nature portfolio

Peer Review File

Comparative Effectiveness of Treatment Strategies for Bipolar Disorder During and After Lithium Treatment

Corresponding Author: Dr Johannes Lieslehto

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
This very large population epidemiological study with extensive longitudinal data examined the utlity of antipsychotics, mood stabilisers and combination therein in people with BP and either intolerance or inadequate response with Lithium. Using data from two large nationwide cohorts in Sweden (105,495 patients) and Finland (60,045 patients) with up to 23 years of follow-up, the study evaluates mood stabilizers, antipsychotics, and their combinations in reducing psychiatric hospitalization risk.
Main findings include:
1. Clozapine-based regimens: Clozapine combined with aripiprazole showed the lowest risk of psychiatric hospitalization compared to lithium monotherapy.
2. 'Lithium-intolerant patients': Among lithium discontinuers, combinations such as quetiapine plus lamotrigine and olanzapine plus valproate were associated with reduced hospitalization risk. LAI antipsychotics combined with valproate showed the lowest point estimate but may be limited due to smaller sample sizes.

However the manuscript needs a number of clarifications on premise and interpretation to be interpretable.

The introduction needs breaking down with clear evidence for definition and position of inadequate response: and the evidence gap needed over recent meta-analyses of augmentation, and then intolerance.

The latter is the more novel section and finding but it is merged in and difficult to parse out.

Please define inadequate response- how does this differ to treatment resistant BP. Please make it clear how the 'intolerance' is defined and used in this study- I believe where >6 months of Li is followed by cessation.

Indeed, how were BP with inadequate response defined and identified? There is an assumption from the within-individual analysis that pattern of prescribing is related inadequate response or intolerance. There are many reasons for medication regimes to change. Relapse and rehospitalisation can also be influenced by other confounds not measured.

The fact that clozapine is most effective in their outcomes is interesting- and may speak to increased monitoring and early intervention or clozapine unique pharmacological properties.

The methods are sound, on the whole, with potentially useful output however the limitations of the data are significant and need greater recognition.

The conclusion and recommendations eg support switching to clozapine-based regimens or augmentation with LAI antipsychotics when lithium monotherapy is inadequate are over-reaching for the level of data on prescribing decisions and choices available.

They do suggest that further trials in lithium resistant BP may indeed investigate LAI and clozapine regimes- and this is a sensible conclusion

Reviewer #2

(Remarks to the Author)
Lieslehto and seven co-authors (last author: H. Taipale) present an analysis of Swedish and Finish register data on the treatment of patients diagnosed with bipolar disorder. Their main interest is in patients who have not responded to lithium or have not tolerated it but in their study they also show results on the effectiveness of lithium. The authors used data from almost 170.000 patients who have been hospitalized more than 60.000 times. Their endpoint is all-cause psychiatric hospitalization, and they find that combinations of mood stabilizers (or aripiprazole) with either clozapine or long-acting antipsychotics are associated with a favorable relapse risk. At the same time, most combinations of two antipsychotics were deemed unfavorable.

The topic certainly warrants attention, and I found the general approach of the study instructive, but the manuscript is a very tough read, mainly due to its several layers of analysis.

Here are some points the editors and/or authors may want to consider:

1. It seems odd that more than 170.000 patients should have been hospitalized for only 60.000 times. Most patients with bipolar disorder I know have been admitted several times, almost always when they were manic, but also frequently when they were depressed. Can it be that admissions are missing from the data? Are the diagnoses correct? I am sure many readers will have difficulties understanding these figures, and authors must explain this clinical-experience-data mismatch.

2. I would strongly argue for restricting the analysis to patients who have not switched diagnoses to unipolar depression, ADHD, or schizophrenia/schizoaffective disorder because mood stabilizers are largely ineffective in patients with these diagnoses, except lamotrigine in patients with unipolar depression. Also, lithium, which is the starting point of the present analysis, is the gold standard in the long-term treatment of bipolar disorder, but nowhere else. This approach will reduce sample size and limit statistical significance but will greatly improve the interpretability of the data.

3. For clarification: In an additional figure, the authors could present examplary illness/treatment courses of patients to illustrate the different approaches they used, that is, within-patient as well as between-patients comparisons. Such a chart could show the timepoint of index diagnosis, treatment changes, durations of treatment and, most importantly, which phases were compared in the various analyses.

4. The authors keep calling what they find in their registers real world data. This, however,
is a difficult term because it pretends that the data truly represent the whole picture. However, as the authors themselves point out, they had to resort to hospitalization as an endpoint just because the more informative endpoint of any mood disorder episode was not available in this register, let alone other key outcomes, such as interepisodic psychopathology or tolerability. Also, RCT data originate from the real world, too, and are even more important than register data. It is precisely for this reason that the pharmaceutical industry very often uses the term real world data when results from RCTs are at variance with their interests. As a result, the term real world data should be avoided because of its imprecision, its ambiguity and because it has become a loaded term.

5. The title suggests that among the patients under study lithium has failed, when in fact it only reflects the authors opinion that lithium is the gold standard and should be used first (a well founded opinion, in my view). However, lithium failure is not a requirement for the main analysis presented in this paper. As a matter of fact, it is not the requirement for the other analyses either, since it is impossible to know from these real world data why lithium was discontinued. To manage expectations among readers, I would rephrase the title and change the introduction.

6. As an analysis of the results' robustness and generalizability: Is the ranking similar in Sweden and Finland? It is awkward for readers to go through the supplement tables themselves. However, figures vary considerably: For example, per supplements 1 & 2, in the Swedish cohort, lithium arrives at an adjusted HR of 0.83, in Finland, however, the respective value is 0.70. The clozapine plus aripiprazole combination reaches 0.26 in Finland and 0.61 in Sweden. These differences are substantial, even though HRs seem to be generally higher in Sweden (which differences may be responsible?). This is why it is so important to show the ranking.

7. In order not to bias the calculation against the index mood stabilizing drug, the authors, in a sensitivity analysis, removed the first 30 days from the analysis. Given that lithium usually needs months to take hold this period may be too short. For example, we advise our patients that they should take into acocunt at least one year before considering going off lithium salts.

8. The authors repeatedly refer to the harmonized nature of their data. Do they mean that they fitted Swedish and Finish register entries so that they could be used in one approach? Please clarify.

9. I did not understand the start (when did the data start?) and stop (when was the last data point?) dates of this analysis, and I would have welcomed these figures as early as in the summary.

10. Similarly, I would have welcomed the average time of follow up, not the upper end („up to 23 years").

11. How were lithium user defined? Anyone who was ever prescribed lithium? The same applies to the other drugs and combintions: any minimum duration of number of prescriptions/refills? Please clarify.

