European Neuropsychopharmacology (2013) 23, 1597–1599

www.elsevier.com/locate/euroneuro

SHORT COMMUNICATION

Limited utility of number needed to treat and the polarity index for bipolar disorder to characterize treatment response

Larry Alphs[a,*], Joris Berwaerts[b,1], Ibrahim Turkoz[b,2]

[a]Janssen Scientific Affairs, LLC, 1125 Trenton-Harbourton Road, Titusville, NJ 08560-2000, USA
[b]Janssen Research & Development, LLC, 1125 Trenton-Harbourton Road, Titusville, NJ 08560-2000, USA

Received 3 December 2012; received in revised form 28 December 2012; accepted 30 December 2012

KEYWORDS
Polarity index;
Number needed to treat;
Bipolar disease;
Limitation

Abstract
The medical community increasingly supports the use of simplifying constructs or ratios to facilitate incorporation of evidence-based medicine into clinical practice such as number needed to treat (NNT) and polarity index (PI). Clinicians and teachers find them to be an appealing, easy-to remember integer that can be readily translated into clinical practice. However, serious questions have been raised with respect to the validity, reliability and value of these descriptors of response. This commentary identifies some of the specific limitations of the NNT and PI constructs when applied to treatments of bipolar disorder.

1. Introduction

The medical community increasingly supports the use of simplifying constructs or ratios to facilitate incorporation of evidence-based medicine into clinical practice. Clinicians and teachers find them to be an appealing, easy-to-remember integer that can be readily translated into clinical practice. Among these constructs are: the number needed to treat (NNT) (Citrome, 2012, 2010; Popovic et al., 2011), number needed to harm (NNH), and the likelihood to be helped or harmed (the ratio of NNH to NNT) (Citrome and Katrowitz, 2008; Citrome, 2012). Recently, Popovic et al. (2012) have proposed the polarity index (PI) as another descriptor for categorizing profiles of drugs used for maintenance treatment of bipolar disorder. This PI is derived by dividing the NNT for the prevention of depressive episodes by the NNT for the prevention of manic episodes (Popovic et al., 2012).

Despite their apparent utility, serious questions have been raised with respect to the validity, reliability and value of these descriptors of response (Hutton, 2000; Thabane, 2003; Stang et al., 2010; Suissa et al., 2012). We argue that particularly in complex diseases such as bipolar disorder, which have a heterogeneous presentation, natural history, and response to treatment (e.g., multiple

*Corresponding author. Tel.: +1 609 730 3693.
E-mail addresses: lalphs@its.jnj.com (L. Alphs), jberwaer@its.jnj.com (J. Berwaerts), iturkoz@its.jnj.com (I. Turkoz).
[1]Tel.: +1 609 730 3540.
[2]Tel.: +1 609 730 7719.

0924-977X
http://dx.doi.org/10.1016/j.euroneuro.2012.12.006

potential outcomes, large variability in response, and numerous competing risks for recurrence of mood episodes during maintenance treatment), the limitations of these constructs, especially those expressed as ratios and ratios of ratios, are exaggerated and severely restrict or nullify their clinical utility. Indeed, it is unlikely that any single measure can ever summarize the full spectrum of therapeutic responses to maintenance treatment for the full population. This commentary identifies some of the specific limitations of the NNT and PI constructs when applied to treatments of bipolar disorder.

Inherent in these constructs is a lack of clarity as to their true statistical properties, especially when presented without confidence intervals (CIs). For example, they provide no information on the normality or skewness of the population under consideration (Hutton, 2000; Lesaffre and Pledger, 1999). Further, without knowledge that treatment response is stable within key population subgroups (like those for age, sex, race, co-morbidities, etc.) the use of these constructs in clinical decision-making or meta-analysis is unwise or even dangerous (Smeeth et al., 1999) as their use may drive decisions contrary to the observed effects in these population subgroups.

A specific example that highlights these challenges is the PI. It has been developed to summarize, in a single, easily remembered number, both the relative antimanic and antidepressant preventive potentials of pharmacologic treatments, and facilitate clinical decision-making regarding selection of maintenance treatment. However, its derivation is based on numerous unsupported assumptions. Among these are that: (1) there is a single, well-accepted definition of relapse that has been consistently applied to the studies from which this number is derived; (2) the risk of relapse is consistent for all members of that population over the course of the disorder; and (3) if a PI is being generated from a meta-analysis of different studies, the true mean for those studies is similar and is derived from similar populations such that data collected in separate studies can be validly pooled.

