Multi-Arm Trial Design with Distinct Biomarkers

April 1, 2025

1 Introduction and Motivation

Biomarkers are variables that can be measured before treatment that are correlated to treatment outcome, ideally above standard of care. The idea in this brief is that we way to power a multi-arm RCT that validates promising biomarkers already reported in the literature. That is, the power calculation is powered to detect the significance of *biomarkers* themselves, rather than treatment effect across the population of study. The point of this is (1) to build off the work of others, as finding biomarkers from scratch could take even more sample size; (2) this would yield an excellent dataset from which to derive a post-hoc multiclass model that, using *all* biomarker data as inputs to a model, could determine which subset of treatments an individual is most indicated for.

Let's now describe the setup for such a trial:

- **Randomization** into one of K experimental arms,
- **Multiple biomarkers** $\{X_1, \ldots, X_K\}$, each potentially relevant to a specific experimental treatment,
- **Outcome** Y (e.g. depression severity) measured post-treatment.

Our main goal is to evaluate, for each experimental treatment k, whether the corresponding biomarker X_k *predicts* an increased treatment effect. In turn, we'll develop the following:

- A simple causal model (DAG) for the multi-arm trial,
- A linear model capturing main and interaction effects,
- A data-generation scheme for simulation studies,
- Methods for multiple testing and power analysis in this multi-biomarker context,
- Empirical estimates of biomarker effects from literature,
- Results from the power analysis, given these assumptions.

2 Model Setup

DAG Representation

We illustrate a basic causal structure for a multi-arm trial with randomization, treatments, correlated biomarkers, and outcome.

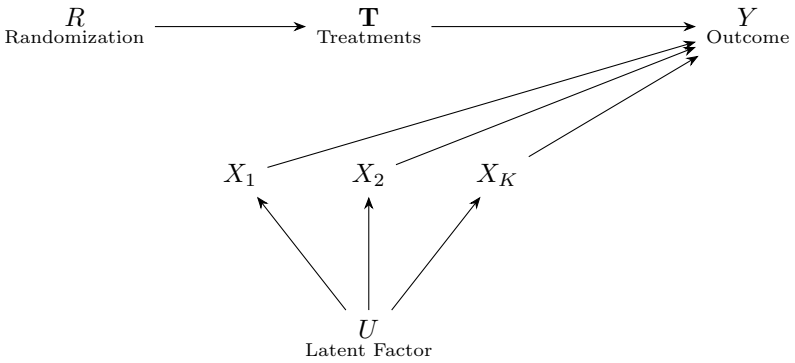


Figure 1: Simple DAG for a multi-arm trial with randomization node R, treatments T, optional latent factor U (introducing correlation among biomarkers), and outcome Y.

Linear Model for the Outcome

We assume a linear model capturing main effects of each treatment and possible interactions with corresponding biomarkers:

$$\begin{aligned} Y_i \; &= \beta_0 \; + \; \sum_{k=1}^{K} \Big[\beta_{1k}\, T_{ik} \; + \; \beta_{2k}\, X_{ik} \; + \; \beta_{3k}\, \big(T_{ik} \cdot X_{ik}\big) \Big] \; + \; \varepsilon_i, \\ &\varepsilon_i \sim \mathcal{N}(0, \sigma^2), \end{aligned}$$

where:

- β_0 is the baseline intercept.
- β_{1k} is the *main effect* of treatment k relative to an implicit baseline condition (e.g., the average of other arms, or arm 1 if using treatment contrasts).
- β_{2k} captures the baseline effect of biomarker X_k (assumed relevant to arm k), with the understanding that β_{2k} can affect outcome under *all* arms.
- β_{3k} measures the interaction (moderation) of treatment k by biomarker X_k. If $\beta_{3k} > 0$, higher values of X_k predict greater treatment benefit for arm k.

Note that because T_{ik} is binary and $\sum_{k=1}^{K} T_{ik} = 1$, every subject belongs to exactly one experimental arm.

