability across the participating research groups. Here we present: 1) the results of the reliability analysis of response to lithium treatment across the participating centres, and 2) the distributional properties of the scale scores. These two sets of findings have been instrumental in obtaining stringent phenotypic definitions of lithium response. These analyses are of particular importance in light of the genome-wide association study (GWAS) currently being undertaken by ConLiGen.

Materials and Methods

Assessment of Clinical Response to Lithium Treatment

The response to lithium treatment was measured using a previously published and validated rating scale: the "Retrospective Criteria of Long-Term Treatment Response in Research Subjects with Bipolar Disorder" [12,28]. Briefly, this scale quantifies the degree of improvement in the course of treatment (A criterion or A score) expressed as a composite measure of change in frequency and severity of mood symptoms. The A score is weighed against 5 factors (B criteria) which allow one to determine if the observed improvement is a result of the treatment rather than a spontaneous improvement or an effect of additional medication. Specifically, the B criteria consider: the number of episodes before/off the treatment (B1), the frequency of episodes before/off the treatment (B2), the duration of the treatment (B3), the compliance during period(s) of stability (B4) and the use of additional medication during the period of stability (B5). The total score (TS) is obtained by subtracting the B score from the A score.

Analysis of the Inter-rater Agreement and Reliability of the Assessment of Lithium Response

The agreement and reliability of the assessment of lithium response between raters of 29 ConLiGen participating centres was measured using a two-stage case-vignette rating procedure (Table 1). Specifically, the study protocol had three phases: 1) twelve standardized case vignettes prepared by investigators (M.A., J.G., C.S.) at Dalhousie University were circulated and rated by 70 investigators; 2) annotated first-round vignettes and rating guidelines were circulated for training purposes after the first stage; 3) sixteen additional more complex vignettes prepared by senior researchers at Dalhousie University, Johns Hopkins University School of Medicine, National Institute of Mental Health (NIMH) and Academia Sinica of Taiwan (M.A., J.G., J.P., T.G.S., F.M., A.C.) were circulated and rated by 48 investigators at the participating sites. The first set of vignettes was based exclusively on BD patients who had been prospectively followed in a specialty program and with detailed clinical information on the course of illness and treatment history. The second set of vignettes was heterogeneous and included patients treated in various settings, some with limited clinical details assessed cross-sectionally. Since raters had no prior knowledge of the rating scale, this design allowed us to estimate the impact of training on agreement and reliability of lithium response assessment. The rating procedure was performed from April 2009 to October 2012.

The degree of concordance of lithium response definition was assessed with Cohen's kappa (κ) [29] and intra-class correlation (ICC) coefficient [30]. These analytical methods were applied to the dichotomous and continuous definition of lithium response, respectively. The κ statistics (multiple raters with two outcomes) were calculated with 95% confidence interval (CI) for each cut off point of the TS scale in the range from 3 (non response to lithium) to 8 (full response to lithium). Interpretation of the strength of agreement was made according to Landis and Koch: poor (κ <0.00), slight (0.00–0.20), fair (0.21–0.40), moderate (0.41–0.60), substantial (0.61–0.80), almost perfect (0.81–1.00) [31].

The quantitative scores of the treatment response scale were analyzed in the first (ICC_1) and second (ICC_2) stage of ratings. Specifically, we analyzed the TS (weighted clinical improvement), the A score (uncorrected clinical improvement), the B score (quantification of confounders), and the A score when B score ≤ 4. The latter measure allows the identification of "valid cases" through selection at the B criteria. Subjects with B score ≤ 4 are likely to have a clinical improvement causally related to lithium treatment. The ICC was tested with the two-way random effects model, that assumes a random sample of K investigators selected from a larger population, and each rates N targets (i.e., case vignettes) altogether, and the two-way mixed effects model, with each target rated by each of the same K investigators, who are the only ones of interest. For both models we calculated the single and average measure reliability.

Analysis of the Distributional Properties of the Treatment Response Scale

For the analysis of the distributional properties, we accessed TS data of 1,308 BD patients from the NIMH centralized ConLiGen phenotypic dataset.

Mixture analysis: frequentist and Bayesian approach. We used mixture analysis to test whether we could identify subgroups of patients according to the degree of response to lithium as expressed by TS. The choice of the mixture model that best fit the distribution of TS was made according to the Akaike's and Schwarz's Bayesian information criteria (AIC and BIC, respectively). The lower values of these two criteria indicated the most parsimonious model that best fit the empirical function of total score distribution. The analysis was performed using the "NMixEM" function implemented in the MixAk package [32] of R software (version 2.13.2).

