Investigating the robustness of a Bayesian meta-analysis of mesdopetam's anti-dyskinetic efficacy to different modelling choices

WERNER E., HANSSON F., WATERS S.

Integrative Research Laboratories (IRLAB), Gothenburg, Sweden · StanCon 2026

PROBABILITY OF BENEFIT

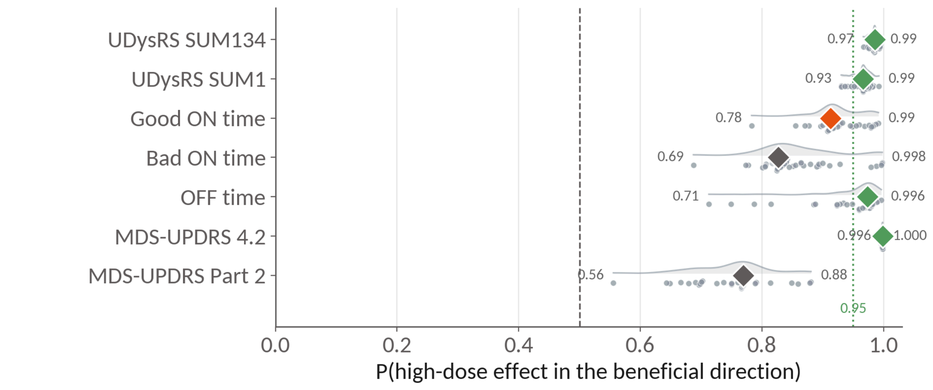


Figure 1: Posterior P(effect in the beneficial direction). Diamond = reference model. The raincloud (half-violin + one dot per specification) shows the full **distribution** of P across the included specifications. End labels give min/max P; the dotted line marks $P = 0.95$.

EFFECT SIZE BY ENDPOINT

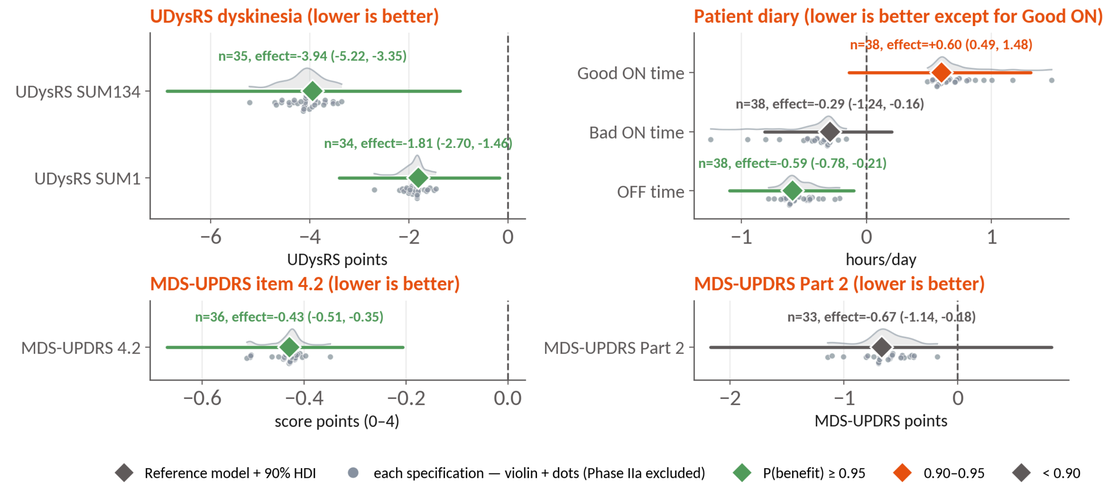


Figure 2: Reference-model effect (± 90% HDI) by measurement scale. Each effect is the **top-dose — placebo** difference in the endpoint's own raw units. The raincloud (half-violin + one dot per specification) shows the full **distribution** of specification means; labels give the number of specifications, the reference **mean effect size**, and the range over included specifications.

INTRODUCTION

Mesdopetam (IRL790) is a dopamine D_3 receptor antagonist in clinical development for **levodopa-induced dyskinesia** in Parkinson's disease (**PD-LID**). PD-LID, which can be described as involuntary movements triggered by anti-parkinsonian medication, is a disabling condition aggravating as the disease advances, and there is a great need for improved medical therapies. The fluctuating nature of PD-LID and the partly subjective nature of the associated functional impairment make the evaluation of therapeutic interventions challenging, although several measurement instruments are available. The clinical efficacy of mesdopetam has been evaluated in a phase IIa and a phase IIb study, both randomized controlled trials comparing mesdopetam against placebo. A Bayesian meta-analysis of the results was previously presented as part of a poster at the 2024 MDS Congress. That work consists of a single meta-analysis; here we instead investigate how results vary under a large number of plausible model specifications.