12. It is interesting that even though the authors reiterate that lithium is the gold standard only relatively few patients received lithium (line 100). How is that?

13. Can the authors differentiate between acute treatment and long-term treatment? For example, a patient in an initial manic phase may have receieved lithium (or olanzapine, to name only one antipsychotic) as an acute treatment, may have been admitted, anyway, but was then maintained on lithium (olanzapine) after his/her succesful treatment in the hospital. Would this count as an hospitalization under lithium (olanzapine)? If so, I the numbers are hard to interpret.

14. Can the authors contribute results to the order of pharmcological treatment? When were the various drugs applied (first drug, second…)? This may help interpreting the relative success/failure of the treatment.

15. Regarding the status „non treatment“: not treated but diagnosed? How many patients were in that group? Or, in within-subjects analyses, how long was this period? Is this a reasonable comparison?

16. There is an interesting meta-analysis supporting clozapine in the field of schizophrenia that the authors may want to mention: PMID: 40535000

17. Was there a minimum observation period? Otherwise, people may appear incredibly ill (if included in their index episode briefly before their inclusion ended).

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
The authors provide a detailed rebuttal and have made several substantial changes that improve clarity. However they continue to overstate the conclusions and clinical inferences that can be drawn. There are two points in the discussion that should be re-written in my opinion:

1. "Our results highlight clozapine-centered regimens as potentially superior alternative over lithium monotherapy, surpassing the scant evidence currently captured in guidelines."

The current guidelines are scant on clozapine recommendations because the RCT's have not been completed. At most, the authors can suggest that these trials need to be completed, in patients who are not able to tolerate or do not respond to lithium. In fact, BAP guidelines do recommend clozapine in refractory BAP, recognising the level of evidence is not strong.

2. "These data suggest that clinicians may consider clozapine-based regimens or augmentation with LAI antipsychotics when lithium monotherapy does not provide sufficient stabilization.' Among patients who discontinue lithium, specific antipsychotic–mood stabilizer combinations may also represent viable alternatives. Prospective trials are needed to confirm our findings"

Again the evidence is not strong enough in this paper to make clinical recommendations, where discontinuation is assumed to equate to intolerance or lack of efficacy. Clozapine is not liscenced for BP, at most the authors again can recommend head to head trials are completed.

The strengths and weaknesses section of the discussion needs to be expanded to greater reflect the level of uncertainty and assumptions in population data.

With these corrections, the manuscript is acceptable for publication, in my opinion. I will leave to the editorial team to decide if these results when accompanied by a balanced and appropriate discusison are novel enough for publication in NMH

Reviewer #2

(Remarks to the Author)
The authors have thoroughly revised and clarified their manuscript, and I am thankful for several explanations that helped me understand their analysis.

On balance, I remain more sceptical than the authors regarding their conclusions. I believe network meta-analyses are superior to investigations like the present one, and one reason is that, in RCTs, the validity of diagnoses and indications is less questionable than in epidemiological data.

Here are some remaining points that should be addressed and possibly added to limitations:

1. As a sensitivity analysis: Is the ranking the same in both countries? If not, what are the differences?

2. To me, a 30-day period seems to be too short to sort out acute from maintenance treatment. At least in my environment, hospitalizations for manic and even more so for depressive episodes can easily last longer than one month. Are there any data on average lengths of stay in Sweden and Finland?
3. The authors emphasize that Swedish and Finnish registers are considered reliable, and this may very well be the case. And yes, at a population level, many people diagnosed with bipolar disorder are doing well w/o the need for inpatient treatment. Still, it may surprise many readers that across a follow-up of 9 years, more than 60% of patients in Sweden and more than 50% in Finland were not hospitalized at all, whereas those admitted had an average number of hospitalizations of 72 and 52, respectively. Possibly, I did not understand the numbers presented, but since other readers may have the same difficulties, the authors need to explain these somewhat odd figures. After all, hospitalization is the endpoint of the present analysis.
4. What I keep finding problematic is the low number of patients receiving lithium at all, making this group hard to assess. Given their modest outcome, it is unlikely they are excellent lithium responders. Can it be that they are patients particularly challenging to treat, and thus the deck is stacked against lithium?
5. In Introduction, the authors mention that about 20% of lithium patients relapse. Can they show the relapse rate of lithium patients in their samples and put these figures into perspective?

Version 2:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
All final points have been sufficiently addressed. The conclusions are now balanced and not overreaching. Thank you

Reviewer #2

(Remarks to the Author)
The authors have done a lot to improve their manuscript, and I have nothing to add.

EDITOR:

Comment 1.01: Please fill out the STROBE checklist providing text excerpts from the manuscript for each point.

Response 1.01: As requested, we have now filled the STROBE checklist accordingly.

Comment 1.02: Please add the flowchart as the main figure reporting numbers of individuals at each stage of study – i.e. numbers potentially eligible, examined for eligibility, confirmed eligible, included in the study, completing follow-up, and analyzed.

Response 1.02: The present study was a register-based study that directly identified eligible individuals. We included all individuals who had bipolar diagnosis at the age of 16-65 years without preceding schizophrenia spectrum disorder (i.e., ICD-10: F20-F29). These inclusion criteria determined the cohort and no exclusion criteria were therefore applied.

The formation of the two cohorts is described elsewhere:

Ermis, C. et al. Real-world effectiveness of pharmacological maintenance treatment of bipolar depression: a within-subject analysis in a Swedish nationwide cohort. The Lancet Psychiatry 12, 198–207 (2025).

Lähteenvuo, M., Paljärvi, T., Tanskanen, A., Taipale, H. & Tiihonen, J. Real-world effectiveness of pharmacological treatments for bipolar disorder: register-based national cohort study. The British Journal of Psychiatry 223, 456–464 (2023).

Comment 1.03: To have more information on how representative the study sample is, please include data on study participants' ethnicity in Methods and Table describing demographic data. If ethnicity is not reported, the reason should be stated in Methods. The Methods section should include an explanation of who identified participant ethnicity and the source of the classifications used (e.g. self-report or selection, investigator, database, electronic health record, survey instrument). Specific ethnic categories are preferred over collective terms, when possible. Categories included in groups labeled as “other” should be defined.

Response 1.03: Data regarding ethnicity is not available in the Swedish or Finnish registers.

We now write in the methods section (page 12, row 426): *“Data regarding ethnicity was not available in the registers of the two cohorts.”*

Comment 1.04: Please provide details on which datasets used in the current study were publicly available and which not; for the latter please explain in the Methods section how the access was obtained and provide details on permission to use the data.

