Validation Between Antidepressant Treatment Failure Proxy and PHQ-9 Score in Major Depressive Disorder Using A Linked Insurance Claims and Electronic Health Record Dataset

EPH277

R. Patel[1], C. White[2], S. St. Rose[3], A. Kilburg[3], S. D. Suessmuth[4], F. de Crescenzo[3] and L. Zhang[2]

[1]Department of Psychiatry, University of Cambridge, Cambridge, UK; [2]Boehringer Ingelheim Pharmaceuticals, Inc., Ridgefield, CT, USA; [3]Boehringer Ingelheim International GmbH, Ingelheim am Rhein, Germany; [4]Boehringer Ingelheim Pharma GmbH & Co. KG, Biberach an der Riss, Germany

Introduction

1 Context:

Major depressive disorder (MDD) is one of the most prevalent mental disorders in the United States, significantly impacting individuals' overall health and well-being.[1]

Although various treatment options are available, many patients continue to struggle with persistent symptoms.[2,3]

2 Unmet need:

Evaluating antidepressant response in patients with MDD using real-world data is challenging due to limited availability of structured symptom measures and variability in symptom presentation.[3,4]

3 Study rationale:

Given the high prevalence and persistent burden of MDD, there is a critical need to better understand the factors contributing to treatment non-response. By evaluating potential proxies of treatment non-response that are widely recorded in real-world data, it may be possible to discover a practical approach to detect patients who may benefit from alternative therapeutic approaches.

Aim

This study aims to assess if antidepressant treatment failure can serve as a proxy for treatment non-response using the Patient Health Questionnaire (PHQ)-9 in patients with MDD.

Trial design

STUDY PERIOD January 1, 2012 to June 30, 2022

Adult patients with MDD (N=2,218) newly diagnosed with ≥1 prescription of antidepressants.

Optum Market Clarity database

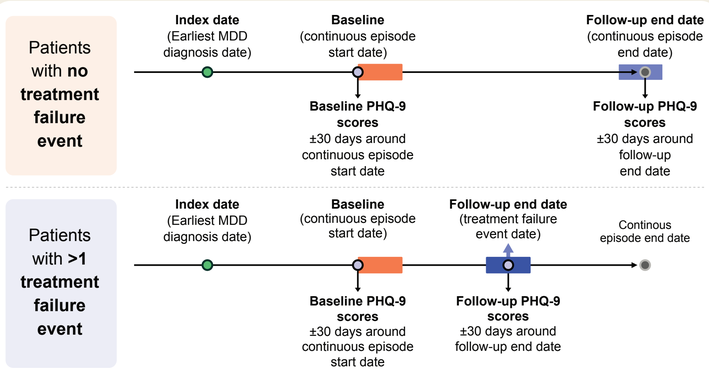
Treatment failure definition:

- **Treatment switch** occurring within 1 to 6 months during a continuous treatment episode.
- **Treatment augmentation** within 1 to 8 months during a continuous treatment episode.

A treatment episode was considered to be continuous if any gaps between treatment lines were less than 120 days apart.

For a detailed explanation of treatment pattern measures and failure definitions throughout the study visit posters **EPH138** and **EE734**.

PHQ-9 COLLECTION DIAGRAM



MODEL ANALYSIS

Regression models

Negative binomial regression estimated differences in **PHQ-9 follow-up scores** between treatment failure and non-failure groups, adjusting for baseline disease characteristics, with log of follow-up time as offset.

Linear regression assessed differences in **PHQ-9 follow-up score change from baseline** between treatment failure and non-failure groups using similar adjustments.

Validation analyses

A validation analysis was conducted to assess the relationship between failure events defined by **PHQ-9 score changes** (yes/no) and **proxy failure events** based on treatment patterns (yes/no).

PHQ-9 score change failure was defined as 'Yes' if the follow up PHQ-9 score increased or failed to decrease by 50% from baseline; otherwise, it was defined as 'No'.

To correct imbalance size between proxy failure vs. proxy non-failure group, **weighted outcomes** were calculated: **weight = number of patients having proxy failure / number of patients having non-proxy failure**.

Logistic regression analyses between PHQ-9 score failure and treatment pattern proxy failure were conducted on the overall population and across **defined MDD subsets**:

Defined MDD subset	Patients with **any baseline PHQ-9 score,** excluding those with no change	**Patients with a baseline PHQ-9 score of ≥5**, excluding those with no change	**Patients with a baseline PHQ-9 score of ≥10**, excluding those with no change	**Patients with a baseline PHQ-9 score of ≥15**, excluding those with no change
Terminology used	**No0**	**Mild MDD**	**Moderate MDD**	**Moderately severe MDD**

Since score failure and proxy failure occurred within similar timeframes, **reciprocal logistic regression** models were applied–alternating each as the dependent outcome. Both crude and adjusted model were used, incorporating **key covariates**: demographics, baseline commodities, medication, psychotherapy treatment and MDD severity status.