To verify the findings from the frequentist mixture analysis, we performed the Bayesian mixture analysis employing a minimum message length approach (MML) [33]. Specifically, we used the Snob software [34] to test whether the distribution results from a union of a number of "classes", where the distributions "within-classes" are homogeneous and have a simple form, but vary significantly "between-classes". The best fitting model was indicated as the most parsimonious model (i.e., the one with the lower cost expressed in nits, a specific measure unit conventionally used to express the length message). The analysis was performed using a measurement error equal to 2.5 empirically estimated by plotting the distribution of TS.

Cut off point calculation. Cut off points were derived using the theoretical TS function and calculating each data point's probability of belonging to each class. Specifically, once the mixture model parameters were estimated, we calculated the posterior probability of any data point x belonging to the i-th class as

$$f(x; x \in Class) = \omega i N(x; \mu i, \sigma i)/f(x)$$

where ω is the weight, μ is the mean, σ is the standard deviation.

The resulting probabilities were then compared in order to establish which class the data point belonged to.

Results

Inter-rater Agreement and Reliability of the Assessment of Lithium Response

Raters agreed to a substantial/moderate (first stage of case-vignettes ratings) and moderate/fair (second stage of case-vignettes ratings) degree in assessing lithium response as a dichotomous variable (response/non response) (Table 2). We did not detect an effect of training as shown by the lack of improvement in κ. Specifically, in the first stage of ratings, the κ score showed a substantial level of agreement when we considered the TS cut off for response to lithium at 6 ($\kappa = 0.65$, 95% CI = 0.36–0.85) and at 8 ($\kappa = 0.61$, 95% CI = 0.33–0.83). The highest κ value was for the TS cut off point of 7 ($\kappa = 0.66$, 95% CI = 0.38–0.86). The second stage of ratings had overall lower κ values than the first indicating a moderate level of agreement in the assessment of lithium response (TS = 6: $\kappa = 0.51$, 95% CI = 0.29–0.73; TS = 7: $\kappa = 0.54$, 95% CI = 0.31–0.76; TS = 8: $\kappa = 0.54$, 95% CI = 0.28–0.76). Again, the highest κ value was found for the TS cut off point of 7. Details can be found in Table 2.

We then analyzed the inter-rater reliability for the continuous definition of lithium response. We found that ICC values (two-way random and mixed effects models, single measure) were higher in the first stage of ratings for TS ($ICC_1 = 0.74$ versus $ICC_2 = 0.55$), for A score ($ICC_1 = 0.66$ versus $ICC_2 = 0.52$) and for total B score ($ICC_1 = 0.59$ versus $ICC_2 = 0.34$). However, the training improved the inter-rater reliability of the A score when B score was ≤ 4 ($ICC_1 = 0.71$ versus $ICC_2 = 0.75$). These results are outlined in Table 2.

Table 2. Inter-rater agreement and reliability of the assessment of lithium response in the two-stage case-vignette rating procedure: kappa and intra-class correlation analysis.

Assessment of lithium response	First stage of ratings[§]			Second stage of ratings[~]		
Dichotomous	**κ (95% CI)**	**z**	**p**	**κ (95% CI)**	**z**	**p**
TS cut off of 8	0.61 (0.33–0.83)	103.50	<0.00001	0.54 (0.28–0.76)	68.06	<0.00001
TS cut off of 7	0.66 (0.38–0.86)	112.18	<0.00001	0.54 (0.31–0.76)	68.71	<0.00001
TS cut off of 6	0.65 (0.36–0.85)	110.52	<0.00001	0.51 (0.29–0.73)	64.54	<0.00001
TS cut off of 5	0.58 (0.29–0.81)	99.23	<0.00001	0.48 (0.25–0.71)	61.25	<0.00001
TS cut off of 4	0.51 (0.20–0.78)	86.83	<0.00001	0.42 (0.18–0.67)	52.94	<0.00001
TS cut off of 3	0.40 (0.10–0.73)	68.25	<0.00001	0.37 (0.13–0.66)	47.46	<0.00001

Continuous	ICC_1 single measure (95% CI)*	ICC_1 average measure (95% CI)*	ICC_2 single measure (95% CI)*	ICC_2 average measure (95% CI)*
TS	0.74 (0.59–0.89)	0.99 (0.98–1.00)	0.55 (0.36–0.80)	0.98 (0.96–0.99)
A score	0.66 (0.49–0.85)	0.99 (0.98–1.00)	0.52 (0.33–0.78)	0.98 (0.96–0.99)
Total B score	0.59 (0.41–0.81)	0.99 (0.98–1.00)	0.34 (0.19–0.64)	0.96 (0.92–0.99)
A score if total B score ≤4	0.71 (0.51–0.91)	0.99 (0.99–1.00)	0.75 (0.51–0.96)	0.99 (0.98–1.00)

