Author’s response to reviews

Title: Treatment Patterns and Sequences of Pharmacotherapy for Patients Diagnosed with Depression in the United States: 2014 through 2019

Authors:

David Kern (dkern2@its.jnj.com)
M. Cepeda (scepeda@its.jnj.com)
Frank Defalco (fdefalco@its.jnj.com)
Mila Etropolski (metropol@its.jnj.com)

Version: 1 Date: 08 Nov 2019

Author’s response to reviews:

Hans Rittmannsberger, Prof. (Reviewer 1)

An observational study on an interesting topic, relying on a huge data base. The methodology seems to be sound, yet the complexity of the issue forces to some simplifications which are discussed by the authors.
There are some points of critique:

The inclusion criteria need to be explained in more detail in the methods section

We’ve added more detail to the Methods section regarding patient selection to make this more clear.

- It remains unclear for what reason two outpatient diagnoses of depression were required to be included into the study.
- p12, line24: "...we required treatment to occur at the time of or following the first diagnosis of depression …" This should be included in the methods section - and how does this pair with the second diagnosis?

The requirement of two diagnoses was applied to be consistent with the validated algorithm for identifying depression patients. Typically in claims data a second diagnosis is required to reduce the number of false positives, i.e., including patients who don’t truly have the condition. It’s possible that a single diagnosis is a rule-out diagnosis or misdiagnosis, and it would be expected that individuals with depression who are seeking treatment would go to the doctor more than once in a one-year period. There is no requirement that the second diagnosis has to be from the same or different provider, only that it occurs on a different date and within one year of the first.

All together the results differ largely from common treatment recommendations
- half of the patients stay without treatment
- one third is treated with other drugs than antidepressants first-line
- the amount of combination or augmentation therapies (which by many authors is preferred compared to a switch) is small.
Could the methodology of the study account for some of these results? Or is there just a gap between theory and practice and then – why? Of course, this question was not the topic of the study, but it seems to be its message and deserves more comments.

These are all very important comments. Note that when updating the analysis to address a comment from reviewer #2 we discovered an error in how combination therapy was captured. Now we observe a higher prevalence of combination therapy. We’ve updated the discussion section to address how our results relate to treatment guidelines and why we might be seeing a high proportion of patients not receiving treatment and a high rate of non-antidepressant use first-line. It may be due to the treatment of comorbidities and symptoms of the depression rather than treatments that are specific to the underlying depression itself. We also added that depression patients who don’t receive any pharmacological treatment may be due to receiving alternative therapies such as psychotherapy.

Early discontinuation of antidepressant treatment is a major problem, not only in the USA (see citation 10), but worldwide 1-3. Given the possibilities of this data base this issue should be covered too.


Unfortunately, looking at how long patients were actively on treatment, the timing of moving from one class to the next, and the amount of time patients may have gone without treatment between various treatment lines was out of scope for this study and would indeed make a compelling manuscript on its own. We’ve added text to the Limitations regarding this and have included the citations above, as well as two for the US.

David Benrimoh (Reviewer 2)

Summary:

This is an interesting paper attempting to reconstruct treatment patterns for patients with MDD using national claims databases. This information would be useful for a number of reasons, from research on clinician patterns and quality of care to work aiming at better quantifying social and economic costs of care. The large sample size is a strength of the study. The reliance on claims data, however is a weakness inherent in this kind of work given that claims do not always accurately reflect diagnostic and clinical reality. While this is counter-balanced by the fact that the data presented represent true clinical practice (and therefore also represent the real diagnostic and management inconsistencies that
happen in real clinical practice), caution should be exercised on the part of the reader when assessing the paper.

Comments:
1. In the introduction (and I'm sure this was just a simple oversight), TMS was classified as an invasive treatment. This should be corrected.

This has been corrected

2. On line 32, reference 2 refers, I believe, to a paper describing the rationale but not the actual results of STAR*D. This reference should be updated and the sentence along with it.

The reviewer is correct this refers to the rationale behind the STAR*D studies. Part of that rationale is the low rate of treatment response and remission which this sentence is referring to.