3 Simulation Framework for Multi-Arm Trials

To evaluate the operating characteristics of a multi-arm design that includes distinct biomarkers and treatments, one can simulate data as follows. Suppose there are K experimental arms, indexed by $k = 1, \ldots, K$. Each subject i belongs to exactly one arm, indicated by the binary variables $\{T_{i1}, \ldots, T_{iK}\}$, with $\sum_{k=1}^{K} T_{ik} = 1$. We also have K biomarkers $\{X_1, \ldots, X_K\}$. Let N be the total number of subjects.

1. **Generate latent factor(s) (optional).** If correlation among the biomarkers is desired, one or more latent factors can be introduced. For instance, draw $U_i \sim \mathcal{N}(0, 1)$ for each subject i. Later steps can incorporate this factor into the biomarker generation process.

2. **Generate biomarkers.** For each subject i and biomarker $k \in \{1, \ldots, K\}$, set

$$X_{ik} \;=\; \mu_k \;+\; \gamma_k\, U_i \;+\; \sigma_{X_k}\, \eta_{ik}, \quad \eta_{ik} \sim \mathcal{N}(0, 1).$$

   Here, γ_k controls how strongly the latent factor U_i influences the kth biomarker, and σ_{X_k} scales any residual variation.

3. **Assign treatments.** Each subject is randomized into exactly one arm. A convenient approach is to sample $\{T_{i1}, \ldots, T_{iK}\}$ from a multinomial distribution with probabilities $(p_1, \ldots, p_K)$, so that one $T_{ik} = 1$ and the rest are 0. Often, $p_1 = \cdots = p_K = \frac{1}{K}$ for equal allocation.

4. **Generate the outcome.** Using the linear model

$$Y_i \;=\; \beta_0 \;+\; \sum_{k=1}^{K} \Big[\beta_{1k}\, T_{ik} \;+\; \beta_{2k}\, X_{ik} \;+\; \beta_{3k}\, \big(T_{ik} \cdot X_{ik}\big)\Big] \;+\; \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2),$$

   each subject's outcome depends on which experimental arm they are assigned to ($T_{ik} = 1$ for exactly one $k \in \{1, \ldots, K\}$), the values of the biomarkers $\{X_{ik}\}$, and the error term ε_i. The intercept β_0 represents a baseline outcome level, β_{1k} is the main effect of arm k (relative to an implicit baseline), β_{2k} captures any overall effect of X_k, and β_{3k} is the interaction term indicating how biomarker X_k moderates treatment k.

By repeating this process for multiple simulated datasets, one can fit the same linear model

$$Y \;\sim\; \{T_1, \ldots, T_K\} \;+\; \{X_1, \ldots, X_K\} \;+\; \{T_k \cdot X_k\}_{k=1}^{K}$$

to each dataset and evaluate how often the estimated interaction terms $\hat{\beta}_{3k}$ are statistically significant. This approach yields estimates of power, type I error rate, or other design metrics, accounting for the correlation structure among biomarkers and multiple treatment arms.

4 Candidate Biomarkers and Empirical Estimates of Effects

Below we consolidate and discuss illustrative findings from the literature, focusing on four candidate biomarkers that may moderate treatment response in depression. We present these examples to highlight the breadth of effect sizes, populations, and methodological nuances that can arise when designing multi-arm biomarker trials.

4.1 Alpha Peak Frequency (iAPF) for rTMS

Three studies have examined whether *individual alpha peak frequency* (iAPF) predicts response to 10 Hz or 1 Hz rTMS protocols in Major Depressive Disorder (MDD), as well as to neurofeedback or medication in ADHD. Voetterl *et al.* (2019) [9], though focused on ADHD, reported that matching patients' iAPF decile to a particular intervention (methylphenidate vs. neurofeedback plus sleep coaching) yielded a sizable remission gain (on the order of 15–30 %). Roelofs *et al.* (2021) [7] and Corlier *et al.* (2019) [1] studied MDD populations receiving rTMS, finding that *proximity* of iAPF to the 10 Hz stimulation frequency explained about 5–10 % of variance in symptom improvement (a correlation of about $r \approx -0.25$ for Roelofs, $r \approx 0.30$ for Corlier), whereas no effect was apparent in the 1 Hz group or in patients who switched off 10 Hz mid-treatment.

