April 6, 2011

Editors
Psychological Medicine
Kenneth S. Kendler, MD
Robin M. Murray, MD

Dear Drs. Kendler and Murray,

I am writing to request that you investigate the Nierenberg et al. article, "*Residual symptoms after remission of major depressive disorder with citalopram and risk of relapse: a STAR*D report,*" and consider retracting it from your journal. I believe that after reviewing this letter, and the articles attached, you will find that the authors' submission and the scientific standards that were used in this article, and many of the STAR*D publications, do not meet accepted standards for quality research and in many cases are in fact inaccurate and have led to unsubstantiated, and potentially harmful, clinical conclusions.

The Nierenberg et al. article uses the clinic-version of the QIDS-SR as the sole measure to report residual symptoms, yet this version is specifically **EXCLUDED** from use as a research measure in STAR*D's Research Protocol since it was used to guide care in this study. This fact is seen in the opening paragraph of the Protocol's **Research Outcomes Assessments** section that states, "*Recall the research outcomes assessments are distinguished from assessments conducted at clinic visits* (such as the QIDS-SR). *The latter are designed to collect information that guides clinicians in the implementation of the treatment protocol. Research outcomes assessments are not collected at the clinic visits. They are **not** collected by either clinicians or CRCs* " (note: the underlined emphasis is part of the Research Protocol's text) [NIMH 2002, pp. 47–48].

As you know, STAR*D is the largest antidepressant trial ever conducted and was overseen by many of America's leading depression researchers. Given the 100+ published peer-reviewed articles by the study's investigators, innumerable citations of STAR*D's findings from other researchers, and wide coverage in the media, STAR*D's published reports have had a significant impact influencing the treatment of major depression; much of it likely wrong though since it is based on the improper and unscientific reporting of STAR*D's results.

I discovered indicators of apparent researcher bias when STAR*D's step-1 results were first published in January 2006. This began a 5+ year effort on my part deconstructing STAR*D by comparing its published methods and findings with STAR*D's pre-specified research measures and analytic plan as described in primary source documents (e.g., the NIMH-approved STAR*D Research Protocol, STAR*D Clinical Procedures Manual, STAR*D Patient Education Manual, and STAR*D's 2004 Controlled Clinical Trials article).

Attached are two articles that I have published from these efforts documenting significant researcher error and unscientific processes in the analysis and reporting of STAR*D's results. Among other forms of apparent bias, these articles document:

- Each of STAR*D's primary source documents clearly distinguished between the assessments obtained at the clinic visits that were used to guide clinicians and the blindly-administered research measures that were to be used to report outcomes. For example, in addition to the Research Protocol cited in the second paragraph above, STAR*D's Controlled Clinical Trials article states, "*At all clinic visits, information is obtained to guide clinicians implementing study treatments (Table 3). Symptomatic status is measured by the QIDS-C_{16} ...The self-rated form of the QIDS (QIDS-SR_{16}) is also administered at all visits*" [p.127]. The next paragraph states, "*Research outcomes (Table 4) are collected by IVR and by telephone interviews conducted by Research Outcomes Assessors (ROAs), independent of and masked to treatment*" [p. 128].
- This same clear distinction is also made in the Background FAQ's NIMH posted regarding STAR*D. FAQ # 5 states "*At each participant visit, STAR*D investigators measured symptoms and side effects to determine when and how much to increase medication doses or change to other treatments*"...and then "*To ensure that there would be **no bias** in assessing how well each treatment worked, the information that was used for measuring the outcome results of the study was collected both by an expert clinician over the phone who had **no knowledge** of what treatment the participants were receiving and by a novel computer-based interactive voice response system*" [NIMH, 2006].
- Despite this clear differentiation between the clinic assessments that were used to guide care and the blinded research outcome measures, in STAR*D's steps 1-4 articles, the clinic-version of the QIDS-SR was used as the secondary measure to report remission rates, and sole measure to report response rates, even though it was not a research measure.
- STAR*D's authors engaged in factual distortions to justify their use of the QIDS-SR as the **sole** measure to report remission and response rates in the summary article. These distortions included falsely asserting that "*the QIDS-SR was not used to make treatment decisions*" [Rush et al., 1908] despite the fact that this assertion is contradicted by the authors themselves when they write in the step-1 article, "*To enhance the quality and consistency of care, physicians used the clinical decision support system that relied on the measurement of symptoms (QIDS-C and QIDS-SR), side effects (ratings of frequency, intensity, and burden), medication adherence (self-report), and clinical judgment based on patient progress'* [Trivedi et al., 2006, p. 30] as well as being contradicted in the primary source documents (e.g., see both the Controlled Clinical Trials citation and the Research Protocol citation cited above).
- To my knowledge, STAR*D's authors have never disclosed in any of their 100+ published articles that the clinic-version of the QIDS-SR was not a research measure.
- In a conference call arranged by NIMH, Stephen Wisniewski, STAR*D's chief biostatistician, acknowledged that the QIDS-SR was not a pre-specified research measure and then when asked by me why this fact was not disclosed in STAR*D's steps 1–4 and summary articles, he responded "***there was not enough journal space***."
- Despite having published over a 100 peer-reviewed articles, STAR*D has to my knowledge never reported the outcomes from any of its 12 pre-specified research measures other than the remission rates for the Hamilton in the initial steps 1-4 articles. Instead, STAR*D's authors repeatedly use the clinic-version of the QIDS-SR to report outcomes as they did in the Nierenberg et al. paper.

