CONFOUNDING BY SEVERITY AND INDICATION IN OBSERVATIONAL STUDIES OF ANTIDEPRESSANT EFFECTIVENESS

Scott B Patten

Department of Community Health Sciences, University of Calgary, Canada

Corresponding Author: patten@ucalgary.ca

ABSTRACT

Background
It has been suggested that antidepressants worsen the course of major depressive disorder. Epidemiological data have sometimes been cited in support of this idea, but such estimates are vulnerable to confounding. The objective of this study was to assess episode incidence and recovery in relation to antidepressant use, adjusting for symptom severity.

Methods
Random digit dialing was used to select a sample of n=3304 community residents. Each respondent was then assessed with a baseline interview followed by a series of six subsequent interviews spaced two weeks apart. The brief Patient Health Questionnaire (PHQ-9) was used to detect depressive episodes during follow-up and to provide ratings of symptom severity. Grouped time proportional hazards models were used to assess confounding by producing estimates of the association between antidepressant use and major depression incidence and prognosis adjusted for baseline symptom severity.

Results
Antidepressant use in initially non-depressed respondents was associated with a markedly higher incidence of depression (Hazard Ratio, HR = 3.9, 95% CI 1.8 – 8.5). With adjustment for the depression severity score in the two weeks preceding the emergence of a new episode, this effect diminished markedly and was no longer statistically significant (HR = 1.2, 95% CI 0.6 – 2.7, p = 0.57). Antidepressant use was also associated with a lower rate of recovery from major depression (HR = 0.8, 95% CI 0.5 – 1.2, p = 0.27), but this effect also moved towards the null value after adjustment for baseline severity (HR = 0.9, 95% CI 0.6 – 1.5).

Conclusions
Antidepressant medication use is confounded with symptom severity. Observational studies seeming to show harmful effects of antidepressants are subject to bias as a result.

Key words: *Antidepressive agent; longitudinal studies; epidemiology; methods*

Antidepressant medications are considered a first-line treatment option for depressive disorders.[1] The efficacy of these medications has been confirmed by many randomized controlled trials; although, methodological features of these trials and publication bias may have led to an exaggerated impression of their efficacy.[2,3]

An opposing argument has also been made: that antidepressant treatment may cause long-term neurobiological changes that may worsen the course of depressive disorders.[4] Canadian epidemiological data indicate that people taking antidepressant medications tend to have longer and more frequent episodes of major depression[5] and an observational clinical study in the UK reported that antidepressant treatment did not lead to improved outcomes in real world clinical practice.[6] Some authors have openly questioned whether these medications have any specific efficacy at all.[7] However, confounding by severity and duration, and potentially by other factors, may explain these epidemiologic findings.[8] Since

treated and untreated episodes are likely to have differing characteristics, antidepressant outcomes suggesting a lack of effectiveness may be due to confounding. A particular concern is confounding by severity, as people seeking treatment for major depression are likely to be more depressed than those not seeking treatment. Also, people on long-term antidepressant treatment are likely to be at high risk of depression and may also be characterized by higher levels of depressive symptoms.

A recent longitudinal telephone survey in the Canadian province of Alberta provided an opportunity to evaluate this issue by examining major depression incidence and recovery in relation to antidepressant treatment with and without adjustment for symptom severity.

METHODS

The study sample consisted of household residents in the Canadian province of Alberta. The sample was selected from a database of residential telephone numbers using random substitution of the final digit to avoid bias that might result from failure to reach unlisted residential telephone numbers. Each number was dialed at least nine times in an attempt to determine whether it reached a residential household or not. When a household was contacted, a household member willing to participate in the study was sought. To be eligible, the volunteering household residents were required to be between the ages of 18 and 65. The study was approved by the University of Calgary Conjoint Health Research Ethics Review Board.