Response 1.04: All datasets used in the present study were derived from nationwide administrative health and social registers in Sweden and Finland. These datasets contain sensitive individual-level information

and are not publicly available under national data protection regulations. Access to both the Swedish and Finnish registers requires formal applications, ethical approval, and authorization from the respective national authorities. To clarify this for readers, we have now expanded the Methods section to describe how data access was obtained.

Added text in the Methods section (page 12, row 416-):
"None of the Swedish or Finnish cohorts used in this study is publicly accessible due to national data protection legislation governing sensitive health information. Ethical approval was obtained from the Regional Ethics Board of Stockholm and the Swedish Ethical Review Authority (2007/762-31; 2024-08708-02) and from the Finnish National Institute for Health and Welfare, the Social Insurance Institution of Finland, the Finnish Centre for Pensions, and Statistics Finland (permissions THL/5279/14.06.00/2023; 31/522/2019; 19023; TK-53-569-19). These permits authorized the use of pseudonymized register data for the present study under secure processing environments."

Comment 1.05: Please add Code Availability sections in the end of the manuscript. If a custom code was created during the analysis, a link to the code should be added in the open repository according to the Nature Mental Health's reporting standards: https://www.nature.com/natmentalhealth/editorial-policies/reporting-standards#availability-of-computer-code.

Response 1.05: The analyses in this study were conducted using standard SAS 9.4 procedures and R packages (meta, metafor) without the development of any custom code or proprietary algorithms. Because no bespoke analytical code was generated, there is no dedicated repository to share. However, the scripts implementing the standard procedures (e.g., model specifications, and meta-analytic routines) can be made available to qualified researchers upon reasonable request, in accordance with Nature Mental Health reporting standards.

To address the reviewer's suggestion, we have added a Code Availability section at the end of the manuscript.

Code Availability: *"No custom computer code was developed for the analyses in this study. All statistical analyses were performed using standard SAS 9.4 procedures and established R packages (meta, metafor). The scripts used to run these conventional procedures are available from the corresponding author upon reasonable request."*

REVIEWER 1

Comment 2.01: his very large population epidemiological study with extensive longitudinal data examined the utlity of antipsychotics, mood stabilisers and combination therein in people with BP and either intolerance or inadequate response with Lithium. Using data from two large nationwide cohorts in Sweden (105,495 patients) and Finland (60,045 patients) with up to 23 years of follow-up, the study evaluates mood stabilizers, antipsychotics, and their combinations in reducing psychiatric hospitalization risk.
Main findings include:
1. Clozapine-based regimens: Clozapine combined with aripiprazole showed the lowest risk of psychiatric hospitalization compared to lithium monotherapy.
2. 'Lithium-intolerant patients': Among lithium discontinuers, combinations such as quetiapine plus lamotrigine and olanzapine plus valproate were associated with reduced hospitalization risk. LAI antipsychotics combined with valproate showed the lowest point estimate but may be limited due to smaller sample sizes.

However the manuscript needs a number of clarifications on premise and interpretation to be interpretable.

The introduction needs breaking down with clear evidence for definition and position of inadequate response: and the evidence gap needed over recent meta-analyses of augmentation, and then intolerance.

The latter is the more novel section and finding but it is merged in and difficult to parse out.

Response 2.01: We thank the reviewer for these thoughtful comments and fully agree that the Introduction required clearer delineation of (1) inadequate response to lithium, (2) lithium intolerance, and (3) the specific evidence gaps that motivate our study. In response, we have reorganized and expanded the Introduction to present these concepts in a more structured and interpretable manner.

Specifically, we have:
1) Added a clearer definition of inadequate lithium response, including the prevalence and clinical implications based on the existing literature.
2) Added a separate paragraph on lithium intolerance, emphasizing its distinct clinical pathway, frequency, and the lack of evidence-based pharmacotherapeutic alternatives.
3) Clarified the limitations of recent meta-analyses, particularly regarding long-term augmentation strategies and combination treatments.
4) Explicitly highlighted the evidence gap that our study addresses, namely, the comparative effectiveness of specific combinations among patients with insufficient response to or discontinuation of lithium.
5) Revised the flow of the Introduction so that the premise and rationale for the two main analyses (inadequate lithium response; lithium intolerance) are clearly separated and easy to follow.

Comment 2.02: Please define inadequate response- how does this differ to treatment resistant BP. Indeed, how were BP with inadequate response defined and identified? There is an assumption from the within-individual analysis that pattern of prescribing is related inadequate response or intolerance. There are many reasons for medication regimes to change. Relapse and rehospitalisation can also be influenced by other confounds not measured.

Response 2.02: Response: We thank the reviewer for raising this important conceptual point. In our study, we do not classify individual patients as having "inadequate response" or "treatment-resistant bipolar disorder". Instead, the concept of inadequate response is operationalized at the analysis level, not at the patient level.

Within-individual models compare each patient's outcomes during periods of lithium monotherapy with outcomes during periods of alternative treatments. Thus, the question is not who inadequate responders are, but rather which regimens provide superior effectiveness over lithium monotherapy in the real-world setting. This differs from treatment-resistant bipolar disorder, which is a clinical diagnosis applied to individuals, typically defined by the failure of multiple adequate therapeutic trials. In contrast, our approach identifies pharmacotherapeutic strategies that outperform lithium monotherapy, thereby informing clinicians about potential alternatives for patients whose symptoms remain unstable or who continue to relapse while taking lithium.

In the Methods section (page 14, row 488-), we describe this analysis: *"A second within-individual analysis compared each alternative pharmacotherapy directly against patients' own periods of lithium monotherapy. The purpose of this analysis was to identify pharmacotherapeutic strategies that could offer superior effectiveness when lithium alone is insufficient."*

Comment 2.03: Please make it clear how the 'intolerance' is defined and used in this study- I believe where >6 months of Li is followed by cessation.

Response 2.03: We thank the reviewer for this comment! We agree that the term "intolerance" may imply a clinical assessment that cannot be directly inferred from register data. For this reason, we have revised the terminology to "lithium discontinuation" throughout the manuscript (including the title), which more accurately reflects how this subgroup is defined in our study.

Operationally, discontinuation subgroup includes individuals who:
1) had ≥6 months of continuous lithium treatment, and 2) subsequently discontinued lithium for ≥30 days.