- STAR*D's authors failed to disclose that all 4,041 patients were started on citalopram/Celexa in their baseline visit and that after up to four treatment trials only 108 patients (2.7%) had a remission and did not relapse and/or drop out during the 12 months of follow-up care. Moreover, it is not known how many of these 108 survivors were one of the 607 patients who, due to a change in eligibility criteria, were allowed into the study despite having a baseline Hamilton score <14 signifying at most only mild symptoms when first started on citalopram/Celexa and therefore who had to score worse during follow-up than when they first entered the study to be counted as relapsed. Nor is it known how many of the 108 patients actually remained "*in remission*" during follow-up care.

As detailed in both of the attached articles, it is critical for journal readers to understand the procedures for administering the clinic-version of the QIDS-SR which invalidated its use as a research measure.

First, STAR*D patients completed a pencil-and-paper version of the QIDS-SR at the beginning of each clinic visit that was overseen by **non-blinded clinical research coordinators** (CRCs). The Clinical Procedures Manual instructed the CRCs to then review the QIDS-SR results to make certain that all items were completed and then see the patient to administer "*the appropriate Patient Education material*" [Trivedi et al., p. 75].

The CRC then administered a multistep educational program for patients and families that was based on the neurochemical imbalance theory of depression and included "*a glossy visual representation of the brain and neurotransmitters*," consistently emphasized that "*depression is a disease, like diabetes or high blood pressure, and has not been caused by something the patient has or has not done. (Depression is an illness, not a personal weakness or character flaw.) The* [CRC] *educator should emphasize that depression can be treated as effectively as other illnesses*," and "*explaining the basic principles of mechanism of action*" of citalopram/Celexa to treat their depression [O'Neal & Biggs, 2001, pp. 4–7].

Next, the CRCs administered the QIDS-C, the clinician-administered version of the QIDS with the identical 16 questions and response options as the QIDS-SR, and were instructed to discuss "*any symptoms and side effects that the patient may be experiencing*" [Trivedi et al., p. 75]. CRCs then recorded both the QIDS-SR and QIDS-C information—*as well as information from four other measures*—on the clinical record form for the treating physician's review before he/she saw the patient. This was done *'to provide consistent information to the clinicians who use this information in the protocol'* [Rush et al., 2004, p. 128].

In light of the above, it is clear why as pre-specified the clinic-version of the QIDS-SR was explicitly excluded from use as a research measure. First, there were significant demand bias effects given the conditions under which it was administered thereby biasing any reporting of QIDS-SR outcomes in STAR*D's open-label study. Second, it is simply inappropriate from a scientific perspective to use a non-blinded self-report measure such as the QIDS-SR to both guide care in every visit as well as to evaluate the outcomes from said care in an open-label trial (or any trial for that matter). Furthermore, the fact that the QIDS

were administered twice in every clinic visit only makes use of it as a research measure doubly absurd in terms of it holding any scientific merit other than documenting the demand bias effects that occur under these circumstances.