The interviews were carried out using a Computer Assisted Telephone Interview (CATI). The measure of depression employed in the study was the brief Patient Health Questionnaire (PHQ-9).[9,10] This is a symptom rating scale whose items cover the nine symptoms comprising the DSM-IV "A" criterion for major depression. The items have a common stem referring to how much a respondent was "bothered by" specific symptoms during the 2-weeks preceding the assessment, also consistent with time frames referenced in the DSM-IV criteria.[11] The PHQ-9 can be scored either as an ordinal depression severity rating (the sum of nine item scores each assigned a value of 0-3, for a total score that can range between 0 and 27) or with an algorithm that is based on the DSM-IV definition.[12] The algorithm requires either depressed mood or loss of interest to be rated at a "most days" level, but accepts suicidal ideation at any level of severity. A total of five of nine symptoms must be endorsed to fulfill the algorithm, as also required by DSM-IV.[11] Additional interview items collected include demographic information and medication use data.

The respondents were recontacted two weeks later for re-administration of the PHQ-9. These follow-up interviews were also conducted over the telephone. A total of six follow-up interviews were conducted so that the longitudinal follow-up stretched over 12 weeks. The analysis initially examined adjusted and unadjusted associations between antidepressant use and (algorithm defined) major depression incidence. Next, the association of antidepressant exposure with recovery from depression was evaluated. Grouped time proportional hazards models were used, as described by Jenkins.* A parametric form of the model was used, a Weibull model. The Weibull model was chosen because incidence and recovery rates in major depression are time dependent in a way that is well described by this type of model: the incidence rate declines with time since a prior episode and the recovery rate declines with increasing episode duration.[13] As an adjustment for severity, PHQ-9 at the baseline of each incidence or recovery interval was added to the models as a covariate. These scores and a variable representing antidepressant exposure were allowed to vary with time in the part of the analysis concerned with incidence, but not in the part of the analysis concerned with prognosis. It should be emphasized that the adjustment for symptom severity in the incidence analysis included the symptom severity measure at the time point before the incident depressive episode; this was not the same rating that was used to determine whether an episode occurred. The analysis therefore determined whether adjustment for baseline symptom severity altered the association of major depression incidence with antidepressant exposure. In the part of the analysis concerned with prognosis, baseline values for

*http://www.iser.essex.ac.uk/teaching/degree/stephenj/ec968/pdfs/STB-39-pgmhaz.pdf

symptom severity were used as a covariate. This was not allowed to vary with time.

RESULTS

During the sampling, a total of 36068 telephone numbers were called. There were 116 indeterminate call dispositions: 68 answering machines, 2 numbers that were always busy and 46 that were never answered. A large proportion of telephone numbers were disqualified: 3264 reached households in which no members were in the eligible age range (18-65 years), 6660 numbers were not in service, 74 numbers did not reach a primary residence (e.g. reached a recreational residence), and the maximum number of call-backs were made to 5801 residences without reaching a person living there. There were 411 blocked calls, 2869 numbers that connected to businesses and 2507 reached fax machines. A language barrier was encountered at 366 households and a "fast busy" signal of uncertain significance was reached at 58. In total, 22010 numbers were disqualified. There were 10608 refusals, such that, in total, the 36068 numbers reached 3334 eligible participants. There were 30 baseline interviews that were only partially complete, such that the final sample consisted of 3304 people.

The sample included 1,068 men and 2,236 women, with a median age of 45. In the sample 72.6% were married. Demographic estimates from Statistics Canada (Summary Tables in Canadian Statistics, CANSIM accessed via E-Stat on July 8, 2008) indicated that the population within the eligible age range in Alberta includes 51% males, has a median age of 39 and is 62% married. As such, the sample over-represented women and married people, and under-represented younger age groups. In view of the study objectives and the nature of the sampling, however, the estimates were not weighted. When asked at the baseline interview whether they were currently taking antidepressants, 11.1% of the sample responded affirmatively. Of the 3304 initial respondents, 2667 (80%) were successfully followed to the 12-week time point.