Even assuming a fixed definition of relapse across studies; the baseline risk of relapse almost certainly varies across subpopulations and at different points within the course of the disease. For instance, one cannot assume that the baseline risk for relapse for adolescents with bipolar disorder is the same as that for elderly persons with a long history of the disease. Nor is the baseline risk for relapse in persons who have less than one month of recovery from a depressive relapse likely to be identical to that for persons who have not had any relapses in the past year. This argues against combining results across trials to generate a meta-analytically-derived PI.

To illustrate and clarify this scenario, let us assume Treatment A for bipolar disorder provides a *constant* 50% risk reduction for relapse across the disease spectrum compared to placebo. If we know that the annualized risk of relapse in an adolescent with a recent relapse treated with placebo is 60%, then the risk for relapse in this person receiving Treatment A is 30%, with an absolute risk reduction of 30% and an NNT of 1/0.3=3.33. However, if the annualized risk of relapse in an elderly person with no history of relapse in the past two years treated with placebo is 20%, then the risk for relapse in this person receiving Treatment A is 10%, with an absolute risk reduction of 10% and an NNT of 1/0.1=10. What, then, is the NNT for Treatment A? Can we combine the results from two such different groups of patients to give an NNT for the entire population? This is complicated even further by emerging evidence that the effect of psychotropic medication on biomarkers reflective of the course of the underlying psychiatric disorder is not constant over the disease spectrum (Bartzokis, 2012). For instance, Treatment A may be more effective early in the course of bipolar disorder than later in the disease course. These considerations raise the requirement for including a CI with each NNT as discussed below. For this example a CI would be required for the constant risk reduction, and the risks for relapse in each subpopulation.

We argue that existing data suggest that the NNT for relapse into mania and relapse into depression differ according to factors such as the polarity of the index episode, number of previous episodes, time since last episode, baseline severity of mood symptoms, and other factors. When identified from clinical trials, these summary characteristics may be further influenced by dropout rates, treatment adherence characteristics, and duration of prospective follow-up (e.g., values may be different after one month of follow up versus one year of follow up). We contend that all of these factors can impact response to treatment and, consequently, could affect NNT-derived constructs. To identify a number that adequately represents the entire population, all risk factors affecting relapse must be known or predicted with significant confidence. We argue that it is impossible to address all of these factors so as to construct a single valid PI for the entire population. Even if this could be achieved, it would be impossible for a clinician to unravel this information when attempting to use the PI to choose between Treatment A and Treatment B for treatment of a substance-abusing 30-year-old bipolar patient with a history of manic relapse six months ago.

Beyond these difficulties, the quantitative/statistical shortcomings of the PI are substantial. Among these are:

1.1. Problems with precision

As noted above, it cannot be assumed that everyone in the population will respond identically at the point estimate identified by the constituent NNTs. Presentation of a PI (composed of two NNTs) as a point estimate without accompanying estimates of precision (e.g., a 95% CI) represents an incomplete description of the therapeutic effect and fails to provide the range of likely values for the population. It also fails to rule out values that are outside of these plausible values. It does not account for the greater certainty in range of possible values that may be established in a large, well-conducted study versus the lesser certainty of values derived from smaller studies. Additionally, as a PI would be developed from separate studies the CI for the numerator may differ substantially from that of the denominator.

1.2. Problems when the CI estimate includes zero

A non-significant NNT will have a CI with two parts: one that describes risk for harm, and the other, potential for benefit.

If the CI for absolute risk reduction includes zero, the corresponding CI for the NNT includes infinity. It is unclear how both risk for harm and benefit can be managed in a single PI computation when either of the CIs includes zero. How informative is a PI of 5, if the NNT for prevention of depressive episodes is 10 (95% CI: −7 to 17) and the NNT for prevention of manic episodes is 2 (95% CI: −15 to 10)? This issue is further highlighted by situations where the treatment effect on the risk for relapse into mania is opposite the effect on depression. A PI so generated would not be informative about which effect was inferior.