Regarding this latter point, in the Nierenberg et al. paper it would have been informative to compare the residual symptoms in depressed patients who remitted as determined by the blindly-administered Hamilton to their self-reports on the clinic-version of the QIDS given their repeated experience of having the CRC subsequently discussing with them "*any symptoms and side effects*" that they reported on this measure. Such knowledge would be of vital importance to clinicians since it would indicate if there are certain depressive symptom domains where such patients underreport their residual symptoms in non-blinded self-reports that they know will be subsequently discussed with their clinician (e.g., suicidality) versus what they disclose in an expert interview in which the interviewer is "*independent of and masked to treatment*" such as occurred with the Hamilton. So while Nierenberg et al. report a rate of only 1.3% of remitters reporting at least mild suicidal ideation based on the QIDS-SR data, this exceptionally low rate may provide false comfort to clinicians reading their paper thereby harming patient care if in fact the actual rate of suicidal ideation is significantly higher when using a more thorough and unbiased assessment such as the gold-standard blindly-administered Hamilton.

The analysis suggested above is just one example of how STAR*D's dataset could be used to advance patient care whereas the current paper may in fact harm patient care by providing false comfort to clinicians. This potential false comfort regarding suicidality resulting from the demand bias effects of how the QID-SR was administered is seen in Nierenberg et al's discussion section where they state, "***Those with the most severe suicidal ideation at baseline, however, had robust improvements, with complete resolution of suicidal ideation*** *and, of the less severe groups, only one participant had a slight worsening*. ***Thus, suicidal ideation was highly responsive to treatment in remitters. Only a very small minority had either persistent or treatment-emergent suicidal ideation***" [pp. 46-47].

The statements above by Nierenberg et al. are also counter to STAR*D's two earlier articles that 120 out of 1,915 patients (6.3%) reported emerging suicidal ideation while taking citalopram/Celexa based on a repeated measures analysis of the QIDS-SR (Laje, Paddock, Manji, **Rush**, et al., 2007) and 124 out of 1,447 patients (8.6%) reported emergent suicidal ideation while taking citalopram/Celexa based on a similar analysis of the QIDS-C (Perlis, Purcell, **Fava**, Fagerness, **Rush**, **Trivedi**, et al., 2007). Being coauthors, Drs. Fava, Rush, and Trivedi clearly knew about the findings from these two studies of emerging suicidal ideation during citalopram/Celexa treatment in STAR*D and it is deeply troubling that this countervailing information was not discussed in the McClintock et al. paper. At a minimum, I would think that if discovered during peer-review, both you and your peer-reviewers would have insisted that the authors address this inconsistency.

Furthermore, while Nierenberg et al. report that "***Only a very small minority had either persistent or treatment-emergent suicidal ideation***" this may in fact be due simply to the demand bias effects of how the QIDS-SR was administered such that some suicidal patients ceased endorsing suicidal ideation because they no longer wanted to discuss this symptom

with the CRC while they were more willing to acknowledge other less evocative symptoms. Simply put, substandard science as evidenced in the Nierenberg et al. paper, and in many of the STAR*D articles, does not only fail to advance the field of depression and its treatment, it may also in fact directly harm patient care.

The Nierenberg et al. paper also includes Drs. Rush, Trivedi, Wisniewski, Fava, and Warden as coauthors, each of whom have been lead authors on one or more STAR*D reports. Furthermore, Dr. Rush was STAR*D's principal investigator.

Despite this impressive listing of coauthors who clearly knew otherwise, the Nierenberg et al. paper adds a new factual distortion to STAR*D's long trail of misrepresentations and repeated inaccuracies regarding their use of the QIDS-SR. This new misrepresentation is their statement that "*The QIDS-SR$_{16}$ was completed by participants at baseline and at every visit to assess depressive symptoms. The self-report FIBSER was completed by participants after every visit to assess side-effects*. ***Both measures were gathered within 72 h of each visit using a telephone-based interactive voice response (IVR) system***." (emphasis added) [p. 43].

This statement in their "**Measures**" section on Drs. Nierenberg, Husain, Trivedi, Fava, Warden, Wisniewski, Miyahara, and Rush's part is not true since there was no "*telephone-based interactive voice response (IVR) system"* version of the QIDS-SR$_{16}$ that was administered *"within 72 h of each visit."* It simply didn't happen. This is fiction; there was no telephonic IVR version of the QIDS-SR that was administered within 72 hours of ***each visit***.