TS: total score.
ICC: intra-class correlation.
CI: confidence interval.
*Mixed and random effects models.
[§] 70 raters.
[¶] 48 raters.
doi:10.1371/journal.pone.0065636.t002

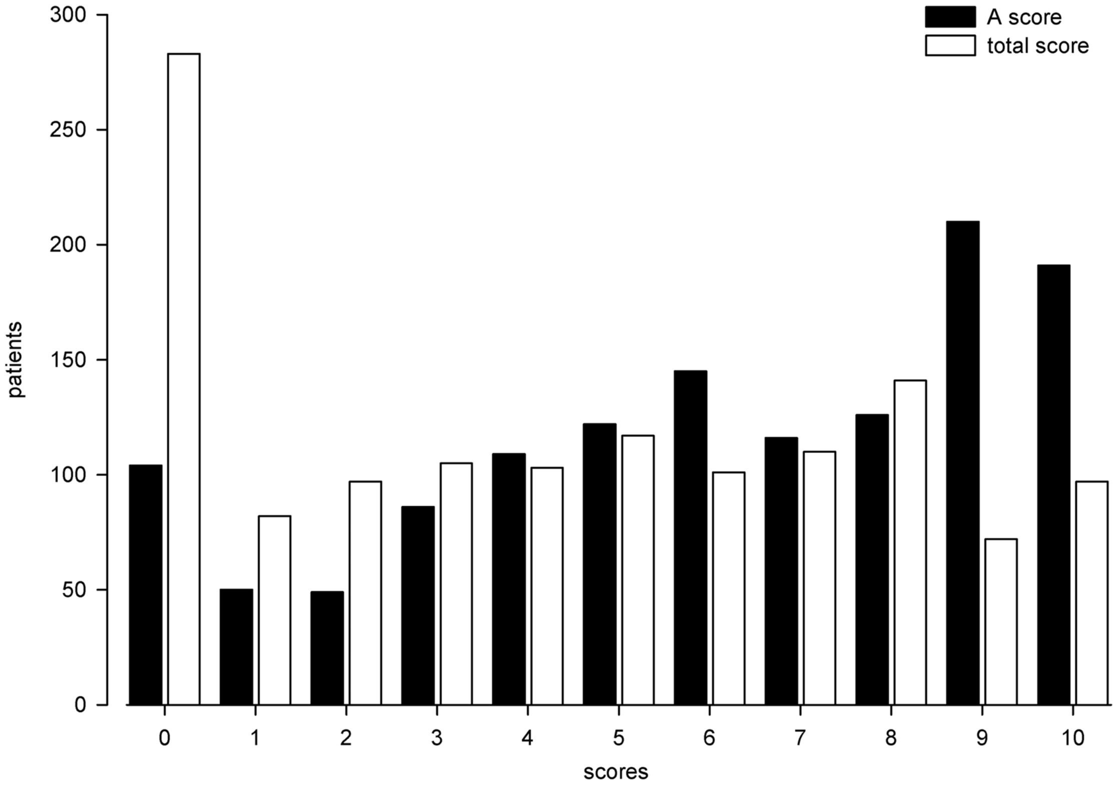

Figure 1. Distribution of total and A scores in the Consortium on Lithium Genetics sample. Histogram plot of the scale scores in 1,308 bipolar disorder patients characterized for response to lithium maintenance treatment.
doi:10.1371/journal.pone.0065636.g001