Among those without major depression at the baseline time point, 1.2% developed an episode of major depression in the subsequent two weeks, with the cumulative incidence climbing to 3.8% over the entire 12 weeks of follow-up. The crude hazard ratio (HR) for new onset major depression associated with antidepressant use was 3.9 (95% CI 1.8 – 8.5), an approximate four fold increase in those being treated with antidepressants. With adjustment for depression score, this effect diminished markedly and no longer achieved statistical significance (HR=1.2, 95% CI 0.6 – 2.7, p=0.57). Addition of other variables, including age and sex, in the model did not result in substantial changes to the adjusted HR for antidepressant use. Having a past history of depression was associated both with major depression incidence and antidepressant use. With inclusion of this variable, the HR moved even closer to the null value, to 1.1 (95% CI 0.5 – 2.5). This suggests the possibility of confounding by indication as well as severity. Non-depressed respondents who report taking antidepressants and who deny a past history of depression are probably taking them for indications other than depressive disorders.

A similar analysis of the recovery pattern was conducted. There were n = 241 respondents who were depressed at the initial interview, and for whom there was successful follow-up for at least one interview so that they could be included in the modeling. Of these 241 respondents, 99 (41.1%) were taking antidepressants. The proportion recovering during each time interval was highest at the first interval, where 43.6% of those with baseline depression no longer fulfilled the requirements of the PHQ-9 algorithm. At this first follow-up interval, those on antidepressants were significantly less likely to have recovered (32.3% versus 51.4%, Fisher's exact p=0.004) than those not on antidepressants. However, there were no significant differences during the subsequent intervals. The unadjusted HR for antidepressant use was 0.8 (95% CI 0.5 – 1.2), which was not significantly different from the null value, p = 0.27 but suggested a lower probability of recovery in those on antidepressants. After adjustment for baseline PHQ-9 score, this effect became closer to the null value (HR = 0.9, 95% CI 0.6 – 1.5). Addition of age and sex to the models did not alter these results.

In the analyses reported above, collinearity is a concern since the PHQ-9 was used both for assessment of major depression status (using the algorithm) and for assessing symptom severity, even though different PHQ-9 ratings were used for this purpose in the analysis.

FIG. 1 PHQ-9 Score Ranges at Baseline for those Recovering, or Failing to Recover, from Major Depressive Episode Present at Baseline

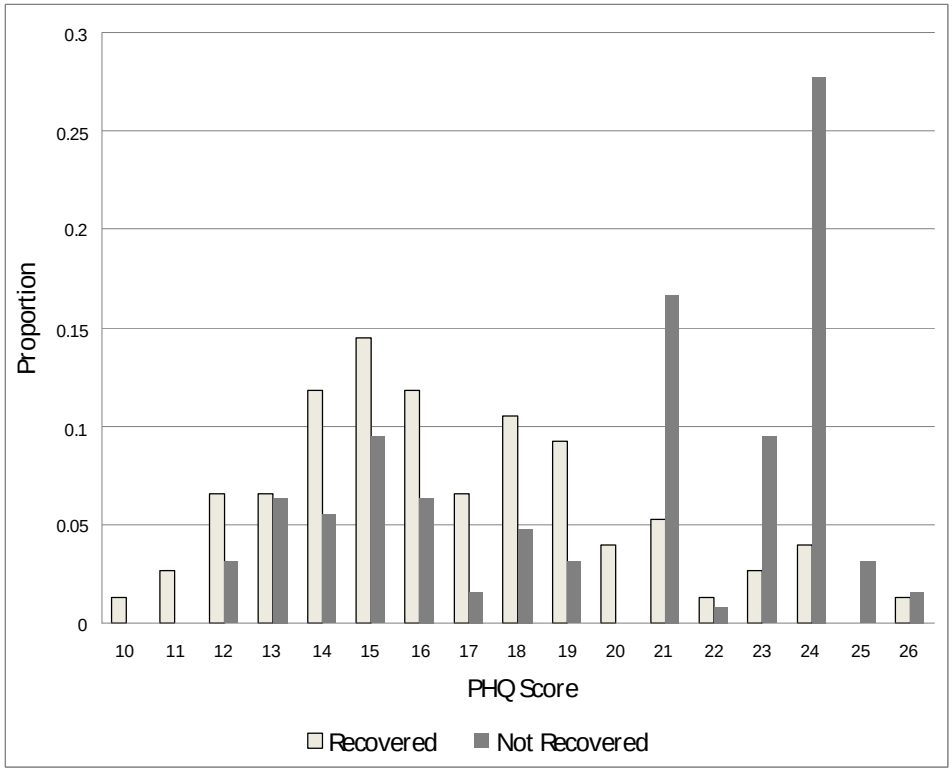

Figure 1 displays the frequency of baseline PHQ-9 scores in respondents who did or did not recover from their major depressive episode during follow-up. Whereas respondents who went on to recover often had lower scores (the basis for the occurrence of confounding in this context), both groups presented a range of scores with extensive overlap.