Drs. Nierenberg et al's most recent factual distortion is easy to validate. I have attached STAR*D's Controlled Clinical Trials article. On page 128, please see table 3 "***Data collection at clinical visits***[a]" and note that the QIDS-SR was completed by the patient **at each clinic visit** and as stated in the footnote "[a]*These measures are used to provide consistent information to the clinicians who use this information in the protocol and are recorded on the CRF. These measures are collected* ***at clinic visits*** *for participants in protocol treatment.*"

Now please see table 4 "***Research outcomes***[a]" on page 129 and note that the QIDS-SR "*Telephone (IVR)*" version was administered on the same schedule as the other research outcome measures as stated in the footnote "[a] *Research outcomes are obtained at entry and exit from each treatment level and in follow-up at months 3, 6, 9, and 12.*" The IVR-version of the QIDS-SR was also administered at week 6 during acute-care as stated two paragraphs below table 4.

Please note that there was no administration of the QIDS-SR's IVR-version "*within 72 h of* ***each visit***." It did not happen. This was not part of STAR*D's methodology as confirmed in the NIMH-approved Research Protocol, Clinical Procedures Manual, and the 2004 Controlled Clinical Trials article that was published well after the study had started. In fact, the Controlled Clinical Trials article states that "*As of June 1, 2003, 2555 subjects had been enrolled into level 1*" [p. 135] indicating that over half of STAR*D's patients had started

citalopram/Celexa treatment adhering to the protocol described in this article; so there was certainly no change to the protocol in which the QIDS-SR was "*gathered within 72 h of each visit using a telephone-based interactive voice response (IVR) system.*" The same is also true for the FIBSER. **Nierenberg et al's description of how both measures were administered is pure fiction on their part**.

While not knowing the authors intent, the effect of Nierenberg et al's fiction was to foster in readers' minds the validity of their use of the QIDS-SR as a research measure to report the residual symptoms' in this article. This observation is supported by the manner in which the authors' misrepresent the QIDS-SR's administration by:

- Disassociating the QIDS-SR from its use to guide care and instead, associate it with the QIDS-SR's IVR-version that was a valid pre-specified research outcome measure; and
- Failing to report in the **Protocol for acute treatment** section that the QIDS-SR was administered at the beginning of every clinic visit as described above and was one of several measures that were used to guide care in this study. Instead, the authors only report the QIDS-C administration.

Additional support that a false legitimacy is being conveyed to your readers that the QIDS-SR was a valid research measure is seen in the "**Definition of residual symptoms**" section where the authors' state, "*Because the most complete data available with the least missing data points were gathered using the QIDS-SR$_{16}$, the presence of individual or domain residual symptoms was categorized using the QIDS-SR$_{16}$.*"

Nierenberg et al. had available the blindly-administered Hamilton that they could have used to report to your readers the rates of remitted patients' residual symptoms using this pre-specified primary research outcome measure. Instead though, they used a self-report measure riddled with potential demand bias effects and then interpreted their "*findings*' in a way that may harm patient care by giving false comfort to clinicians regarding the likelihood of persistent suicidal ideation in remitted patients.

Another indicator of apparent researcher bias in the Nierenberg et al. article is their report of a 32% remission rate for step-1's citalopram/Celexa treated patients. As Pigott et al. document, these patients' actual rate was only 25.4% when using STAR*D's pre-specified primary outcome measure and analytic plan [Pigott et al. 2010, p. 274].

Furthermore, while Nierenberg et al's analysis found that, "*Participants who reached remission within the first 6 weeks had fewer residual symptoms compared to those who reached remission after 6 weeks (Fig. 1)*" [p. 44], this findings is nothing more than an artifact of their decision to use the sham QIDS-SR to report residual symptoms versus the pre-specified Hamilton. This is easy to verify by simply comparing step-1's table 2 with figure 3 [Trivedi et al., 2006, p. 32 & 33]. In table 2, 93% of the 790 patients who had a Hamilton-defined remission, their exit Hamilton assessment confirming their remission occurred after 8 or more weeks of treatment. In contrast, figure 3 indicates that approximately 50% of the 943 QID-SR defined remissions occurred by the 6th week of treatment. This difference is due to the fact that the vast majority of the additional 153

'*remissions*' that were identified by using the sham QIDS-SR were patients who dropped out of treatment at the week 2, 4, or 6 visit whereas those staying longer in treatment were less likely to dropout and more likely to take the exit Hamilton. Demand bias alone among these early dropouts likely accounts this difference in residual symptoms.