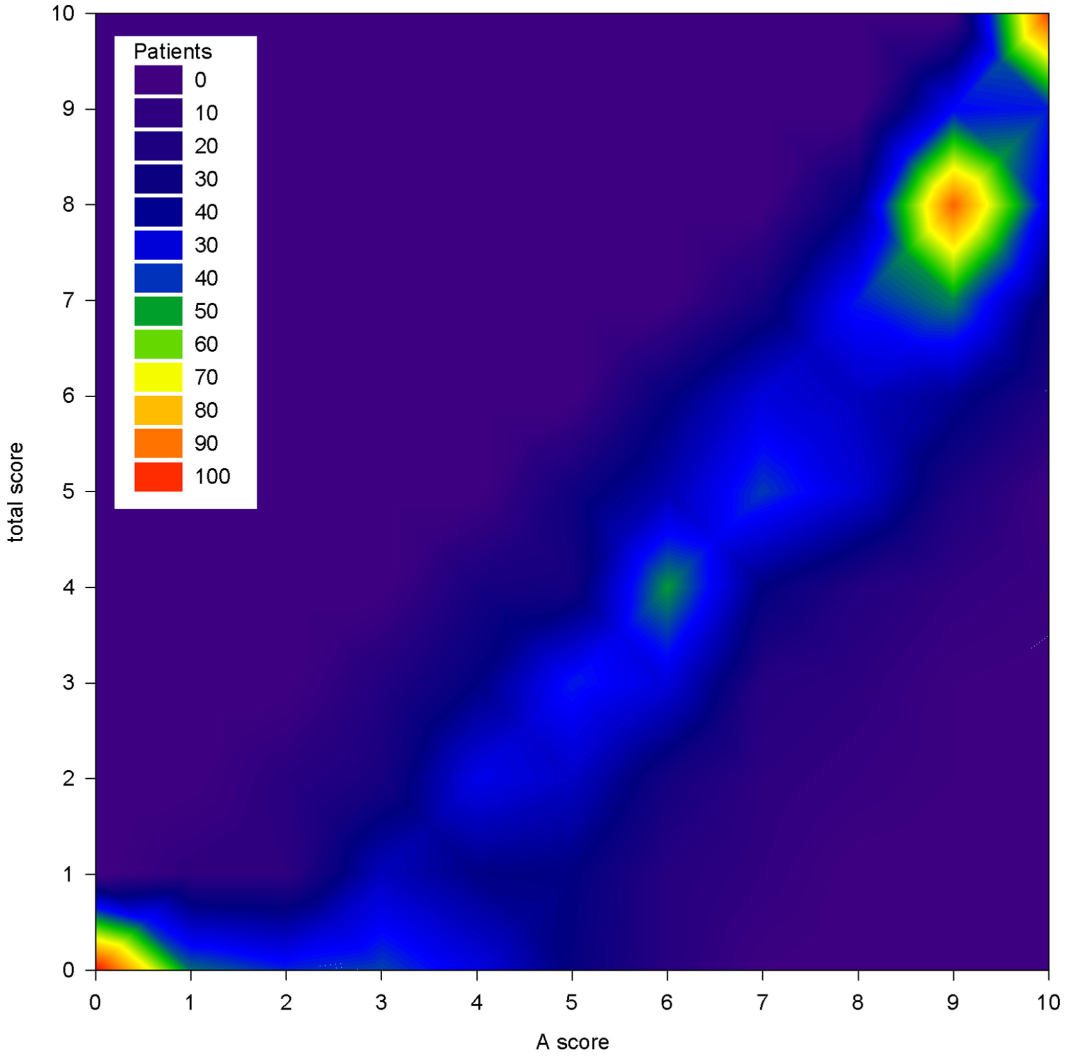

Figure 2. Joint distribution of total and A score in the Consortium on Lithium Genetics sample. Contour plot of the scale scores in 1,308 bipolar disorder patients characterized for response to lithium maintenance treatment.
doi:10.1371/journal.pone.0065636.g002

Analysis of the Distributional Properties of the Treatment Response Scale

Distribution of the TS and joint distribution with score A. Figure 1 illustrates the distribution of TS and A score in 1,308 BD patients characterized for lithium response. Two hundred eighty three patients (21.6%) had TS equal to 0 and 104 patients (8%) had A score equal to 0. In the whole sample the mean A score ± standard deviation] was 6.1±3.1 and the mean TS was 4.4±3.1. The joint distribution of TS and A scores is represented in Figure 2. It illustrates the presence of two frequency peaks at the extreme ends of the scale, namely at 0 and in the area comprised between score A equal to 9 and TS equal to 8–10. A third peak is present at the intersection of A score equal to 6 and TS of 4.

Mixture analysis: frequentist and bayesian approach. The frequentist mixture analysis on TS showed a best-fitting theoretical model of three normal components (AIC = 6467.69, BIC: 6498.75) (Figure 3A). A model with four components did not improve the fit (AIC = 6471.68, BIC = 6513.09, respectively). The mean TS was 0.76±1.15 for the non responder component, 4.6±1.15 for the partial responder component and 8.3±1.15 for the full responder component, with 37%, 30%, and 33% of the population proportion, respectively.

The MML mixture analysis identified the most parsimonious model of three normal components [mean, (SD), (proportion of population)]: 0.5, (1.00), (32%); 4.5, (1.7), (38%); 8.4, (1.2), (30%)], representing the non responder, the partial and the full responder groups of patients. The model is displayed in Figure 3B.

Cut off point calculation. The functions of TS identified with the two different mixture analysis approaches (frequentist and Bayesian) were used to derive the probability of belonging and to calculate the cut off point between the components. The frequentist mixture model suggested two cut off points at TS = 3 and TS = 6.4. Considering the Bayesian MML theoretical function, we obtained two cut off points at 2 and 7. These results confirmed that TS ≥7 is the most appropriate cut off for the definition of full response to lithium prophylaxis as suggested in previous studies [12,13].