A joint analysis of the two studies must address the fact that they differ in many important respects, including the patient population, the duration of the study and the dosing regimen. Further, even for a single study there are a large number of researcher degrees of freedom in terms of e.g. the statistical model used for each endpoint, the selection of a prior distribution, and the handling of missing data and dropouts. In this analysis, we construct and sample from a large number of Bayesian models to investigate how sensitive our efficacy estimates are to different modelling choices. Models are implemented in PyMC and sampled using Nutpie.

THE TWO STUDIES

Feature	Phase IIa	Phase IIb
Patients (FAS)	42 *	140
Countries	UK, Sweden	FR, IL, IT, PL, RS, US
Age, yr	67 ± 7	65 ± 9
Female	38%	46%
PD duration, yr	11 ± 6	10 ± 5
Hoehn & Yahr	3.2 ± 0.5	2.5 ± 0.7
Duration	4 weeks	12 weeks
Dosing	7.5 mg BID, titrated up or down	2.5/5/7.5 mg BID, titrated down only

* Phase IIa's amantadine-free subset; 30 amantadine users (amantadine is itself anti-dyskinetic) are excluded from the headline analyses.

MODEL UNIVERSE

Each endpoint is modelled as a hierarchical regression on a **latent severity**. Each visit-t severity carries the latent baseline b_i forward and adds a **change** predictor η_{it}: a baseline-severity term βb_i, a study-specific dose effect, visit and dose×visit terms, country/study offsets and a patient intercept; the latent scale is mapped to the observed endpoint through D and link g. The baseline b_i is not plugged in as data — it is itself a latent score, observed through the **same** measurement model D (errors-in-variables, so measurement noise does not attenuate β):

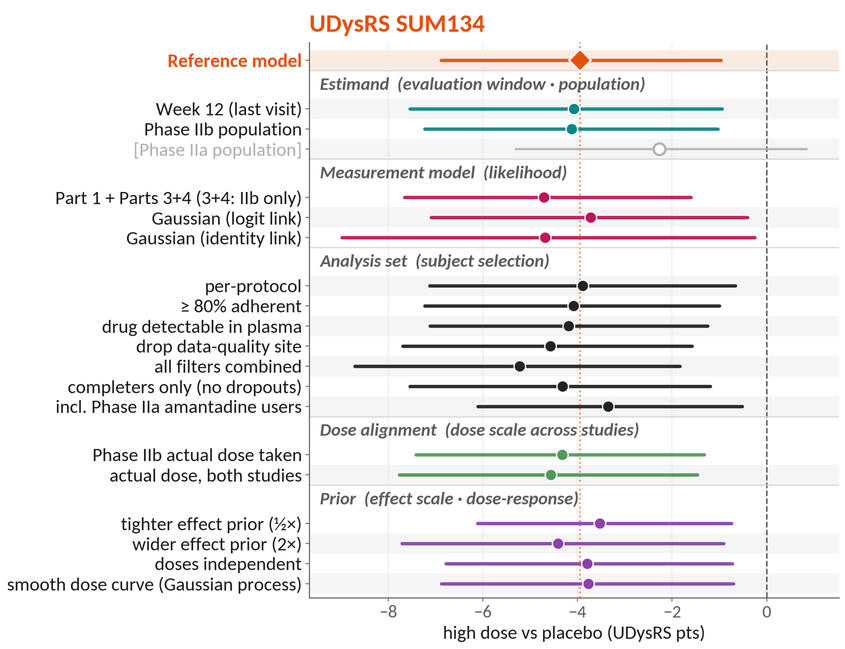
$$\eta_{it} = \alpha + \beta b_i + \delta_{dl(i)} + \tau_t + (\delta\tau)_{dt} + c_{k(i)} + s_{l(i)} + u_i$$
$$\delta_{dl} \sim N\left(\bar{\delta}_d, \tau_{\text{dose}}\right), \quad \delta_{0l} = 0$$
$$b_i \sim N(\mu_b, \sigma_b)$$
$$y_{i0} \sim D(g^{-1}(b_i)), \quad y_{it} \sim D(g^{-1}(b_i + \eta_{it}))$$