Your journal is not the only one in which STAR*D's lead investigators have made this new fictious claim that the QIDS-SR was administered "*within 72 hours of each clinic visit*" [see attached McClintock et al., 2011, p. 181]. I have written a letter to the Journal of Clinical Psychopharmacology's editors requesting retraction for this article as well.

The fact is that we would be far further ahead today in improving the outcomes for patients suffering with major depression if STAR*D's authors had reported their findings as pre-specified in 2006. Such honesty would have increased the urgency to research alternative models of psychopharmacological and non-pharmacological care for such patients. For example, investigating Dr. Ghaemi's hypothesis that major depression should be viewed as more analogous to an infectious disease with antidepressants prescribed for shorter durations similar to antibiotics versus the diabetes/insullin model advocated by STAR*D [Ghaemi, 2008, p. 965].

For major depression treatment to advance, STAR*D's authors' continuing pattern of substandard science, factual distortions, and lack of forthrightness to journal readers has to end. If not now under your co-editorship, then when?

This letter and supporting materials document significant misrepresentations and substandard science in the Nierenberg et al. paper. If these are shown to be true when investigated by you, it clearly warrants this article's retraction. The authors' misrepresentations warranting investigation includes:

- Not disclosing that the measure used to report residual symptoms in this article was explicitly excluded from use as a research outcome measure in STAR*D; and
- Falsely describing how the QIDS-SR was administered in a way that misleads journal readers into believing that it was a valid measure to report residual symptoms in this article.

Thank you for your prompt attention to this important matter.

Sincerely,

H. Edmund Pigott, Ph.D.

References:

Ghaemi SN. *Why antidepressants are not antidepressants: STEP-BD, STAR*D, and the return of neurotic depression*. Bipolar Disorders 2008: 10: 957–968.

Laje, G., Paddock. S., Manji, H., Rush, A.J., et al. *Genetic markers of suicidal ideation emerging during citalopram treatment of major depression*. American Journal of Psychiatry 2007; 164:1530–1538.

National Institute of Mental Health. *Sequenced Treatment Alternatives to Relieve Depression (STAR*D) Research Protocol* (Rev. ed.). Retrieved October 1, 2010, from Freedom of Information Act.

National Institute of Mental Health. (January 2006). *Questions and answers about the NIMH Sequenced Treatment Alternatives to Relieve Depression (STAR*D) study—Background*. @: http://www.nimh.nih.gov/trials/practical/stard/backgroundstudy.shtml

Nierenberg et al., *Residual symptoms after remission of major depressive disorder with citalopram and risk of relapse: a STAR*D report*. Psychological Medicine 2010; 40: 41–50.

O'Neal, B., & Biggs, M. (2001). *STAR*D patient education manual.* Retrieved from http://www.edc.pitt.edu/stard/public/study_manuals.html.

Perlis, R. H., Purcell, S., Fava, M., Fagerness, J., Rush, A. J., Trivedi, M. H., et al. (2007). *Association between treatment-emergent suicidal ideation with citalopram and polymorphisms near cyclic adenosine monophosphate response element binding protein in the STAR*D study*. Archives of General Psychiatry 2007; 64(6): 689–697.

Rush et al., *Sequenced Treatment Alternatives to Relieve Depression (STAR * D): rationale and design*. Controlled Clinical Trials 2004; 25: 119–142.

Rush A. J., Trivedi, M. H., Wisniewski S. R., Nierenberg A. A., Stewart J. W., Warden D., et al. (2006). *Acute and longer-term outcomes in depressed outpatients requiring one or several treatment steps: a STAR*D report*. The American Journal of Psychiatry, 163(11), 1905–1917.

Trivedi et al. *STAR * D clinical procedures manual*. July 31, 2002. Retrieved from www.edc.pitt.edu/stard/public/study_manuals.html.

Trivedi et al. Evaluation of outcomes with citalopram for depression using measurement-based care in STAR * D: implications for clinical practice. American Journal of Psychiatry 2006; 163: 28–40.