Although the exact magnitudes differ by population and methodology (and ADHD vs. MDD is an imperfect comparison), a recurring theme is that *alignment* or *closeness* between one's alpha peak and the rTMS frequency can moderate clinical outcome. Correlations in the range of 0.20–0.30 (i.e. explaining roughly 5–10 % of outcome variance) are consistent with a moderate interaction effect ($\beta_3 \approx 0.2$–0.3 if the outcome is continuous). Stronger effects have been reported in some settings (e.g. ADHD). However, non-linear patterns (peaks around 10 Hz rather than strictly linear) may matter, complicating a simple linear interaction model. Hence, an RCT seeking to confirm iAPF as a predictive biomarker for rTMS would likely assume at least a modest but clinically meaningful interaction size.

Caveats. Voetterl's ADHD sample differs in population, comorbidity, and outcome measures relative to MDD rTMS trials. Corlier's retrospective design (with possible protocol switches) may attenuate effect estimates. Nonetheless, across all three studies, alpha-frequency alignment emerges as a plausible biomarker worthy of prospective validation.

4.2 EHR-Based Predictors of Ketamine Response

A series of four small-to-moderate trials [4, 5, 3, 6] found that *family or personal history of alcohol use disorder (AUD)* predicts a more robust antidepressant response to ketamine in both MDD and bipolar depression. For instance, Niciu *et al.* (2015) reported that a combination of family history of AUD, BMI, and prior suicide attempt explained up to 36 % of variance in outcome by Day 7 post-infusion. Meanwhile, Permoda–Osip *et al.* (2014) and Luckenbaugh *et al.* (2012) observed that 45–60 % of responders (vs. about 15–20 % of non-responders) had a positive AUD history, corresponding to odds ratios around 3–5. Phelps *et al.* (2009) noted an absolute difference of nearly 50 % in response rates (67 % vs. 18 %), yielding an even larger odds ratio.

Such large contrasts suggest a potentially strong *treatment* $\times$ *AUD-history* interaction. In a logistic model, an odds ratio of about 3 corresponds to a log-odds coefficient $\ln(3) \approx 1.1$. Even if some estimates are inflated by small sample sizes or retrospective biases, these consistent findings highlight that certain EHR-derived variables can have substantial predictive value for ketamine response. A well-powered multi-arm trial (one of the arms being ketamine) that collects family/personal AUD

data could detect this interaction with fewer subjects than is typically required for more modest biomarkers.

Caveats. These results come from relatively small samples and mixed unipolar/bipolar populations. Confirmatory RCTs with strict randomization and larger N are still needed to rule out confounding or selection biases. In practice, investigators might combine AUD history with other EHR factors into a multi-variable biomarker.

4.3 Inflammatory Markers for ECT (Dellink et al., 2025)

A recent meta-analysis by Dellink *et al.* (2025) [2] aggregated 14 studies of ECT in depression ($n = 556$) and found that patients with higher baseline levels of *C-reactive protein* (CRP) or *interleukin-6* (IL-6) consistently showed greater symptom reduction from acute ECT courses. The effect, however, was modest: correlations of about $r \approx 0.20$, implying these markers explained about 4 % of the variance in outcomes. Changes in these markers *during* the course of ECT did not correlate strongly with final outcomes, suggesting a potential *trait-level* rather than *state-level* predictive mechanism.
For multi-arm trials that include an ECT arm, this finding supports the possibility of an *inflammatory-subtype* of depression that responds preferentially to ECT. Given $r \approx 0.20$, one can anticipate a smaller interaction coefficient (e.g. $\beta_3 \approx 0.2$ in a standardized linear framework). Detecting such an effect reliably would likely demand a larger sample size than is needed for strong predictors like AUD history.

Caveats. Heterogeneity exists in how each study measured CRP/IL-6 and defined "high inflammation." The $n = 556$ total still spans multiple ECT protocols and populations. Nonetheless, the consistent direction of effect (higher inflammation, better ECT response) warrants further prospective confirmation.