Discussion

The purpose of this study was to assess the key phenotypic measures of response to lithium treatment in the large interna-

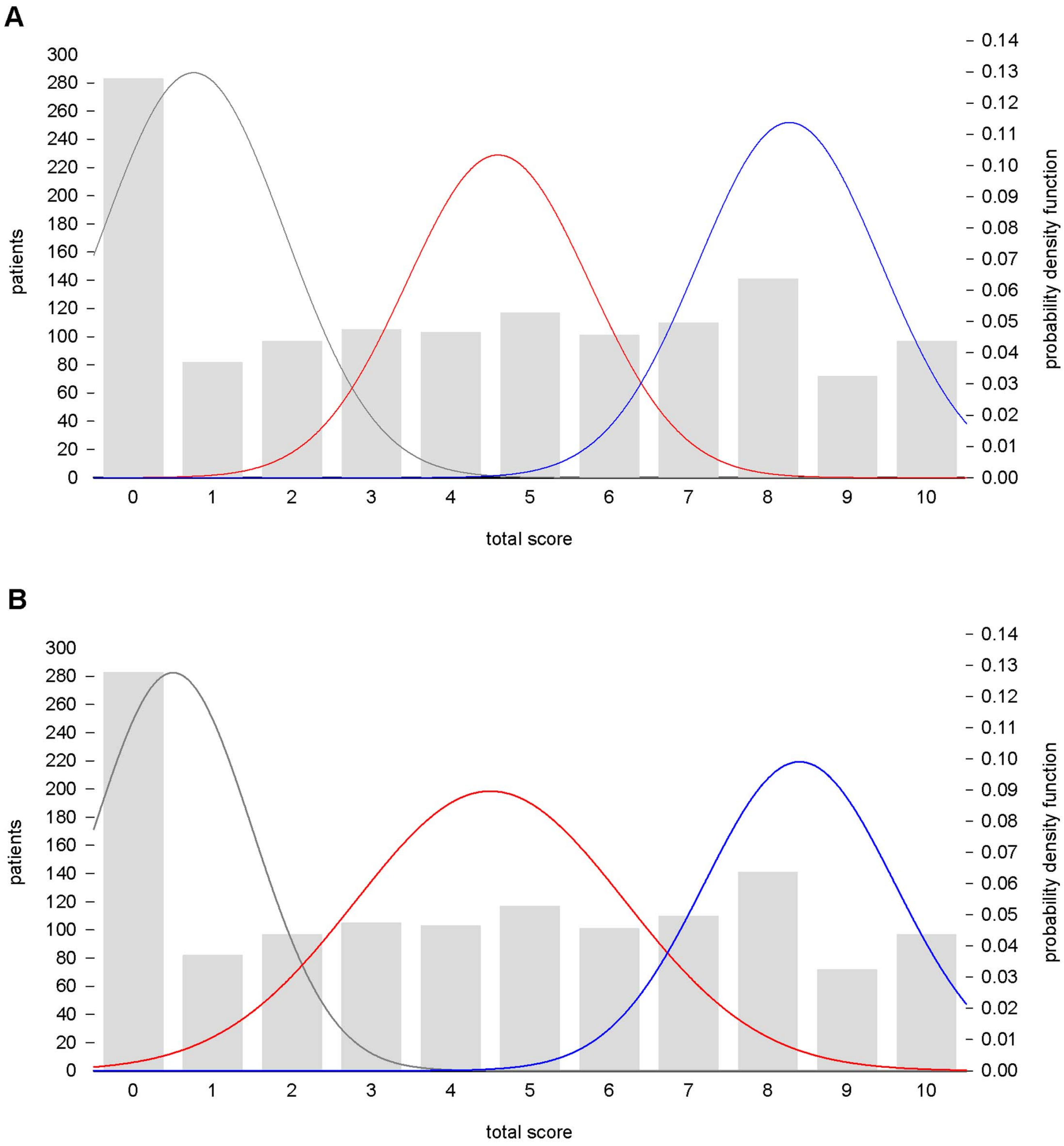

Figure 3. Empirical and theoretical distributions of the total score in the Consortium on Lithium Genetics sample. Frequentist, **A**, and Bayesian minimum message length, **B**, mixture modeling identify three subpopulations of non responders (grey), partial responders (red), and full responders (blue) in total scores of 1,308 bipolar disorder patients characterized for response to lithium maintenance treatment.
doi:10.1371/journal.pone.0065636.g003