1.3. Problems with competing risks

The endpoint for the PI consists of two (not including a separate category for mixed episodes) distinct events of interest (relapse into mania and relapse into depression). The eventual treatment failure is attributed to one event exclusive of the other. This raises a problem of 'competing risks.' Depressive and manic NNTs in clinical trial settings are not independent. If a failure is attributed to one event (e.g., depression), the chance for relapse into the other event type (i.e., mania) is excluded. This is a typical example of dependent censoring, and represents a version of ascertainment bias in epidemiology.

1.4. Problems with using a simple proportion

The PI construct represents ratios of simple proportions that capture event rates at the end of a defined period. However, it does not adjust for dropouts or censoring mechanisms that are likely to have occurred during the period of observation.

Historically, constructs such as the NNT, NNH, and their ratios have been advocated for ascribing risks and benefits to treatments in acute disorders such as infection that involve a homogeneous population with a circumscribed outcome of interest (e.g., *Helicobacter pylori* eradication). These characteristics and a uniform population response have mitigated against misinterpretation of these simplified constructs. However, with more complex disorders like bipolar disorder, the variability in therapeutic effect across patients and over time, the multiplicity of effects and resulting permutations of possible results as well as the sampling and statistical problems of the NNT and the PI limit the utility of PI as a single conceptually meaningful measure. Indeed, an adequate presentation of these values would require a spectrum of values that are likely to be so broad and require so many qualifiers as to provide minimal clinical utility.

Role of the funding source

Support for this commentary was provided by Janssen Scientific Affairs, LLC.

Contributors

All authors contributed to the development and preparation and review of this commentary.

Conflict of interest

Conflict of interest statements will be provided off-line.

Acknowledgment

The authors acknowledge the editorial assistance of Susan Ruffalo, PharmD.

References

Bartzokis, G., 2012. Neuroglialpharmacology: myelination as a shared mechanism of action of psychotropic treatments. Neuropharmacology 62, 2137–2153.

Citrome, L., Katrowitz, J., 2008. Antipsychotics for the treatment of schizophrenia: likelihood to be helped or harmed, understanding proximal and distal benefits and risks. Expert Rev. Neurother. 8 (7), 1079–1091.

Citrome, L., 2010. Adjunctive aripiprazole, olanzapine, or quetiapine for major depressive disorder: an analysis of number needed to treat, number needed to harm, and likelihood to be helped or harmed. Postgrad Med 122 (4), 39–48.

Citrome, L., 2012. Lurasidone for the acute treatment of adults with schizophrenia: what is the number needed to treat, number needed to harm, and likelihood to be helped or harmed? Clin. Schizo. Rel. Psych., 75–85.

Hutton, J.L., 2000. Number needed to treat: properties and problems. J. R. Stat. Soc. A 163 (Part 3), 403–419.

Lesaffre, E., Pledger, G., 1999. A note on the number needed to treat. Control Clin. Trials 20 (5), 439–447.

Popovic, D., Reinares, M., Amann, B., Salamero, M., Vieta, E., 2011. Number needed to treat analyses of drugs used for maintenance treatment of bipolar disorder. Psychopharmacology (Berl). 213 (4), 657–667 Epub 2010 October 31.

Popovic, D., Reinares, M., Goikolea, J.M., Bonnin, C.M., Gonzalez-Pinto, A., Vieta, E., 2012. Polarity index of pharmacological agents used for maintenance treatment of bipolar disorder. Eur. Neuropsychopharmacol. 22 (5), 339–346.

Smeeth, L., Haines, A., Ebrahim, S., 1999. Numbers needed to treat derived from meta-analyses—sometimes informative, usually misleading. Br. Med. J. 318 (7197), 1548–1551.

Stang, A., Poole, C., Bender, R., 2010. Common problems related to the use of number needed to treat. J. Clin. Epidemiol. 63, 820–825.

Suissa, D., Brassard, P., Smiechowski, B., Suissa, S., 2012. Number needed to treat is incorrect without proper time-related considerations. J. Clin. Epidemiol. 65 (1), 42–46 Epub 2011 August 4.

Thabane, L., 2003. A closer look at the distribution of number needed to treat (NNT): a Bayesian approach. Biostatistics 4 (3), 365–370.