Key Conclusions

- Baseline PHQ-9 severity and treatment failure status were the strongest predictors of non-response.
- Patients with treatment failure, defined by a switch or augmentation within 1–8 months, showed less improvement in PHQ-9 scores compared to non-failure patients, supporting its validity as a proxy for treatment effectiveness.
- The findings suggest that in the absence of data on treatment response, treatment failure may be partly correlated with non-response but also explained by other factors.

Results

When analyzing **patient characteristics**, 78.5% of patients with MDD experienced treatment failure. Of those, a higher proportion were Black and Asian patients (**Table 1**).

Table 1. Demographic characteristics at baseline in patients with MDD with and without treatment failure

	No treatment failure (n=477)	Treatment failure (n=1,741)	*P*-value
Age			1.00
Mean (SD)	42.0 (13.90)	42.2 (13.61)	
Median (IQR)	42.00 (30.00, 55.00)	43.00 (30.00, 54.00)	
Sex (%)			1.00
Male	119 (24.9%)	431 (24.8%)	
Race (%)			0.015
Asian	9 (1.9%)	8 (0.5%)	
Black	73 (15.3%)	205 (11.8%)	
Caucasian	374 (78.4%)	1,430 (82.1%)	
Missing	21 (4.4%)	98 (5.6%)	
Region (%)			0.014
Midwest	247 (51.8%)	996 (57.2%)	
Northeast	65 (13.6%)	152 (8.7%)	
South	103 (21.6%)	340 (19.5%)	
West	27 (5.7%)	152 (8.7%)	
Unknown	35 (7.3%)	101 (5.8%)	
Type of Health Insurance (%)			0.613
Commercial	328 (68.8%)	1,138 (65.4%)	
Medicaid	78 (16.4%)	257 (14.8%)	
Medicare	26 (5.5%)	145 (8.3%)	
Other	11 (2.3%)	30 (1.7%)	
Unspecified	27 (5.7%)	151 (8.7%)	
Uninsured	7 (1.5%)	20 (1.1%)	

Some differences were observed in **baseline psychiatric comorbidities** and psychiatric **medication intake** between patients with treatment failure and those without (**Table 2**).

Table 2. Clinical characteristics and HCRU at baseline in patients with MDD with and without treatment failure

	No treatment failure (%) (n=477)	Treatment failure (%) (n=1,741)	*P*-value
Psychiatric diagnosis			
MDD	389 (81.6%)	1,484 (85.2%)	1.000
Anxiety	211 (44.2%)	866 (49.7%)	1.000
Adjustment disorder	72 (15.1%)	239 (13.7%)	1.000
Substance-related	42 (8.8%)	149 (8.6%)	1.000
ADHD	40 (8.4%)	85 (4.9%)	0.128
Psychiatric medications			
Antidepressants	469 (98.3%)	1,715 (98.5%)	1.000
Anxiolytics	178 (37.3%)	782 (44.9%)	0.117
Analgesics	168 (35.2%)	706 (40.6%)	1.000
Anti-convulsant	120 (25.2%)	523 (30.0%)	1.000
Mood stabilizer	61 (12.8%)	424 (24.4%)	<0.001
Annualized HCRU			
Outpatient	475 (99.6%)	1,734 (99.6%)	1.000
Emergency department	78 (16.4%)	258 (14.8%)	1.000
Inpatient	50 (10.5%)	125 (7.2%)	0.693

Patients with **treatment failure** had **significantly higher PHQ-9 scores** at baseline (continuous) and follow-up compared to scores of patients with no treatment failure (**Table 3**).

Table 3. PHQ-9 scores from index date to censured exposure end date in patients with MDD with and without treatment failure

	No treatment failure (%) (n=477)	Treatment failure (%) (n=1,741)	*P*-value
Baseline PHQ-9 Score (Continuous)			0.047
Mean (SD)	11.91 (6.18)	12.82 (6.27)	
Median (IQR)	10.00 (9.00, 16.00)	12.00 (9.00, 18.00)	
Baseline PHQ-9 Score (Categorical)			1.00
No MDD (Score 0–4)	53 (11.1%)	147 (8.4%)	
Mild (Score 5–9)	158 (33.1%)	529 (30.4%)	
Moderate (Score 10–14)	104 (21.8%)	385 (22.1%)	
Moderate to Severe (Score 15–19)	92 (19.3%)	383 (22.0%)	
Severe (Score 20–27)	70 (14.7%)	297 (17.1%)	
Follow-up PHQ-9 Score			<0.001
Mean (SD)	9.47 (6.15)	11.93 (6.36)	
Median (IQR)	9.00 (5.00, 12.00)	10.00 (8.00, 17.00)	
PHQ-9 Change Score			<0.001
Mean (SD)	−2.44 (6.12)	−0.89 (6.27)	
Median (IQR)	0.00 (−6.00, 0.00)	0.00 (−4.00, 1.00)	
Follow-up duration (months)			0.217
Mean (SD)	6.60 (8.81)	5.65 (7.77)	
Median (IQR)	3.37 (2.03, 7.30)	2.93 (1.63, 6.23)	