Definition of the discontinuation group is described in the Methods section (page 14, rows 492-): *"In the third within-individual analysis, comparative effectiveness was assessed among patients who had maintained lithium therapy for ≥6 months and then discontinued it for at least 30 days. During these lithium-free intervals, exposure to antipsychotics, mood stabilizers, and their combinations was contrasted with periods of nonuse of mood stabilizers and antipsychotics, with follow-up censored at death, at the end of the study, or upon reinitiation of lithium."*

Comment 2.04: Indeed, how were BP with inadequate response defined and identified? There is an assumption from the within-individual analysis that pattern of prescribing is related inadequate response or intolerance. There are many reasons for medication regimes to change. Relapse and rehospitalisation can also be influenced by other confounds not measured.

Response 2.04: As noted previously (please see **Response 2.02**), our study does not label individual patients as having inadequate response to lithium. Instead, the concept of inadequate response is inferred analytically, based on whether an alternative regimen is associated with a lower risk of psychiatric hospitalization for the same individual during periods in which the person is not on lithium monotherapy.

The within-individual design is specifically intended to address the concerns highlighted by the reviewer:

Each patient serves as his or her own control, thereby removing all time-invariant confounders (e.g., disorder severity, comorbidities, socioeconomic status, genetic risk). The models were adjusted for time-varying covariates, including sequence of medications, time since cohort entry, and concurrent psychotropic use. Because treatment changes in BD can occur for multiple clinical and non-clinical reasons, the within-individual framework minimizes bias from such factors by comparing only different treatment periods within the same person, rather than across different individuals.

We agree that relapse and rehospitalization can be influenced by unmeasured factors. However, the direction of residual confounding is expected to work against the observed findings: treatment combinations are typically initiated during clinical deterioration or after failure of prior regimens, which would bias estimates toward worse outcomes during combination therapy. Therefore, our finding that some regimens outperform lithium monotherapy likely represents a conservative estimate of their effectiveness.

We have added the following text in the Discussion section (page 8, row 309-):
"Although treatment changes may reflect factors unrelated to lithium response, the within-individual design minimizes confounding by comparing medication periods within the same patient. Any remaining time-varying confounding, such as clinical deterioration leading to treatment escalation, would tend to bias results against intensified or combination regimens, rendering our estimates conservative."

Note also that the sensitivity analysis, where 30 days of exposure periods were removed, aligned with the original analyses (please see page 6 and Supplementary Figure 1).

Comment 2.05: The fact that clozapine is most effective in their outcomes is interesting- and may speak to increased monitoring and early intervention or clozapine unique pharmacological properties.

Response 2.05: We agree with the reviewer. The observed effectiveness of clozapine-centered regimens may reflect both its unique pharmacological properties and the increased clinical monitoring typically accompanying clozapine treatment.

We now acknowledge this possibility in the Discussion (page 8, row 319-): *"Nevertheless, the superiority of clozapine over lithium may also partially stem from regular contact in healthcare via blood tests."*

Comment 2.06: The methods are sound, on the whole, with potentially useful output however the limitations of the data are significant and need greater recognition.

Response 2.06: We agree that the register-based data have several limitations.

In response, we have expanded the Limitations section in the Discussion (pages 10-11, row 379-): *"Furthermore, psychiatric hospitalization, our primary outcome, serves as a proxy for severe relapse rather than a direct measure of symptom recurrence or functional status. Therefore, studies using more granular outcomes in the future are needed. In addition, hospital admissions for any psychiatric diagnosis do not distinguish symptom types, and we lack data to confirm treatment resistance among medication users. Register-based data lack detailed clinical information, such as symptom severity, adherence, and reasons for treatment changes."*

Comment 2.07: The conclusion and recommendations eg support switching to clozapine-based regimens or augmentation with LAI antipsychotics when lithium monotherapy is inadequate are over-reaching for the level of data on prescribing decisions and choices available.

Response 2.07: We appreciate this important comment! We agree that the wording of the Conclusions could overstate the strength of evidence given the observational nature of the data and the lack of direct information on prescribing decisions. In response, we have revised the concluding statements to be more cautious and now frame the findings as options that clinicians may consider rather than firm recommendations.

We have now reworded the conclusions more cautiously (page 11, page 398-):
"These data suggest that clinicians may consider clozapine-based regimens or augmentation with long-acting injectable antipsychotics when lithium monotherapy does not provide sufficient stabilization. Among patients who discontinue lithium, specific antipsychotic–mood stabilizer combinations may also represent viable alternatives. Prospective trials are needed to confirm our findings."

Comment 2.08: They do suggest that further trials in lithium resistant BP may indeed investigate LAI and clozapine regimes- and this is a sensible conclusion

Response 2.08: We agree and have retained this point in the Conclusions.

REVIEWER 2

Comment 3.01: Lieslehto and seven co-authors (last author: H. Taipale) present an analysis of Swedish and Finish register data on the treatment of patients diagnosed with bipolar disorder. Their main interest is in patients who have not responded to lithium or have not tolerated it but in their study they also show results on the effectiveness of lithium. The authors used data from almost 170.000 patients who have been hospitalized more than 60.000 times. Their endpoint is all-cause psychiatric hospitalization, and they find that combinations of mood stabilizers (or aripiprazole) with either clozapine or long-acting antipsychotics are associated with a favorable relapse risk. At the same time, most combinations of two antipsychotics were deemed unfavorable.

The topic certainly warrants attention, and I found the general approach of the study instructive, but the manuscript is a very tough read, mainly due to its several layers of analysis.

Response 3.01: We thank the reviewer for this constructive comment! We recognize that the multiple analytic layers can make the manuscript challenging to follow. In response, we have revised the Introduction to provide a clearer conceptual structure (see **Response 2.01**) and improved readability throughout the manuscript by reorganizing key sections, simplifying transitions, and clarifying the rationale for each analytical step. We hope these revisions make the overall presentation more accessible.

Comment 3.02: It seems odd that more than 170.000 patients should have been hospitalized for only 60.000 times. Most patients with bipolar disorder I know have been admitted several times, almost always when they were manic, but also frequently when they were depressed. Can it be that admissions are missing from the data? Are the diagnoses correct? I am sure many readers will have difficulties understanding these figures, and authors must explain this clinical-experience-data mismatch.

Response 3.02: We thank the reviewer for raising this important point! The sentence was slightly misphrased. The total number of hospitalizations over the total follow-up was several times higher than 60,000.

We have rephrased the following sentence (page 5, row 225-): *"Over the total available follow-up, 36,732 patients had psychiatric hospitalizations in the Swedish cohort (2,634,928 psychiatric hospitalizations in total) and 26,464 in the Finnish cohort (1,381,684 in total)."*

Note that the sample represents a very heterogeneous population. While many clinicians primarily encounter individuals with frequent relapses, nationwide registers also include large numbers of patients with mild, infrequent, or well-controlled illness, some of whom may never require hospitalization after diagnosis.