4.4 Speech Latency as a Voice Biomarker (Siegel et al., 2024)

Siegel *et al.* (2024) [8] explored *speech latencies*—the time between interviewer and patient speech turns—in a 3-arm bipolar depression trial of SEP-4199 vs. two comparators. Baseline psychomotor slowing, captured by longer latencies, correlated with more robust improvement under the *active* agent compared to placebo ($r \approx 0.30$). Patients with normal latencies tended to improve rapidly even on placebo, reducing any measured drug–placebo gap.
In a multi-arm design that includes a novel medication alongside other experimental treatments, such voice-based measures could be an inexpensive, objective biomarker. An observed partial correlation of 0.30 corresponds to about 9 % variance explained and a moderate $\beta_3 \approx 0.3$ in a linear model. Because speech recording imposes minimal burden, it may be an attractive option for biomarker collection in large trials.

Caveats. Psychomotor slowing may not be the only driver of delayed turn-taking; anxiety, cognitive impairment, or environment (e.g. telehealth) could also influence latencies. Further validation in bigger samples is needed to confirm specificity for medication response.

Overall, these four examples illustrate a range of plausible biomarker strengths and methodological complexity. Effect sizes vary from modest ($r \approx 0.20$) to quite large (odds ratios of 3–5). Different diagnostic groups (unipolar vs. bipolar depression, ADHD), sample sizes, and outcome measures each add layers of heterogeneity. For planning future *multi-arm RCTs*, careful synthesis of these prior estimates, along with explicit simulation (as outlined in previous sections), can guide sample size requirements and statistical methods for reliably detecting treatment–biomarker interactions.

5 Multiple Testing and Power Analysis

In a multi-arm, multi-biomarker setup, each $k = 1, \ldots, K$ generates a hypothesis $H_{0,k} : \beta_{3k} = 0$. To control the familywise error rate or false discovery rate, common adjustments include:

- **Holm's step-down (FWER):** Sort p-values in ascending order, apply successively stricter thresholds, and stop once a test fails rejection.
- **Benjamini–Hochberg (FDR):** Sort p-values and apply a threshold that controls the false-discovery rate at α.

One can define various *success criteria*, for example:

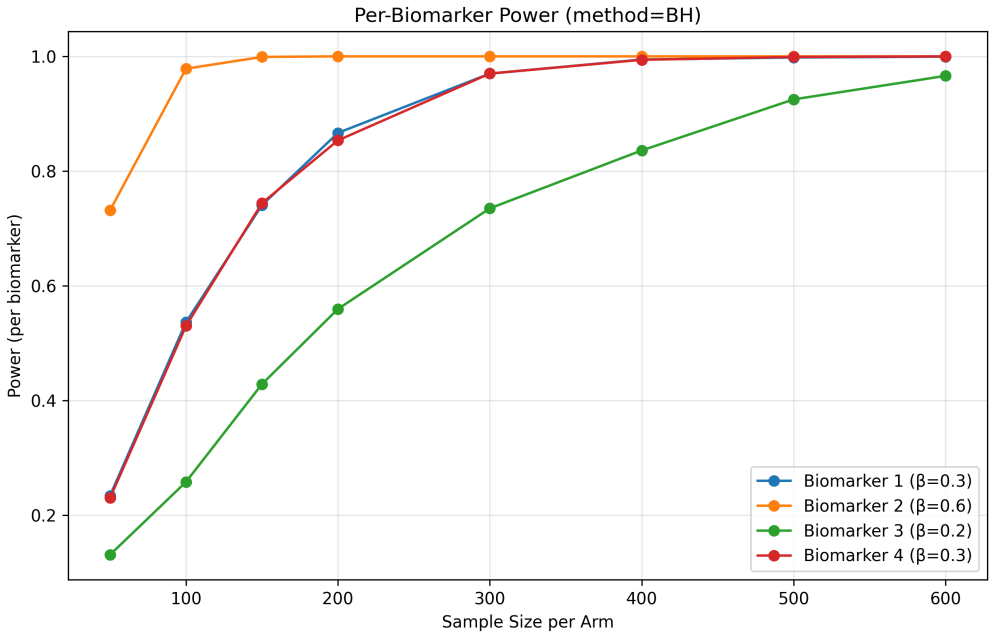
- `any` significant biomarker ($\exists k : p_k < \alpha^*$),
- `all` biomarkers significant ($\forall k : p_k < \alpha^*$),
- **Other subsets** (e.g. at least two biomarkers pass; unimplemented for now).