When stratified by baseline PHQ-9 score to define MDD severity, the results indicated a **trend of increased follow-up PHQ-9 scores in line with MDD severity (Fig. 1)**. Across disease severity groups, patients with treatment failure showed higher mean PHQ-9 scores compared to those with no failure. A greater score reduction was observed among patients with moderate to severe MDD and treatment failure vs. those without (**Fig. 2**).

Fig. 1. Mean follow-up PHQ-9 scores between patients with and without treatment failure events stratified by baseline MDD severity

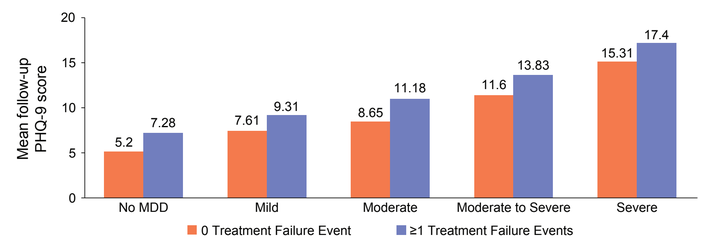
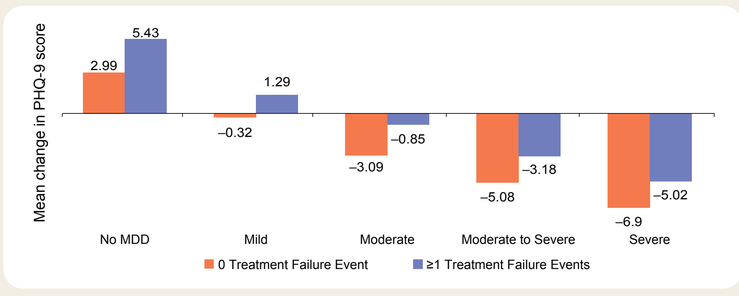


Fig. 2. Comparison of mean change of PHQ-9 from baseline to follow-up time between patients with and without treatment failure events stratified by baseline MDD severity



In **regression analyses**, failure status (follow-up score model *P*-value = 8.54×10^{-8}; score change model *P*-value = 5.13×10^{-13}) and **baseline PHQ-9 severity** (follow-up score model *P*-value ≤9.18×10^{-4}; score change model *P*-value <2.25×10^{-17}) were the only variables **significantly associated with PHQ-9 follow-up scores or score change in overall MDD dataset**, respectively (**Table 4**).

Table 4. Follow-up and change from baseline PHQ-9 models results across failure status and in overall MDD datasets

	Follow-up PHQ-9 score		PHQ-9 score change from baseline	
	Negative binomial ratio vs. reference group (95% CI)		Linear model results vs. reference group (mean [SE])	
	Crude	Adjusted	Crude	Adjusted
Failure status "Yes" ("No fail" as reference)	1.26 (1.14–1.4)	1.32 (1.19–1.47)	2.02 (0.29)	2.16 (0.3)
Baseline PHQ-9 according to MDD severity				
Mild	1.28 (1.08–1.51)	1.31 (1.11–1.56)	−4.03 (0.45)	−3.95 (0.46)
Moderate	1.82 (1.54–2.16)	1.86 (1.56–2.21)	−6.12 (0.47)	−6.2 (0.48)
Moderately severe	2.61 (2.2–3.09)	2.64 (2.21–3.14)	−8.2 (0.47)	−8.32 (0.48)
Severe	3.5 (2.93–4.18)	3.55 (2.96–4.25)	−10 (0.49)	−10.2 (0.5)
No MDD (reference)	1	1	0	0

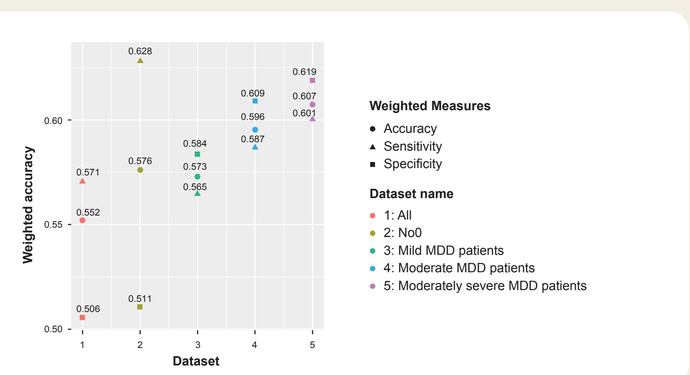
In **subset analyses**, effect sizes were calculated for overall and specific datasets. Negative binomial model performance was defined by Nagelkerke R^2 score (**Table 5**).