3. The terms "adjunctive therapy" and "adjuvant therapy" are both used. I suggest picking one and running with it for clarity's sake.

The text has been revised for consistency, per the reviewer’s suggestion.

4. You explain why the end date of data collected was 1/31/2019, but it is not clear why the start date was chosen (i.e. 1/1/2014) Is it because these databases did not exist before then? Please specify.

We chose the start date of 1/1/2014 so that our treatment patterns would include the most recent 5-year period. We wanted our results to reflect current treatment practices rather than historical ones. We’ve added text to the methods section explaining this and added start dates of each of the data sources in their respective sections, as all available data was used when excluding patients who received prior treatment.

5. In your specification of the depression cohort, I believe the reader deserves more explanations about why you required two outpatient diagnoses, why patients already on treatment were excluded, and, most importantly, why patients needed to be consistently enrolled in the database for the time period specified. This last point is especially important, given that there may be systemic bias introduced if patients with depression, or with specific depression courses, are less likely to be enrolled in the database (which is likely to be true). These are not deal-breakers for the study, but do require full discussion in the limitations section and brief justifications in the methods section.

Two outpatient diagnoses were required because we are following the only validated algorithm for identifying depression patients in claims data, which also used two outpatient diagnoses. The requirement of two diagnoses is common practice in claims data as it improves specificity and makes it less likely to include patients who received a rule-out diagnosis or single misdiagnosis for the condition. We excluded patients with prior treatments because we wanted to include patients who were newly diagnosed and wanted to start our analysis at the first evidence of treatment, as our treatment pattern analysis begins with first treatment received. Had we included patients with previous treatment
without a sufficient wash-out period we could not be certain that they were truly new depression patients and that their observed treatment was the first received. The point about requiring a minimum amount of follow-up, and in this case a significant one, is well-taken. When choosing this follow-up requirement we wanted enough follow-up across the population that would allow us to see multiple lines of therapy and various treatment changes. It’s possible, and even likely, that by doing so we are excluding a certain subset of individuals with depression, but the alternative of requiring too short of a follow-up period would prevent us from seeing what happens during later lines of therapy and would distort the observed treatment patterns in the population as a whole. Text in the methods and limitations have been added to address these points.

6. Why was a switch from one class to another and then back to another class recorded as only one switch? Clinically, it is meaningful for a patient to be on, say, sertraline, then venlafaxine, then citalopram. That sequence might be explained for example by a patient experiencing bad side effects with the first two drugs and then being switched to a third that is seen by clinicians as having fewer side effects. It might also represent loss of insurance for non-generic drugs. In sum, it might be meaningful so why ignore it? Please provide a rationale.

The scope of this work was to examine the order in which various treatment classes were first received and whether they were received in combination with each other. Had we done this at the ingredient level, then changes like the one described above would have been captured. However, by doing that the data would have become incredibly granular as there are dozens of different ingredients that would be included across all the treatment classes. Future analysis may either look more granularly at individual drugs to capture in-class switching and/or examine switching back to a previously used medication class, but for simplicity and clarity the scope of our work was limited to capturing the first exposure to each medication class.

7. Are the databases mutually-exclusive? Is there certainty that there is no double-counting?

The IBM MarketScan databases are mutually-exclusive, in that a single patient can only exist in one of the three databases at any given time. However, it is possible there may be overlap between the Optum and IBM CCAE databases. The overlap between databases is unknown and trying to identify overlap between the databases would be a violation of the data use agreements.

8. I am somewhat concerned by some of the results from table 1 and believe they should be further discussed. First, aside from anxiety disorders, dysthymia, and drug dependence, I don't see other psychiatric comorbidities one would expect (PTSD, psychotic disorders, OCD, and especially personality disorders); I also do not see a discussion of bipolar disorder anywhere, and this should be discussed especially if bipolar disorder was not ruled out.