DISCUSSION

Poor outcomes observed in epidemiological studies of antidepressant treatment have raised concerns about the effectiveness of these medications in real world settings. However, it has been pointed out that confounding by severity and duration of episodes could account for these findings.[8] The current analysis found evidence for confounding by severity. This study was a telephone survey which needed to rely on a brief measure of symptom severity, although the PHQ-9 is considered a valid instrument.[9,14] Usually, an inaccurate measure of a potentially confounding variable would lead to incomplete control of confounding. For this reason, the results presented here probably fall short of providing a fully adjusted estimate; yet nevertheless, the changes observed with adjustment for symptom severity do document the occurrence of confounding.

It should also be acknowledged that this study fails to provide evidence of the effectiveness of antidepressant medications. The adjusted estimates did not confirm superior outcomes, either in terms of incidence or recovery, for those under treatment. The results suggest, however, that evidence for or against the efficacy of these medications must derive from studies that employ highly effective strategies for control of confounding. Since this may be a difficult objective to achieve in observational

studies (even with advanced techniques such as propensity score analysis), the results suggest that randomization may be an essential design feature in studies assessing antidepressant effectiveness.

Acknowledgments

This study was a secondary analysis of data collected in a prior survey and did not receive funding. The data derived from the Alberta Surveys Initiative.

REFERENCES

1. CANMAT Working Group. Clinical guidelines for the treatment of depressive disorders. Can J Psychiatry 2001;46(Suppl. 1):1S-92S.
2. Moncrieff J. Are antidepressants overrated? A review of methodological problems in antidepressant clinical trials. J Nerv Ment Dis 2001;189:288-95.
3. Turner EH, Matthews AM, Linardatos E, et al. Selective Publication of Antidepressant Trials and Its Influence on Apparent Efficacy. New Eng J Med 2008;358:252-60.
4. Fava GA. Can long-term treatment with antidepressant drugs worsen the clinical course of depression? J Clin Psychiatry 2003;64:123-33.
5. Patten SB. The impact of antidepressant treatment on population health: Synthesis of data from two national data sources in Canada. Population Health Metrics 2004;2/1/9.
6. Brugha TS, Bebbington PE, MacCarthy B, et al. Antidepressants may not assist recovery in practice: a naturalistic prospective survey. Acta Psychiatr Scand 1992;86:5-11.
7. Moncrieff J. Are antidepressants as effective as claimed? No, they are not effective at all. Can J Psychiatry 2007;52:96-7.
8. Patten SB. In debate: Are antidepressants as effective as claimed? (Letter). Can J Psychiatry 2007;52:750.
9. Spitzer RL, Kroenke K, Williams JBW, et al. Validation and utility of a self-report version of the PRIME-MD. The PHQ Primary Care Study. JAMA 1999;282:1737-44.
10. Spitzer RL, Williams JBW, Kroenke K, et al. Validity and utility of the PRIME-MD Patient Health Questionnaire in assessment of 3000 obstetric-gynecologic patients: The PRIME-MD Patient Health Questionnaire Obstetric-Gynecology Study. Am J Obstet Gynecol 2000;183:759-69.
11. American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders (DSM-IV-TR). Washington: American Psychiatric Association; 2000.
12. Kroenke K, Spitzer RL, Williams JBW. The PHQ-9. Validity of a brief depression severity measure. J Gen Intern Med 2001;16:606-13.
13. Patten SB. A visual depiction of major depression epidemiology. BMC Psychiatry 2007;7:23.
14. Löwe B, Spitzer RL, Gräfe K, et al. Comparative validity of three screening questionnaires for DSM-IV depressive disorders and physicians' diagnosis. J Affect Disord 2004;78:131-41.