tional collaborative Consortium on Lithium Genetics. To this end, two main analyses have been carried out: the inter-rater agreement and reliability of lithium response definition across the ConLiGen participating sites, and the analysis of the distributional properties of the lithium treatment response scale [12]. We found that two definitions of lithium response, one dichotomous and the other continuous had moderate to substantial inter-rater agreement and reliability. Specifically, the two-stage case vignettes inter-rater reliability analysis pointed to the measure of clinical improvement under lithium treatment expressed by the A score and with selection of "valid cases" through a total B score ≤ 4. This phenotypic definition of lithium response had a substantial inter-rater reliability in the first stage of ratings ($ICC_1 = 0.71$) with further improvement in the second stage ($ICC_2 = 0.75$).

Regarding the dichotomous definition of lithium response, a scale TS ≥7 was identified as the best cut off as shown by inter-rater agreement κ scores in the first ($\kappa = 0.66$) and second ($\kappa = 0.54$) stages of case vignette ratings. Further, the analysis of the distributional properties of the treatment response scale further supported this dichotomous definition. In addition, this same measure of lithium response has been previously proposed in several clinical and genetic papers [12,13,35,36].

Some methodological considerations need to be made. For the analysis of the distributional properties, we applied mixture modeling, a method that has been extensively used in psychiatry for the identification of patient subgroups, reducing phenotypic heterogeneity and ultimately helping genetic research [37–39]. It should be noted that this method is exploratory and it does not identify the factors determining the differences between the identified subgroups [40]. A validation of the model can be obtained by comparison of the characteristics of each subgroup. In the ConLiGen study, we plan to use the clinical correlates of lithium response as external validators of the phenotypic measure suggested by the mixture modeling. Such analysis will test and compare the direction and magnitude of the association of a number of clinical variables with lithium response in its dichotomous and continuous definition.

Notably, the analysis of inter-rater reliability and agreement has involved investigators belonging to different research groups with different clinical backgrounds and training. Nevertheless, the use of standardized case vignettes and the training procedures has produced moderate to substantial agreement in the assessment of lithium response. These findings are of importance, given the evidence that even in the context of inpatient unit settings the inter-rater agreement can be unsatisfactory [41].

We performed a two-stage case-vignettes procedure aimed at testing the effect of training on the assessment of lithium response. Contrary to our expectations, we only detected improvement in the inter-rater reliability of lithium response expressed by the A score and with selection of "valid cases" through a total B score ≤4, but not in that expressed by TS or A score. Arguably, the second set of vignettes described more complicated clinical cases with comorbidities, lack of compliance and multiple treatments, all factors that could have influenced the scoring of the B criteria. Indeed, the ICC for the total B score decreased noticeably in the second stage of ratings, implying an increased variability in rating that impacted the discrimination among cases [42]. This explanation is corroborated by the finding of the higher ICC_2 of A score with total B score ≤4. By applying this cut-off we decreased the assessment variability ultimately increasing the discrimination among cases.

Further, these findings confirm that patients with short duration of lithium treatment, poor compliance, and concomitant medications are unlikely to be assessed reliably. This argues against the inclusion of such complex, non-standard cases in pharmacogenomic studies of lithium response. Finally, the higher inter-rater agreement and reliability found in the first set of vignettes suggests that the assessment of lithium response is reliable if sufficient clinical details are available. On the other hand if the information is limited, additional rater training will be of little help.

In conclusion, our findings support the use of two definitions of lithium response for the pharmacogenomic GWAS currently being performed by ConLiGen. Accurate phenotypic definitions of treatment response are crucial in pharmacogenomic studies [43,44]. Heterogeneity in the phenotype definition of treatment response can be a problem especially when in the context of psychiatric disorders. In the absence of other reliable clinical measures of response to lithium, this study has suggested two plausible phenotypic definitions that await application and validation in other samples.

Author Contributions

Conceived and designed the experiments: M. Alda TGS FJM MB MR. Performed the experiments: M. Manchia M. Alda RA JMA LB CEMB BTB FB SB CBP EB CVC ATAC CC S. Clarke PMC CD MDZ JRD BE PF LF MAF JF SG JG FSG PG OG R. Hashimoto JH R. Hoban SJ JPK LK TK JRK SKS SK PHK IK GL CL ML SGL CALJ M. Maj AM LM TM PBM FM PM AN MN TN CO UO NO RHP AP JBP DRE AR ER SR JKR MS PRS OKS BS FS MGS GS LRS CS JWS AS TS PS SKT AT AW DZ MB MR TGS M. Adli. Analyzed the data: M. Manchia M. Alda. Contributed reagents/materials/analysis tools: NA JMB S. Cichon SDD UH LH GAR GT JS NRW PPZ. Wrote the paper: M. Manchia M. Alda.