5.1 Illustrative Simulation Results

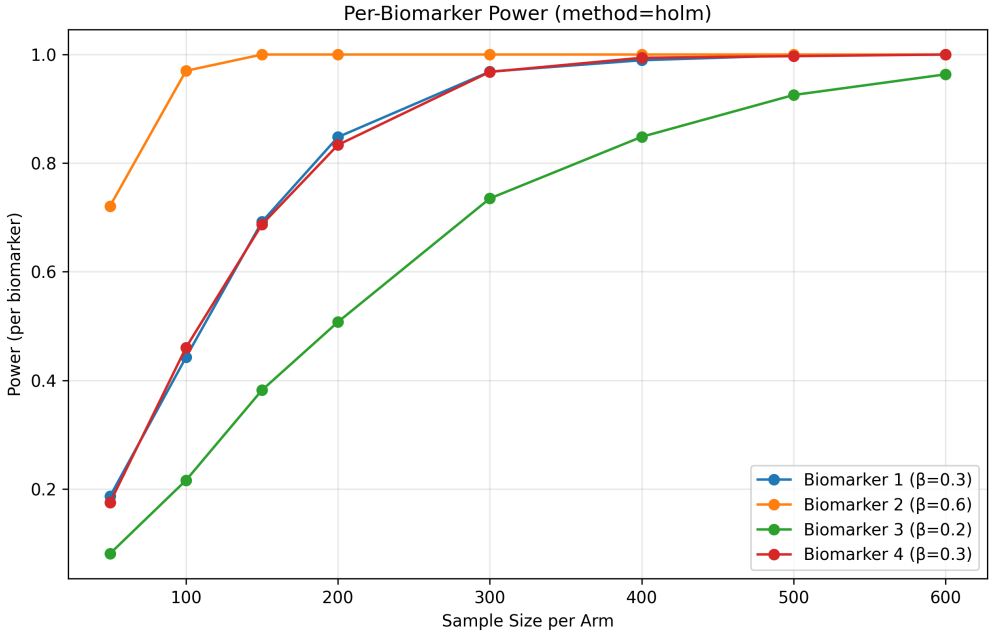
The following are the resulting power curves using the assumptions developed above.

Observations.

- **Per-Biomarker Power** (Figure 2) reveals stark differences in detection rates:
  - The strongest biomarker (AUD history for ketamine, $\beta_3 \approx 0.6$) achieves 80% power with just 100 subjects per arm.
  - Moderate biomarkers (iAPF for rTMS, $\beta_3 \approx 0.3$) require 300–400 subjects per arm for similar power.
  - Weaker biomarkers (inflammatory markers for ECT, $\beta_3 \approx 0.2$) need 500+ subjects per arm.
- **Overall Success** (Figure 3) shows dramatic differences between criteria:
  - Under `any`, power grows rapidly once the strongest biomarker achieves moderate power, reaching 80% with 300 subjects per arm.
  - Under `all`, power is dominated by the weakest biomarker, requiring 2000+ subjects per arm to achieve 80% power.
  - The gap between `any` and `all` widens as sample size increases, reflecting the challenge of simultaneous detection.
- **Multiple Testing Methods** show minimal impact:
  - BH and Holm corrections yield nearly identical power curves, especially under `any`.
  - With only $K = 4$ hypotheses, the numerical differences between correction methods are negligible.
  - The choice of success criterion (`any` vs. `all`) matters far more than the correction method.