Table 5. Adjusted negative binomial model performance and results in the overall dataset and across defined MDD subsets

	Follow up PHQ-9 difference by failure proxy (yes or no) (95% CI)	Follow up PHQ-9 ratio by failure proxy (yes or no) (95% CI)	Formula	Nagelkerke R^2	Cohen's d
Overall dataset	6.12 (4.03–8.2)	1.32 (1.19–1.47)	Follow-up PHQ-9 score ~ baseline PHQ-9 severity status + fail proxy (yes or no) + baseline covariates (Dx, Rx, encounters + psychotherapy treatment)	0.26	−0.39
No0	11.18 (8.66–13.7)	1.71 (1.49–1.97)		0.19	−0.54
Mild MDD	11.77 (9.14–14.41)	1.76 (1.51–2.04)		0.19	−0.55
Moderate MDD	10.31 (7.27–13.36)	1.61 (1.36–1.89)		0.18	−0.52
Moderately severe MDD	10.45 (6.5–14.41)	1.59 (1.29–1.94)		0.16	−0.48

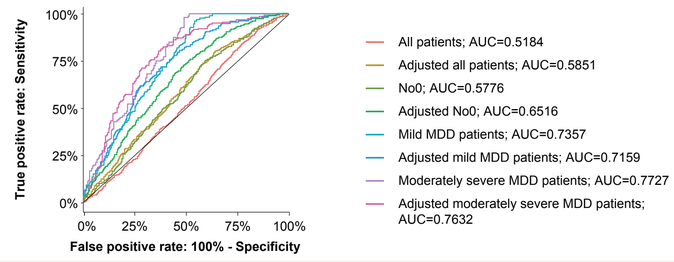
Weighted **validation analyses results** between proxy failure vs. no failure group showed the **highest accuracy** in moderately severe MDD patients (baseline PHQ-9 ≥15) (**Fig. 3**).

Fig. 3. Validation between score failure event (gold standard outcome) and treatment pattern proxy failure event



Logistic regression models of PHQ-9 score failure yielded a maximum **area under the curve (AUC) of 0.76** in moderately severe MDD patients (**Fig. 4**), suggesting that the score failure measure may be a key factor associated with increased odds of proxy failure events. In addition to disease severity, the use of anxiolytics (OR, 95% CI: 2.03 [1.4–2.96]) or mood stabilizers (OR, 95% CI: 2.68 [1.59–4.74]) were also identified as significant risk factors.

Fig. 4. Association between treatment failure and PHQ-9 score change from baseline using a logistic regression model



Additional Conclusion

Treatment complexity, defined by the number of treatment lines or generic drugs used, was associated with increased symptom burden and greater changes in PHQ-9 scores.

Abbreviations

ADHD, attention deficit hyperactivity disorder; AUC, area under the curve; CI, confidence interval; Dx, diagnosis; HCRU, healthcare resource utilization; IQR, interquartile range; MDD, major depressive disorder; OR, odds ratio; PHQ-9, Patient Health Questionnaire 9; Px, prescription; SD, standard deviation; SE, standard error.

References

1. GBD 2017 Disease and Injury Incidence and Prevalence Collaborators. *Lancet*. 2018;392(10159):1789–1858.
2. Zhdanava M, et al. *J Clin Psychiatry*. 2021;82(2):20m13699.
3. Koch E, et al. *Lancet Psychiatry*. 2025;12(6):457–468.
4. van Dijk DA, et al. *J Affect Disord*. 2025;382:68–84.

Disclosures

LZ, CW, SSR, FDC, AK, and SDS are full-time employees of Boehringer Ingelheim. RP has participated in Scientific Advisory Boards for Boehringer Ingelheim and Teva, has received grant funding from Janssen, and has received consulting fees from Holmusk, Akrivia Health, Columbia Data Analytics, Clinilabs, Social Finance, Boehringer Ingelheim, Bristol Myers Squibb, Supernus, Teva, and Otsuka.

Acknowledgements

The authors meet the criteria for authorship as recommended by the International Committee of Medical Journal Editors. This study was funded by Boehringer Ingelheim (1447-0011). Writing, editorial support, and formatting assistance were provided by Sandra Calcat i Cervera, PhD.

Boehringer Ingelheim

Presented at the ISPOR Europe 2025, 9–12 November 2025, Glasgow, Scotland, UK