References

1. Goodwin FK, Jamison KR (2007) Manic-depressive illness. New York: Oxford University Press.
2. Yatham LN, Kennedy SH, O'Donovan C, Parikh S, MacQueen G, et al. (2005) Canadian Network for Mood and Anxiety Treatments (CANMAT) guidelines for the management of patients with bipolar disorder: consensus and controversies. Bipolar Disord 7 Suppl 3: 5–69.
3. Fountoulakis KN, Kasper S, Andreassen O, Blier P, Okasha A, et al. (2012) Efficacy of pharmacotherapy in bipolar disorder: a report by the WPA section on pharmacopsychiatry. Eur Arch Psychiatry Clin Neurosci 262 Suppl 1: 1–48.
4. Muller-Oerlinghausen B, Ahrens B, Grof E, Grof P, Lenz G, et al. (1992) The effect of long-term lithium treatment on the mortality of patients with manic-depressive and schizoaffective illness. Acta Psychiatr Scand 86: 218–222.
5. Cipriani A, Pretty H, Hawton K, Geddes JR (2005) Lithium in the prevention of suicidal behavior and all-cause mortality in patients with mood disorders: a systematic review of randomized trials. Am J Psychiatry 162: 1805–1819.
6. Muller-Oerlinghausen B, Felber W, Berghofer A, Lauterbach E, Ahrens B (2005) The impact of lithium long-term medication on suicidal behavior and mortality of bipolar patients. Arch Suicide Res 9: 307–319.
7. Prien RF, Caffey EM, Klett CJ (1974) Factors associated with treatment success in lithium carbonate prophylaxis: report of the Veterans Administration and National Institute of Mental Health Collaborative Study Group. Arch Gen Psychiatry 31: 189–192.
8. Solomon DA, Keitner GI, Miller IW, Shea MT, Keller MB (1995) Course of illness and maintenance treatments for patients with bipolar disorder. J Clin Psychiatry 56: 5–13.
9. Maj M, Pirozzi R, Magliano L, Bartoli L (1998) Long-term outcome of lithium prophylaxis in bipolar disorder: a 5-year prospective study of 402 patients at a lithium clinic. Am J Psychiatry 155: 30–35.
10. Baldessarini RJ, Tondo L (2000) Does lithium treatment still work? Evidence of stable responses over three decades. Arch Gen Psychiatry 57: 187–190.
11. Rybakowski JK, Chlopocka-Wozniak M, Suwalska A (2001) The prophylactic effect of long-term lithium administration in bipolar patients entering treatment in the 1970s and 1980s. Bipolar Disord 3: 63–67.
12. Grof P, Duffy A, Cavazzoni P, Grof E, Garnham J, et al. (2002) Is response to prophylactic lithium a familial trait? J Clin Psychiatry 63: 942–947.
13. Garnham J, Munro A, Slaney C, Macdougall M, Passmore M, et al. (2007) Prophylactic treatment response in bipolar disorder: results of a naturalistic observation study. J Affect Disord 104: 185–190.
14. Chillotti C, Deiana V, Manchia M, Lampus SF, Ardau R, et al. (2009) [Evaluation of lithium treatment response in Sardinian bipolar patients]. Riv Psichiatr 44: 28–35.
15. Grof P, Alda M, Grof E, Fox D, Cameron P (1993) The challenge of predicting response to stabilising lithium treatment. The importance of patient selection. Br J Psychiatry Suppl 16–19.
16. Kleindienst N, Engel R, Greil W (2005) Which clinical factors predict response to prophylactic lithium? A systematic review for bipolar disorders. Bipolar Disord 7: 404–417.
17. Alda M (2001) Genetic factors and treatment of mood disorders. Bipolar Disord 3: 318–324.
18. Alda M, Grof P, Rouleau GA, Turecki G, Young LT (2005) Investigating responders to lithium prophylaxis as a strategy for mapping susceptibility genes for bipolar disorder. Prog Neuropsychopharmacol Biol Psychiatry 29: 1038–1045.
19. Duffy A, Alda M, Milin R, Grof P (2007) A consecutive series of treated affected offspring of parents with bipolar disorder: is response associated with the clinical profile? Can J Psychiatry 52: 369–376.