The Swedish and Finnish registers used here are generally considered reliable and have been extensively validated and are known to reliably capture all inpatient psychiatric admissions (e.g., Lähteenvuo et al., BJP 2023; Ermis et al., Lancet Psychiatry 2025).

Comment 3.03: I would strongly argue for restricting the analysis to patients who have not switched diagnoses to unipolar depression, ADHD, or schizophrenia/schizoaffective disorder because mood stabilizers are largely ineffective in patients with these diagnoses, except lamotrigine in patients with unipolar depression. Also, lithium, which is the starting point of the present analysis, is the gold standard

in the long-term treatment of bipolar disorder, but nowhere else. This approach will reduce sample size and limit statistical significance but will greatly improve the interpretability of the data.

Response 3.03: We thank the reviewer for this comment! We agree that diagnostic comorbidity can complicate interpretation; however, restricting the cohort to individuals who never receive other psychiatric diagnoses would not reflect real-world clinical practice, where comorbidity is common rather than exceptional (as shown in Table 1). Thus, excluding these patients would induce selection bias and thereby substantially limit the generalizability of our findings.

Conversely, we agree that excluding individuals who later convert to schizophrenia is an appropriate sensitivity analysis, given that several antipsychotics examined (e.g., clozapine) are predominantly used for schizophrenia. These results were highly consistent with our main findings (Supplementary Figure 4).

Importantly, we analyzed the data using within-individual analysis that compares each individual to themselves, which eliminates the selection bias (e.g., between individuals with comorbidities vs. without comorbidities). Please see **Response 2.04.**

Comment 3.04: For clarification: In an additional figure, the authors could present examplary illness/treatment courses of patients to illustrate the different approaches they used, that is, within-patient as well as between-patients comparisons. Such a chart could show the timepoint of index diagnosis, treatment changes, durations of treatment and, most importantly, which phases were compared in the various analyses.

Response 3.04: We thank the reviewer for this helpful suggestion. In response, we have added a new figure (Supplementary Figure 8, referred on page 13 [row 451-]) that visually illustrates both the within-individual and between-individual analytical frameworks, including example treatment timelines, periods of exposure and non-exposure, and how comparison windows were constructed. We hope this clarifies our analytic approaches for readers.

Comment 3.05: The authors keep calling what they find in their registers real world data. This, however, is a difficult term because it pretends that the data truly represent the whole picture. However, as the authors themselves point out, they had to resort to hospitalization as an endpoint just because the more informative endpoint of any mood disorder episode was not available in this register, let alone other key outcomes, such as interepisodic psychopathology or tolerability. Also, RCT data originate from the real world, too, and are even more important than register data. It is precisely for this reason that the pharmaceutical industry very often uses the term real world data when results from RCTs are at variance with their interests. As a result, the term real world data should be avoided because of its imprecision, its ambiguity and because it has become a loaded term.

Response 3.05: We thank the reviewer for this thoughtful comment! RCTs and register-based cohorts offer complementary perspectives. RCTs provide evidence whether a given treatment is efficious but these studies are highly selective reagarding exclusion criteria (e.g., often individuals with comorbidities or suicidal behavior are excluded). Note, however, that RCTs monitor adherence to medications very closely and offer medications free of charge and thus, represent more ideal circumstances than every day clinical practice. Nationwide register studies on the other hand cannot directly assess efficacy but they capture the broad clinical heterogeneity of routine clinical practice. We agree that the term ‘real-world data’ is imprecise and carries ambiguous connotations. To avoid misunderstanding, we now use the more neutral terms ‘nationwide register-based data’ or ‘population-based data’ throughout the manuscript.

Comment 3.06: The title suggests that among the patients under study lithium has failed, when in fact it only reflects the authors opinion that lithium is the gold standard and should be used first (a well founded opinion, in my view). However, lithium failure is not a requirement for the main analysis presented in this paper. As a matter of fact, it is not the requirement for the other analyses either, since it is impossible to know from these real world data why lithium was discontinued. To manage expectations among readers, I would rephrase the title and change the introduction.

Response 3.06: We appreciate the reviewer's thoughtful observation. We agree that the original title could imply that all included patients had experienced lithium failure, which is not accurate given that our register-based data cannot determine the clinical reasons for treatment changes or discontinuation. As noted in our earlier responses (please see **Responses 2.02-2.04**), we have revised the terminology throughout the manuscript to use the more neutral term "lithium discontinuation" rather than "intolerance" or "failure."

To align the title with the study design, we have rephrased the title to remove any suggestion of lithium failure:
"Comparative Effectiveness of Treatment Strategies for Bipolar Disorder ***During and After Lithium Treatment****: Analysis of Two Nationwide Cohorts"*

Comment 3.07: As an analysis of the results' robustness and generalizability: Is the ranking similar in Sweden and Finland? It is awkward for readers to go through the supplement tables themselves. However, figures vary considerably: For example, per supplements 1 & 2, in the Swedish cohort, lithium arrives at an adjusted HR of 0.83, in Finland, however, the respective value is 0.70. The clozapine plus aripiprazole combination reaches 0.26 in Finland and 0.61 in Sweden. These differences are substantial, even though HRs seem to be generally higher in Sweden (which differences may be responsible?). This is why it is so important to show the ranking.

Response 3.07: Indeed, point estimates differ slightly between Sweden and Finland, as expected given differences in prescribing patterns, population characteristics, and sample sizes for specific regimens (particularly those with a low number of events). This is precisely why we applied a random-effects meta-analysis to obtain a pooled estimate across cohorts. The main results of the present study are the meta-analysis-based results, which capture the "overall picture" across the two cohorts. Individual estimates from each cohort provided in the Supplementary Material are intended for interested readers who, e.g., might want to combine them with additional cohorts.

Comment 3.08: In order not to bias the calculation against the index mood stabilizing drug, the authors, in a sensitivity analysis, removed the first 30 days from the analysis. Given that lithium usually needs months to take hold this period may be too short. For example, we advise our patients that they should take into acocunt at least one year before considering going off lithium salts.

Response 3.08: We thank the reviewer for this comment! The 30-day omission was not intended to represent the expected onset of lithium's therapeutic effect. Rather, this sensitivity analysis was performed to reduce protopathic bias, which occurs because medications are often initiated during acute clinical deterioration (i.e., before the drug has any chance to exert benefit). Excluding the first 30 days helps ensure that early relapse events occurring immediately after treatment initiation do not artificially inflate the hazard associated with a given medication. Also, this omission analysis excluded the first 30 days following medication discontinuation, thereby accounting for potential carry-over effects, as treatment effects do not cease immediately upon stopping a given medication.