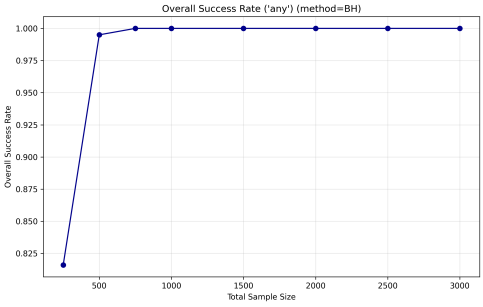


(a) Biomarker power: BH

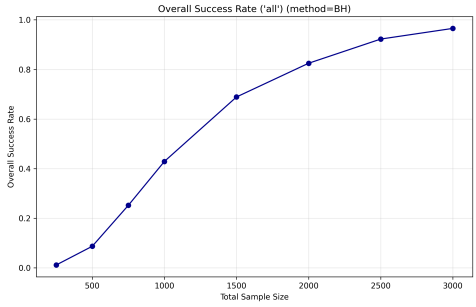


(b) Biomarker power: Holm

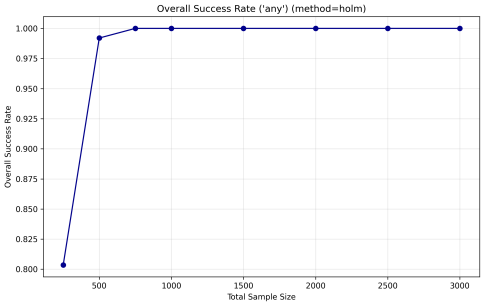
Figure 2: Example per-biomarker power curves under BH and Holm corrections.



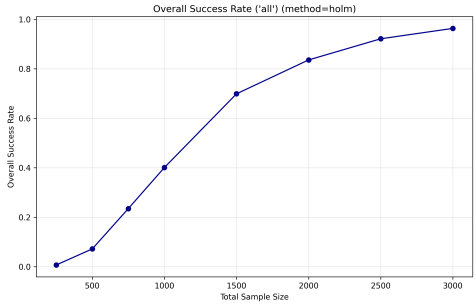
(a) Overall success (BH, `any`).



(b) Overall success (BH, `all`).



(c) Overall success (Holm, `any`).



(d) Overall success (Holm, `all`).

Figure 3: Example overall success rates under BH vs. Holm correction, comparing two success criteria: `any` significant biomarker vs. `all` biomarkers significant.

6 Summary

A multi-arm trial that evaluates distinct treatment–biomarker interactions can efficiently explore multiple hypotheses in a single protocol. Our simulation framework reveals several key insights for trial design:

- **Sample Size Requirements Vary Dramatically:**
  - Strong biomarkers (e.g., AUD history for ketamine) can be validated with modest samples ($n \approx 100$ per arm).
  - Moderate biomarkers (e.g., iAPF for rTMS) require 3–4x larger samples.
  - Weak biomarkers (e.g., inflammatory markers) may need 5x larger samples.
- **Success Criteria Drive Design:**
  - The `any` criterion is practical for exploratory trials, achieving 80% power with 300 subjects per arm.
  - The `all` criterion is extremely conservative, requiring 2000+ subjects per arm.
  - Investigators should carefully consider whether detecting all biomarkers is truly necessary.
- **Multiple Testing is Secondary:**
  - With few hypotheses ($K = 4$), the choice between BH and Holm has minimal impact.
  - The success criterion dominates the power analysis.
  - More sophisticated corrections may be unnecessary in this context.

Real-world biomarkers for TRD, such as iAPF alignment, alcohol-use history, inflammatory markers, or speech latencies, illustrate a range of plausible effect sizes. In practice, investigators should:

- Prioritize biomarkers based on prior evidence strength and clinical importance.
- Consider whether detecting all biomarkers is necessary or if finding any significant biomarker is sufficient.
- Use the simulation framework to explore different sample sizes and success criteria.
- Be prepared for substantial sample size requirements if all biomarkers must be validated simultaneously.

By examining the resulting power curves, one can plan a multi-arm trial that is appropriately powered to detect meaningful biomarker–treatment interactions while balancing statistical rigor with practical feasibility.

References

[1] Juliana Corlier, Linda L. Carpenter, Andrew C. Wilson, Eric Tirrell, A. Polly Gobin, Brian Kavanaugh, and Andrew F. Leuchter. The relationship between individual alpha peak frequency and clinical outcome with repetitive transcranial magnetic stimulation (rTMS) treatment of major depressive disorder (MDD). *Brain Stimulation: Basic, Translational, and Clinical Research in Neuromodulation*, 12(6):1572–1578, November 2019. doi: 10.1016/j.brs.2019.07.018. URL `https://doi.org/10.1016/j.brs.2019.07.018`.