20. Cruceanu C, Alda M, Turecki G (2009) Lithium: a key to the genetics of bipolar disorder. Genome Med 1: 79.
21. McCarthy MJ, Leckband SG, Kelsoe JR (2010) Pharmacogenetics of lithium response in bipolar disorder. Pharmacogenomics 11: 1439–1465.
22. Angst J, Sellaro R (2000) Historical perspectives and natural history of bipolar disorder. Biol Psychiatry 48: 445–457.
23. Sachs GS, Rush AJ (2003) Response, remission, and recovery in bipolar disorders: what are the realistic treatment goals? J Clin Psychiatry 64 Suppl 6: 18–22.
24. Turecki G, Grof P, Grof E, D'Souza V, Lebuis L, et al. (2001) Mapping susceptibility genes for bipolar disorder: a pharmacogenetic approach based on excellent response to lithium. Mol Psychiatry 6: 570–578.
25. Lopez de Lara C, Jaitovich-Groisman I, Cruceanu C, Mamdani F, Lebel V, et al. (2010) Implication of synapse-related genes in bipolar disorder by linkage and gene expression analyses. Int J Neuropsychopharmacol 13: 1397–1410.
26. Coppen A, Peet M, Bailey J, Noguera R, Burns BH, et al. (1973) Double-blind and open prospective studies on lithium prophylaxis in affective disorders. Psychiatr Neurol Neurochir 76: 501–510.
27. Maj M, Del Vecchio M, Starace F, Pirozzi R, Kemali D (1984) Prediction of affective psychoses response to lithium prophylaxis: the role of socio-demographic, clinical, psychological and biological variables. Acta Psychiatr Scand 69: 37–44.
28. Schulze TG, Alda M, Adli M, Akula N, Ardau R, et al. (2010) The International Consortium on Lithium Genetics (ConLiGen): an initiative by the NIMH and IGSLI to study the genetic basis of response to lithium treatment. Neuropsychobiology 62: 72–78.
29. Cohen J (1960) A Coefficient of Agreement for Nominal Scales. Educational and Psychological Measurement 20: 37–46.
30. Shrout PE, Fleiss JL (1979) Intraclass correlations: uses in assessing rater reliability. Psychol Bull 86: 420–428.
31. Landis JR, Koch GG (1977) The measurement of observer agreement for categorical data. Biometrics 33: 159–174.
32. Komarek A (2011) Mixture of methods including mixtures. CRAN.
33. Wallace CS, Dowe DL (2000) MML clustering of multi-state, Poisson, von Mises circular and Gaussian distributions. Statistics and Computing 10: 73–83.
34. Wallace CS, Boulton DM (1968) An Information Measure for Classification. Computer Journal 11: 185–194.
35. Squassina A, Manchia M, Congiu D, Severino G, Chillotti C, et al. (2009) The diacylglycerol kinase eta gene and bipolar disorder: a replication study in a Sardinian sample. Mol Psychiatry 14: 350–351.
36. Squassina A, Manchia M, Borg J, Congiu D, Costa M, et al. (2011) Evidence for association of an ACCN1 gene variant with response to lithium treatment in Sardinian patients with bipolar disorder. Pharmacogenomics 12: 1559–1569.
37. Bellivier F, Golmard JL, Rietschel M, Schulze TG, Malafosse A, et al. (2003) Age at onset in bipolar I affective disorder: further evidence for three subgroups. Am J Psychiatry 160: 999–1001.
38. Etain B, Mathieu F, Rietschel M, Maier W, Albus M, et al. (2006) Genome-wide scan for genes involved in bipolar affective disorder in 70 European families ascertained through a bipolar type I early-onset proband: supportive evidence for linkage at 3p14. Mol Psychiatry 11: 685–694.
39. Hamshere ML, Gordon-Smith K, Forty L, Jones L, Caesar S, et al. (2009) Age-at-onset in bipolar-I disorder: mixture analysis of 1369 cases identifies three distinct clinical sub-groups. J Affect Disord 116: 23–29.
40. Lubke G, Neale M (2008) Distinguishing between latent classes and continuous factors with categorical outcomes: Class invariance of parameters of factor mixture models. Multivariate Behav Res 43: 592–620.
41. Cheniaux E, Landeira-Fernandez J, Versiani M (2009) The diagnoses of schizophrenia, schizoaffective disorder, bipolar disorder and unipolar depression: interrater reliability and congruence between DSM-IV and ICD-10. Psychopathology 42: 293–298.
42. Burdock EI, Fleiss JL, Hardesty AS (1963) A new view of inter-observer agreement. Personnel Psychology 16: 373–384.
43. Daly AK (2010) Genome-wide association studies in pharmacogenomics. Nat Rev Genet 11: 241–246.
44. Motsinger-Reif AA, Jorgenson E, Relling MV, Kroetz DL, Weinshilboum R, et al. (2010) Genome-wide association studies in pharmacogenomics: successes and lessons. Pharmacogenet Genomics.