Importantly, this omission analysis produced results highly consistent with the main findings, indicating that our conclusions are not driven by early crisis-related symptom aggravation (please see Supplementary Figure 1).

Description of the rational of 30-day omission danalysis is provided in the Methods section (page 14, row 469-): *"We first re-ran the within-individual models after excluding the first 30 days of each exposure (and nonexposure) period, thereby reducing the risk of protopathic bias resulting from treatment initiation during acute deterioration as well as potential carry-over effects of previous medication use to subsequent nonuse periods."*

Comment 3.09: The authors repeatedly refer to the harmonized nature of their data. Do they mean that they fitted Swedish and Finish register entries so that they could be used in one approach? Please clarify.

Response 3.09: With "harmonization", we mean that the inclusion and exclusion criteria, variable definitions, medication exposure modelling (PRE2DUP), and analytic procedures were made consistent across the Swedish and Finnish datasets so that identical analyses (and meta-analysis) could be performed in both cohorts. The underlying registers differ in structure, but all relevant elements (diagnosis codes, medication classifications, outcome definitions, and covariates) were aligned to ensure comparability. Note, also that Sweden and Finland have relatively similar publically funded healthcare.

Comment 3.10: I did not understand the start (when did the data start?) and stop (when was the last data point?) dates of this analysis, and I would have welcomed these figures as early as in the summary.

Response 3.10: As noted in the Methods section (page 12), each cohort has well-defined observation windows (Sweden: July 1, 2006–December 31, 2021; Finland: January 1, 1996–December 31, 2018), and individuals contribute follow-up time differently depending on their treatment trajectories. We did not set a minimum follow-up as an inclusion criterion. Note, however, that individuals with short follow-up and no exposure vs. non-exposure periods do not contribute to the within-individual analyses.

To improve transparency, we have added the total person-years of exposure for each medication category (Supplementary Tables 1-2), which hopefully helps readers better understand the temporal structure of the dataset.

Due to the word count limitations (max 150 words), we could not fit the exact dates in the abstract.

Comment 3.11: Similarly, I would have welcomed the average time of follow up, not the upper end („up to 23 years").

Response 3.11: The duration of follow-up from each cohort is provided in the Results section (page 5, row 222-): *"We identified 105,495 patients with BD (65,607 women, mean [SD] age=44.23 [18.79] years, lithium users=27,058, **mean [SD] follow-up=9.09 [5.15] years**) from Sweden and 60,045 individuals (33,859 women; mean [SD], age=41.7 [15.8] years, lithium users=7,700, **mean [SD] follow-up=9.45 [6.01] years**) from Finland."*

In the abstract, we have rephrase the following sentence: *"Using harmonized nationwide cohorts from Sweden (n = 105,495) and Finland (n = 60,045) **with an average of nine years of follow-up**, we assessed the comparative effectiveness of mood stabilizers, antipsychotics, and their combinations in preventing psychiatric hospitalization."*

Comment 3.12: How were lithium user defined? Anyone who was ever prescribed lithium? The same applies to the other drugs and combintions: any minimum duration of number of prescriptions/refills? Please clarify.

Responsee 3.12: Lithium users were defined based on modeled periods of lithium treatment (dispensed prescriptions). As with all medications in this study, exposure periods were reconstructed using the validated PRE2DUP method, which estimates continuous drug use from dispensing data by accounting for dose, refill patterns, stockpiling, and inpatient days. Thus, a patient was considered a lithium user only during intervals in which the PRE2DUP model indicated lithium treatment.

Comment 3.13: It is interesting that even though the authors reiterate that lithium is the gold standard only relatively few patients received lithium (line 100). How is that?

Response 3.13: This is an important and worrysome observation! The relatively low proportion of lithium users in our cohorts reflects a well-documented decline in lithium prescribing over the past two decades, despite its strong evidence base. This trend has been reported in several studies, including Poranen et al. (2022), and might be attributed to concerns about side effects, monitoring requirements, and increasing use of alternative agents.

We have acknowledged this in the Introduction section (page 3, row 55-): *"Additionally, lithium prescription rates have fallen significantly in recent years."*

Comment 3.14: Can the authors differentiate between acute treatment and long-term treatment? For example, a patient in an initial manic phase may have receieved lithium (or olanzapine, to name only one antipsychotic) as an acute treatment, may have been admitted, anyway, but was then maintained on lithium (olanzapine) after his/her succesful treatment in the hospital. Would this count as an hospitalization under lithium (olanzapine)? If so, I the numbers are hard to interpret.

Response 3.14: We thank the reviewer for raising this important point. Unfortunately, nationwide register cohorts do not allow us to distinguish reliably between acute treatment initiation and long-term maintenance treatment as clinical intentions were not available. As a result, an admission occurring shortly after treatment initiation could fall within a modeled exposure window.

To mitigate this issue, we used two strategies:
1) Within-individual analysis, which compares different treatment periods within the same patient and therefore reduces confounding from illness severity or prior treatment history (the effect of time is controlled in the models), 2) A 30-day omission sensitivity analysis, where the first 30 days of each treatment period were excluded to reduce protopathic bias (i.e., the tendency for medications to be initiated during acute deterioration before they have time to exert benefit). The results of the 30-day omission analysis were consistent with the main findings, suggesting that our conclusions are not driven by early hospitalizations immediately following acute treatment initiation (please see Supplementary Figure 1).

Overall, our findings apply to long term outcomes rather than short term drug effects, as noted in the Discussion section (page 11): *"Additionally, our findings pertain to long-term outcomes rather than acute treatment effects…"*

Comment 3.15: Can the authors contribute results to the order of pharmcological treatment? When were the various drugs applied (first drug, second…)? This may help interpreting the relative success/failure of the treatment.

Response 3.15: We appreciate the reviewer's interest in treatment sequencing. Unfortunately, given the size of the cohort and the substantial heterogeneity in individual treatment trajectories, it is not feasible to summarize meaningful "first–second–third line" patterns across nearly 170,000 patients from two countries. Treatment order varies widely, reflecting patient-specific clinical histories, comorbidities, prescriber preferences, and changes over decades of practice.

Importantly, treatment sequence is incorporated into the within-individual models, where we adjust for time since cohort entry and account for the chronological ordering of medication periods for each patient.

Comment 3.16. Regarding the status „non treatment": not treated but diagnosed? How many patients were in that group? Or, in within-subjects analyses, how long was this period? Is this a reasonable comparison?