We originally included conditions that were prevalent in more than 10% of the population overall and the comorbidities noted above didn’t meet this threshold. However, because all of the conditions noted by the reviewer are important for providing additional context about the observed treatment patterns, we added each of them as “additional conditions of interest”. All are relatively uncommon, occurring in <5% of patients, with the exception of psychotic disorder in the MDCD population (6.9%), which could account for the increased use of antipsychotics in this population. We’ve added text about these results to the Results section and included these results in Table 1.
We’ve also added text to the methods section about bipolar disorder. Codes for bipolar disorder were not included in the validated depression definition; however, these patients were also not explicitly excluded from the analysis, resulting in roughly 2% of patients having a diagnosis of comorbid bipolar disorder. This proportion increases slightly when expanding the time frame for identifying comorbid conditions to include one year following the index date (365 days prior to 365 days following the index depression diagnosis), but still accounts for only a small fraction of patients, with 3.5% of patients having at least one diagnosis code for bipolar disorder in CCAE, 8.4% in MDCD, 2.3% in MDCR, and 3.3% in Optum.

Second, the anxiety comorbidity rate is lower than I would expect, raising the very real possibility that the claims data is not complete.

Patients were classified as having anxiety disorder if they visited a physician seeking care for their anxiety which would lead to the provider filing a claim with an ICD-9/10 code for anxiety disorder. Like any other condition, patients who do not seek care for their anxiety will not be captured in the claims database and thus the observed prevalence of anxiety will be lower than the truth. It’s important to note that the time period for capturing comorbidities includes the year prior to the index date, i.e., the baseline period. If an individual developed anxiety after their diagnosis of depression they would not be captured in this calculation. Had we included the year prior to and the year following the index date, the prevalence of anxiety disorder would have increased from 25.7% to 42.9% in CCAE, 21.3% to 34.5% in MDCD, 19.8% to 31.0% in MDCR, and 27.3% to 41.6% in Optum which may be closer to expectation.

Third, you note an increased rate of drug dependence in Medicaid patients, but you don’t offer a statistical test for this (it’s likely to be significant but it would be good form to do it because I’m not sure how variable the datasets are); this same concern can be levelled at your result about suicide. These comments should be addressed in the results and discussion.

We didn’t include statistical testing for a few reasons: 1) there is no a priori hypothesis that is being tested; 2) with 4 different databases there are 6 different pairwise comparisons [(n-1)!] which is too many to include in any concise way. We could have done a single test to see if there was any difference between all groups, but that wouldn’t tell us exactly which groups are different from each other; and, perhaps most importantly, 3) with such a large number of patients even very small differences will be statistically significant and so the more important inference is the qualitative difference between groups. The “variability” of databases (assuming the reviewer is referring to variation within a given database) will have no impact on testing differences in proportions, as the standard errors are determined solely according to the proportions and the number of individuals. To illustrate this point we calculated the z-score for the difference in the two proportions of female patients between CCAE and Optum. The proportions are very similar (60.4% vs. 62.4%); however, the z-score is -8.6912 which yields a p-value of <0.00001, extremely significant at any reasonable alpha level, and provides no useful information. Text has been added to the methods section detailing how we captured baseline covariates and the lack of statistical testing.

9. The high proportion of people receiving non-SSRI initial therapy is a problem for the claim that these are all really patients with depression. For example, these could be patients with psychotic disorders or bipolar disorder who happen to have experienced a depression in the period you studied; this might also explain partly why Medicaid patients received different treatment types.
We do not believe presence of psychotic disorder or bipolar disorder is the reason for seeing the high rates of non-SSRI use. Very few of these patients had psychotic or bipolar disorders as evidenced by the new data added to Table 1. All patients met the depression diagnostic criteria regardless of their comorbid burden, and given the high PPV of the algorithm used, we are confident that nearly all patients included truly had depression. This doesn’t mean that the treatments are not being used for conditions other than depression, but there isn’t evidence that these are treatments for conditions such as psychotic or bipolar disorders.

Also the use of non-standard treatments or combination treatments might relate to patients with a history of treatment resistant depression.