[2] Annelies Dellink, Gertjan Vanderhaegen, Violette Coppens, Karen M. Ryan, Declan M. McLoughlin, Jennifer Kruse, Eric van Exel, Linda van Diermen, Jean-Baptiste Belge, Tore Ivar Malmei Aarsland, and Manuel Morrens. Inflammatory markers associated with electroconvulsive therapy response in patients with depression: A meta-analysis. *Neuroscience & Biobehavioral Reviews*, 170:106060, 2025. doi: 10.1016/j.neubiorev.2025.106060. URL `https://www.sciencedirect.com/science/article/pii/S0149763425000600`.

[3] David A. Luckenbaugh, Lobna Ibrahim, Nancy Brutsche, Jose Franco-Chaves, Daniel Mathews, Craig A. Marquardt, Christy Cassarly, and Jr. Zarate, Carlos A. Family history of alcohol dependence and antidepressant response to an n-methyl-d-aspartate antagonist in bipolar depression. *Bipolar Disorders*, 14(8):880–887, December 2012. doi: 10.1111/bdi.12003. Epub 2012 Sep 14.

[4] Mark J. Niciu, David A. Luckenbaugh, Dawn F. Ionescu, Sara Guevara, Rodrigo Machado-Vieira, Erica M. Richards, Nancy E. Brutsche, Neal M. Nolan, and Jr. Zarate, Carlos A. Clinical predictors of ketamine response in treatment-resistant major depression. *The Journal of Clinical Psychiatry*, 75(5):e417–e423, May 2014. doi: 10.4088/JCP.13m08698. Epub ahead of print 2014. PMID: 24922494.

[5] Agnieszka Permoda-Osip, Maria Skibińska, Alicja Bartkowska-Śniatkowska, Sebastian Kliwicki, Maria Chłopocka-Woźniak, and Janusz K. Rybakowski. [factors connected with efficacy of single ketamine infusion in bipolar depression]. *Psychiatria Polska*, 48(1):35–47, 2014. In Polish, with English abstract.

[6] Laura E. Phelps, Nancy Brutsche, Jazmin R. Moral, David A. Luckenbaugh, Husseini K. Manji, and Carlos A. Zarate. Family history of alcohol dependence and initial antidepressant response to an n-methyl-d-aspartate antagonist. *Biological Psychiatry*, 65(2):181–184, 2009. doi: 10.1016/j.biopsych.2008.09.029. URL `https://www.sciencedirect.com/science/article/pii/S0006322308011694`.

[7] Charlotte L. Roelofs, Noralie Krepel, Juliana Corlier, Linda L. Carpenter, Paul B. Fitzgerald, Zafiris J. Daskalakis, Indira Tendolkar, Andrew Wilson, Jonathan Downar, Neil W. Bailey, Daniel M. Blumberger, Fidel Vila-Rodriguez, Andrew F. Leuchter, and Martijn Arns. Individual alpha frequency proximity associated with repetitive transcranial magnetic stimulation outcome: An independent replication study from the icon-db consortium. *Clinical Neurophysiology*, 132(2):643–649, 2021. doi: 10.1016/j.clinph.2020.10.017. URL `https://www.sciencedirect.com/science/article/pii/S1388245720305320`.

[8] Joshua S. Siegel, Alex S. Cohen, Steven T. Szabo, Sasagu Tomioka, Mark Opler, Brian Kirkpatrick, and Seth Hopkins. Enrichment using speech latencies improves treatment effect size in a clinical trial of bipolar depression. *Psychi-*

atry Research, 340:116105, 2024. doi: 10.1016/j.psychres.2024.116105. URL https://www.sciencedirect.com/science/article/pii/S0165178124003901.

[9] Helena Voetterl, Guido van Wingen, Giorgia Michelini, Kristi R. Griffiths, Evian Gordon, Roger DeBeus, Mayuresh S. Korgaonkar, Sandra K. Loo, Donna Palmer, Rien Breteler, Damiaan Denys, L. Eugene Arnold, Paul du Jour, Rosalinde van Ruth, Jeanine Jansen, Hanneke van Dijk, and Martijn Arns. Brainmarker-I Differentially Predicts Remission to Various Attention-Deficit/Hyperactivity Disorder Treatments: A Discovery, Transfer, and Blinded Validation Study. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, 8(1):52–60, 2023. doi: 10.1016/j.bpsc.2022.02.007. URL https://www.sciencedirect.com/science/article/pii/S2451902222000465.