Response 3.16: Thank you for raising this point. In our analyses, "non-treatment" refers to periods during which a patient had a bipolar disorder diagnosis but was not using any mood stabilizer or antipsychotic according to the PRE2DUP exposure modeling. These untreated intervals occur naturally in longitudinal register data, for example, due to medication discontinuation, gaps between prescriptions, patient non-adherence, or clinical decisions to pause a given regimen.

In the within-individual framework, the duration of non-treatment periods varies across patients and is modeled precisely based on dispensing records via PRE2DUP. This comparison is methodologically appropriate because each patient serves as their own control, allowing us to estimate how relapse risk changes when the same individual is on a specific medication versus when they are not on any mood stabilizer or antipsychotic.

We also note that in a separate analysis, we compared all treatment strategies directly against lithium monotherapy.

Comment 3.17: There is an interesting meta-analysis supporting clozapine in the field of schizophrenia that the authors may want to mention: PMID: 40535000

Response 3.17: We thank the reviewer for the suggestion! However, given the reviewer's previous observation that the manuscript is already dense, we opted to maintain the focus on bipolar disorder, as expanding into additional diagnostic groups would further stray the narrative.

For this reason, although an excellent read in the field of schizophrenia, we have refrained from discussing its findings in the manuscript.

Comment 3.18: Was there a minimum observation period? Otherwise, people may appear incredibly ill (if included in their index episode briefly before their inclusion ended).

Response 3.18: Thank you for raising this point. We did not impose a minimum observation period at cohort entry, because the within-individual design inherently accounts for differences in follow-up length: each patient contributes only the time during which they are under observation, and comparisons are made within the same individual across different treatment periods. Thus, short initial observation windows do not bias the results toward appearing more ill or more relapse-prone. Importantly, if a person has a very short follow-up without medication changes, he/she does not contribute to the analysis.

To improve transparency, we now report the total person-years contributed for each medication category (Supplementary Tables 1-2), which reflects the underlying exposure time and ensures that readers can assess the robustness of each estimate.

REVIEWER 1:

Comment 1.01: The authors provide a detailed rebuttal and have made several substantial changes that improve clarity. However they continue to overstate the conclusions and clinical inferences that can be drawn. There are two points in the discussion that should be re-written in my opinion:

1. "Our results highlight clozapine-centered regimens as potentially superior alternative over lithium monotherapy, surpassing the scant evidence currently captured in guidelines."

The current guidelines are scant on clozapine recommendations because the RCT's have not been completed. At most, the authors can suggest that these trials need to be completed, in patients who are not able to tolerate or do not respond to lithium. In fact, BAP guidelines do recommend clozapine in refractory BAP, recognising the level of evidence is not strong.

Response 1.01: We thank the reviewer for this comment! We have now downplayed the clinical implications of our results given the observational nature of our work.

Accordingly, we have now rephrased this sentence (page 8): *"Our results highlight the importance of conducting RCTs on clozapine to assess its efficacy in bipolar disorder, given the scant evidence on its use currently captured in guidelines."*

Comment 1.02: "These data suggest that clinicians may consider clozapine-based regimens or augmentation with LAI
antipsychotics when lithium monotherapy does not provide sufficient stabilization.' Among patients who discontinue lithium, specific antipsychotic–mood stabilizer combinations may also represent viable alternatives. Prospective trials are needed to confirm our findings"

Again the evidence is not strong enough in this paper to make clinical recommendations, where discontinuation is assumed to equate to intolerance or lack of efficacy. Clozapine is not liscenced for BP, at most the authors again can recommend head to head trials are completed.

The strengths and weaknesses section of the discussion needs to be expanded to greater reflect the level of uncertainty and assumptions in population data.

With these corrections, the manuscript is acceptable for publication, in my opinion. I will leave to the editorial team to decide if these results when accompanied by a balanced and appropriate discusison are novel enough for publication in NMH

Response 1.02: We have now downplayed the clinical implications of our results.

We now write more cautiously in the conclusion section (page 11): *"We found that clozapine-based regimens and lithium accompanied by LAI antipsychotics demonstrated a lower risk of psychiatric hospitalization compared to lithium monotherapy. Among patients who discontinued lithium, only specific antipsychotic–mood stabilizer combinations were related to reduced psychiatric hospitalization risk. Prospective trials are needed to decipher the efficacy of these regimens."*

Also, we have reformatted the following section in the strengths and weaknesses section (page 10): *"However, residual confounding by protopathic bias cannot be ruled out. Indeed, since pharmacotherapy is often initiated or intensified at times of symptom exacerbation, indication severity (rather than the indication per se) can introduce confounding factors that make combination regimens appear less effective than they are. Thus, the observed associations between specific combination regimens and lower hospitalization risk compared with lithium monotherapy should be interpreted cautiously, as residual bias related to treatment escalation during symptom exacerbation cannot be excluded."*

REVIEWER 2

Comment 2.01: The authors have thoroughly revised and clarified their manuscript, and I am thankful for several explanations that helped me understand their analysis.

On balance, I remain more sceptical than the authors regarding their conclusions. I believe network meta-analyses are superior to investigations like the present one, and one reason is that, in RCTs, the validity of diagnoses and indications is less questionable than in epidemiological data.

Response 2.01: We thank the reviewer for these comments. We highly agree that RCTs are required to demonstrate the efficacy of a given medication regimen. However, RCTs are highly selective with their stringent inclusion criteria and often have short follow-up durations. In this light, nationwide observational studies that include all patients with bipolar disorder and with decades of follow-up are crucial to supplement the results from RCTs. Also, observational studies such as ours can help guide the development of new RCTs to confirm their findings (hypothesis generation studies).

We have now rephrased the conclusion section to be more cautious (page 11): *"We found that clozapine-based regimens and lithium accompanied by LAI antipsychotics demonstrated a lower risk of psychiatric hospitalization compared to lithium monotherapy. Among patients who discontinued lithium, only specific antipsychotic–mood stabilizer combinations were related to reduced psychiatric hospitalization risk. Prospective trials are needed to decipher the efficacy of these regimens."*

Comment 2.02: 1. As a sensitivity analysis: Is the ranking the same in both countries? If not, what are the differences?

Response 2.02: While the primary results of the study are based on random-effects meta-analyses across the two cohorts, we agree that assessing the consistency of treatment rankings between Sweden and Finland might be informative to some readers. Therefore, as a sensitivity analysis, we examined whether the ranking of adjusted hazard ratio point estimates for medications and their combinations was similar across the two cohorts. For this purpose, we calculated Spearman's rank correlation coefficient using the cohort-specific point estimates. We observed a strong correlation between the rankings in Sweden and Finland (Spearman's rho = 0.79), indicating substantial alignment in the relative ordering of treatment effectiveness despite differences in absolute effect sizes.