It is highly unlikely that these treatments being used as the first or second option are for treatment resistant depression (TRD) patients. We started follow-up on the first observed diagnosis of depression in the database with at least one year of prior observation, so these are newly diagnosed depression patients that would not have met criteria for TRD at the time treatment patterns begin to be captured.

These concerns betray the fundamental issue of using claims data- it is not complete and does not necessarily provide an accurate look at all patients and it may be less accurate for certain more complex patients (so the inaccuracy in itself may be biased). A full discussion of this in the limitations is warranted because without this discussion to balance your claims the reader could be misled.

The authors respectfully disagree that claims data do not provide an accurate assessment of the study question. In fact, we would argue that claims data offer the best data for assessing real-world treatment practices within large, diverse populations. While we cannot be 100% certain that an individual receiving a diagnosis code for some condition truly has that condition (in this case, depression), using a validated algorithm increases our confidence. Additionally, one of the biggest strengths of claims data is the pharmacy data, as it provides records of the exact medication that was dispensed to a patient; there is no ambiguity. Various additions to the limitations of claims data have been added to the Limitations section.

10. You absolutely must further discuss the algorithm (ref 8) you used to identify depression in claims data in the methods and you should discuss it and report all of the metrics. Sure, it has a high PPV- but what is the NPV? And the PPV depends on the sample the algorithm was trained in- are you sure your sample is equivalent? I believe you need to work much harder to convince the reader of the accuracy of your depression diagnosis, because the prescribing patterns you discovered have two explanations: bad practice by clinicians (certainly possible) or bad sampling (very possible). Please expand on this in the methods and the discussion. Also, the trouble of working with claims data should also be discussed up front in the introduction.

This is the only validated definition of depression patients available, and rather than rely on gut instinct or blindly follow what others have done in the past we wanted our definition of depression to be supported by data. Only PPV was assessed in the cited validation study (see Solberg et al reference for complete details). This is because the design of the validation started with cases observed in claims and then sought to confirm those diagnoses via review of the medical charts. The validation study identified cases of depression from an unnamed administrative claims database using the same algorithm applied in this study (2 outpatient diagnosis or one inpatient diagnosis with one of the ICD-9 codes listed in the Methods section). The absence of including non-cases from the claims in the validation precluded the calculation of additional metrics such as sensitivity, specificity and NPV. Unfortunately, there is no
additional information on the characteristics of the population used in the validation. It is theoretically possible that this definition may not be portable to some degree to other sources of administrative claims data, but it is hard to imagine why this would be the case.

In our study the main concern is whether the patients included in our study are truly depression patients, and while it is only one small validation study, it is the only hard evidence we have, and it shows that the definition does a very good job at including only patients who have depression. However, typically when algorithms have such high measures of PPV, the sensitivity is relatively low, and thus we may be missing some depression patients in our analysis. This is a trade-off we are comfortable with; we believe it is more important that those included truly have the condition rather than attempting to include all depression patients at the expense of including a large number of individuals who don’t have the condition. Additional text on the limitations of claims data has been added to the discussion section. The Introduction is not the appropriate section for this discussion.

11. In your limitations you say that you tried to limit "misclassification due to receiving therapy for reasons unrelated to depression, we required treatment to occur at the time of or following the first diagnosis of depression with no prior history of treatment"- however, this is misleading, because patients may very well have had treatment prior to the time window you are looking at, and went off medication, then came back on when they relapsed.

The index date for patients was the date at which the first diagnosis of depression occurred. All patients had at least one year of look-back but theoretically could have had up to 16 years of prior observation as the data in most of the databases goes back to 2000. Patients were excluded if they had prior history of treatment at any time during this lookback period. So, yes, while they may have had treatment before they entered the database this would have had to occurred more than one year ago and before they entered the database, and for most patients it is much longer than that.

In addition, their "reason unrelated to depression" might very well have relapsed at the same time as the depression (for example, a schizoaffective disorder patient might relapse into a psychotic episode with depressive features, and the depression might be what is picked up by the ER doctor.