We now write (page 5): *"Across the two cohorts, the rank ordering of treatment-specific aHR estimates was highly aligned (Spearman's rho = 0.79)."*

Comment 2.03: To me, a 30-day period seems to be too short to sort out acute from maintenance treatment. At least in my environment, hospitalizations for manic and even more so for depressive episodes can easily last longer than one month. Are there any data on average lengths of stay in Sweden and Finland?

Response 2.03: We did not try to distinguish between acute and maintenance treatment in our analyses. The primary outcome (i.e., psychiatric hospitalization) represents a long-term rather than an acute outcome. The 30-day omission analysis was employed to reduce protopathic bias. Protopathic bias refers to the tendency to initiate medications during acute symptom exacerbations, and since it will take time before they have time to exert benefit, this bias might result in too pessimistic statistical estimates. By removing the acute phase (i.e., 30 days after the start of treatment or non-treatment periods), we were able to reduce the effect of protopathic bias on our results.

In the within-individual analyses, medication exposure is modelled based on dispensing data, and inpatient treatment periods themselves are not explicitly modelled. Consequently, the exact duration of hospital stays does not directly affect the estimation of treatment effects in this framework. We therefore did not use the 30-day window as a proxy for episode duration, but rather as a sensitivity analysis to minimize bias related to treatment initiation during acute deterioration (protopathic bias).

Within-individual modelling is described on page 13: "The outcomes in the models were modeled as recurring occurrences. The follow-up time was reset to zero after each outcome (i.e., psychiatric hospitalization), and the models were adjusted for the sequence of treatments, time since cohort entry, and time-varying use of other psychotropics (antidepressants, benzodiazepines and related agents, ADHD medications, and medications for substance use disorders)."

We also provide a graphical description of the analyses of our work in Supplementary Figure 9.

Comment 2.04: The authors emphasize that Swedish and Finnish registers are considered reliable, and this may very well be the case. And yes, at a population level, many people diagnosed with bipolar disorder are doing well w/o the need for inpatient treatment. Still, it may surprise many readers that across a follow-up of 9 years, more than 60% of patients in Sweden and more than 50% in Finland were not hospitalized at all, whereas those admitted had an average number of hospitalizations of 72 and 52, respectively. Possibly, I did not understand the numbers presented, but since other readers may have the same difficulties, the authors need to explain these somewhat odd figures. After all, hospitalization is the endpoint of the present analysis.

Response 2.04: We thank the reviewer for highlighting these points! Unfortunately, we've noticed that there was an error in the total number of hospitalizations. Correct numbers are now presented on page 5: *"Over the total available follow-up, 36,732 patients had psychiatric hospitalizations in the Swedish cohort (119,059 psychiatric hospitalizations in total) and 26,464 in the Finnish cohort (93,922 in total). "*

Note, however, that the reported averages were calculated across highly heterogeneous follow-up times, ranging from very short observation periods to several decades. Consequently, individuals with long follow-up and frequent admissions contribute disproportionately to the total number of hospitalizations, which can inflate mean values when aggregated across the entire cohort.

To provide a clearer and more interpretable description of hospitalization risk across the follow-up, we have now plotted the cumulative incidence of first psychiatric hospitalization after cohort entry in both countries (Supplementary Figure 1). This figure depicts the proportion of patients experiencing at least

one severe relapse (i.e., psychiatric hospitalization) during follow-up and is not influenced by extreme values among patients with long observation times as censoring due to loss of follow-up is considered.

We have added these cumulative incidence figures to the Results section to improve clarity (page 5): *"Cumulative incidence of psychiatric hospitalizations over the follow-up across the two cohorts is depicted in Supplementary Figure 1."*

Finally, we want to emphasize that psychiatric hospitalization represents a relatively severe outcome, and many patients with milder symptom exacerbations are treated in outpatient settings and therefore do not contribute to the hospitalization endpoint.

Comment 2.05: What I keep finding problematic is the low number of patients receiving lithium at all, making this group hard to assess. Given their modest outcome, it is unlikely they are excellent lithium responders. Can it be that they are patients particularly challenging to treat, and thus the deck is stacked against lithium?

Response 2.05: The number of lithium users has indeed declined in recent years (please see Poranen et al., 2022, Acta Psychiatr Scand). We agree with the reviewer that the relatively small proportion of lithium users raises the possibility that these individuals may represent a clinically more challenging subgroup. Nevertheless, such an indication bias is likely to shift statistical estimates against lithium. Despite this, however, lithium monotherapy was the most effective treatment among mood stabilizer monotherapies (Figure 1).

In the present study, we primarily used within-individual analyses, in which each individual is compared with themselves across different treatment periods. This approach mitigates selection bias related to stable patient characteristics, including illness severity and genetics. A limitation of the within-individual analysis is that individuals must have outcome events and changes in medication patterns over their follow-up time to contribute to the results (i.e., individuals with constant use of a given treatment and no outcome events do not contribute to the within-individual results). To test whether the results "hold" across the whole cohort, we conducted complementary between-individual analyses, adjusted for measured covariates (detailed on page 14). Importantly, the results of these two analytic approaches were highly consistent (Supplementary Figures 3–4), suggesting that our findings are not driven solely by between-patient differences in case mix.

We now write on page 14: "Next, we performed traditional between-individual Cox regression modeling, adjusted for age, sex, … and ADHD drugs (N06BA), to assess whether our findings are driven solely by between-patient differences in case mix."

Comment 2.06: In Introduction, the authors mention that about 20% of lithium patients relapse. Can they show the relapse rate of lithium patients in their samples and put these figures into perspective?

Response 2.06: We thank the reviewer for this important question! In the present study, we are unable to report a single "relapse rate" for lithium-treated patients that would be directly comparable to relapse rates reported in randomized controlled trials. This is because our primary analyses are based on a within-individual design, in which individuals contribute person-time to multiple treatment states over follow-up rather than belonging to a fixed "lithium group."

In this framework, patients switch treatments or experience periods without mood stabilizers or antipsychotics, and individuals who remain continuously on a single regimen without switching do not

contribute to the within-individual comparisons (see Supplementary Figure 9 for more details). As a result, lithium exposure is modeled as time-varying, and outcomes are evaluated relative to other treatment periods within the same individual, rather than as cumulative relapse proportions within a predefined lithium cohort.

REVIEWER 1:

Comment 1.01: All final points have been sufficiently addressed. The conclusions are now balanced and not overreaching. Thank you

Response 1.01: We thank the reviewer for this comment!

REVIEWER 2

Comment 2.01: The authors have done a lot to improve their manuscript, and I have nothing to add.

Response 2.01: We thank the reviewer for these comments.