This may be true, but it is also one of the reasons why we required two outpatient (which includes ER) diagnoses of depression on different dates to be more certain that this is a patient receiving care for depression over time. It’s also true that a patient might simultaneously receive one of the treatments of interest for another condition at the time they are newly diagnosed with depression, but looking at rates of comorbid conditions such as psychotic disorder, PTSD, and others, this doesn’t appear to be what is driving the observed treatment patterns.

As such I would recommend amending this to "…with no prior history of treatment within the time window examined."

Amended to say “… with no prior history of treatment in the database” as the lookback window varies according to how long the patient was in the database prior to being diagnosed with depression (min of 1 year).

And I'd also add something about how hard it is to know for sure that the only thing being treated was a depression.
This is a good point and it is what we were driving at in our original limitations statement. We’ve added more text to the limitations section to expand on this point.

12. Table 3- how do you differentiate anxiolytics from sedatives? Benzodiazepines are both…

We used the ATC classification for the classification of anxiolytics (ATC: N05B) and hypnotics/sedatives (N05C). Benzodiazepines fall in both categories, though the same benzodiazepine can only occur in one class according to the classification system. We have added an appendix with each of the individual medications corresponding to each class.

13. Where is the data on stimulants? These are often used as adjuncts. What about Lithium? Which is also used as an adjunct.

We’ve added stimulants and lithium to the analysis. Lithium use was very rare and didn’t impact results but use of psychostimulants was somewhat prevalent.

14. In figure 1, you discuss how you identify combinations and switches. Is this how you do it in your analysis? If so please put in methods. Also I'm not convinced by the definitions; I'm a clinician and if I forget to discontinue a drug (formally, I mean; I still tell the patient not to take it) after starting a new one because side effects were intolerable (and this happens), then if the overlap was 30 days then it's combination? Or what about cross-tapering? We often switch from one drug to another over a long time period by gradually reducing the dose of one and increasing the dose of the other; this would mean when we're actually switching, you might be calling it combination! In other words, this is a (very) rough definition of switch vs. combo and should be discussed in the results. This might explain, for example, the odd finding of switching from an SSRI to an anxiolytic- which just doesn't make sense. More likely, the patient could not tolerate the SSRI, got an anxiolytic to help with sleep, stopped the SSRI because of the side effects, but stayed on the anxiolytic because they are hard to stop. As such calling this switching without describing the (major) caveats is misleading.

Yes, Figure 1 is how we identified combinations and switches. How continuous use and combination use are defined was explained in the methods section, but we’ve added additional text regarding switches. Note that when we re-ran the analysis to include stimulants and lithium we discovered a bug in the software used to categorize combination therapy. The new analysis, after the bug was resolved, shows higher use of combination therapy that increases in later lines of therapy, which is more in-line with expectations.

The reviewer raises very good, important points regarding potential misclassification of switching and adding on and further text has been added to the limitations. In defining any variable, including the identification of switching and add-on therapy, there has to be some rule that is consistently applied across all data. We cannot review charts for a quarter-million patients to understand the nuances of why these patterns exist, so we need to have an algorithm that minimizes misclassification as much as possible, but with the understanding that removing all misclassification is impossible – even if we were to review individual charts misclassification would still exist. We feel that the 30-day overlap offered us the cleanest way to identify combination therapy versus a switch because most prescriptions are 30 days [aside, to highlight this point: of the 50,827,025 prescriptions for an SSRI in the CCAE database since 2014, 37,879,871 (75%) had exactly a 30-day supply]. Presence of a 30-day overlap in the context of 30-day prescriptions would require continuing to fill one drug while also filling a second,
new drug. While this method isn’t perfect, and will lead to some misclassification, requiring a shorter overlap period would lead to a higher rate of falsely classifying switches as combination use, and requiring a longer overlap period would miss a larger number of true short-term combination use by classifying it instead as switching.

Overall, this is an interesting paper with useful information, but the methods and limitations need much more beefing up and the claims made should be contextualized given these limitations.

Thank you for your comments which have helped make this a stronger manuscript.