Here δ_{dl} is the default hierarchical study×dose interaction: a shared curve plus a study deviation, with τ_{dose} as the borrowing dial ($\tau_{\text{dose}} \to 0$ gives full pooling). Bounded item sums use Beta-Binomial models, the 0–4 UPDRS item uses ordinal logit, and diary hours use a generalized Dirichlet-multinomial composition. This default at each endpoint — the study×dose interaction, the measurement families just named, random-walk dose and visit curves and weakly-informative priors, read **time-averaged over Weeks 4·8·12 on the pooled two-study population** — is our **reference model**: the diamond in every forest.

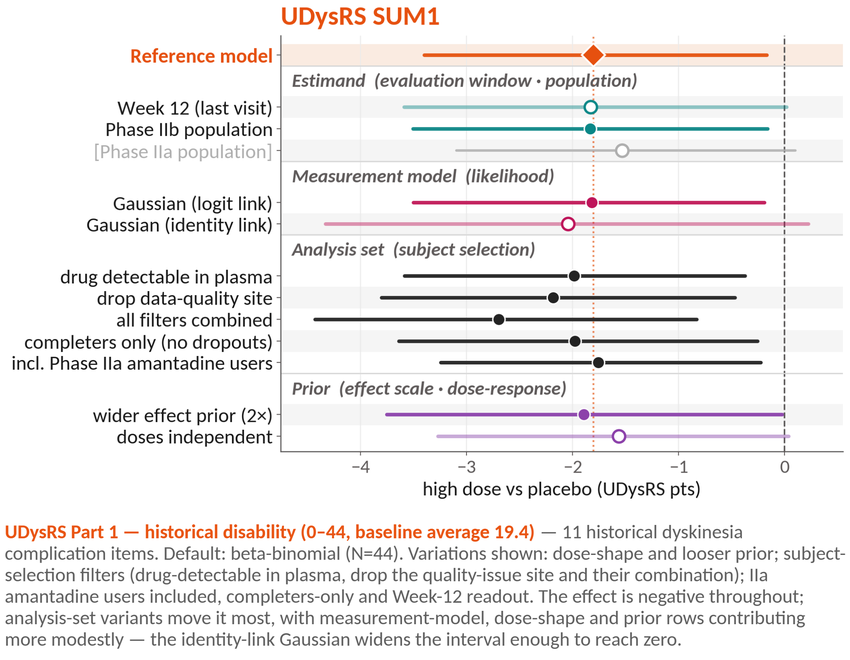
We then refit a curated sensitivity grid, varying one main modelling choice at a time where possible. The **measurement family** D is the widest axis: Gaussian linear models under a logit or an identity link, beta-binomial models for bounded item sums, binomial, ordinal logit **and** probit for the single 0–4 item, and a zero-inflated Beta alongside Dirichlet, Dirichlet-multinomial, logistic-normal, generalized Dirichlet and generalized Dirichlet-multinomial families for the diary endpoints. We compare a number of different priors for the **dose-response relationship** δ: random walk (the reference), second-order random walk, Gaussian process, monotone, Emax, quadratic or fully independent. The **visit-effect curve** is likewise a random walk or independent. We also **rescale the priors themselves**: the overall effect scale halved and doubled, the dose×time interaction widened, the country- and study-heterogeneity scales tightened and loosened (from near-complete pooling to weak meta-analytic borrowing), the nuisance hierarchy flattened, and the between-study borrowing dial τ_{dose} weakened or strengthened. **Dose alignment** d is taken as-randomized or swapped for the dose actually received in Phase IIa, Phase IIb or both, or reweighted by compliance. **Analysis-set** filters restrict the sample by per-protocol adherence, dose-modification, site quality, compliance, plasma exposure and their combination, with a complementary add-back that restores the Phase IIa amantadine users. For **missing data** we compare available-case (MAR) against a completers-only sample. Finally the **estimand** varies the *readout* of the contrast against placebo: pooled, Phase IIb-only or Phase IIa-only population (not included in the headline range), time-averaged over Weeks 4·8·12 or read at Week 12. Priors are weakly informative throughout.

PER-ENDPOINT FORESTS — SOME INTERESTING VARIATIONS

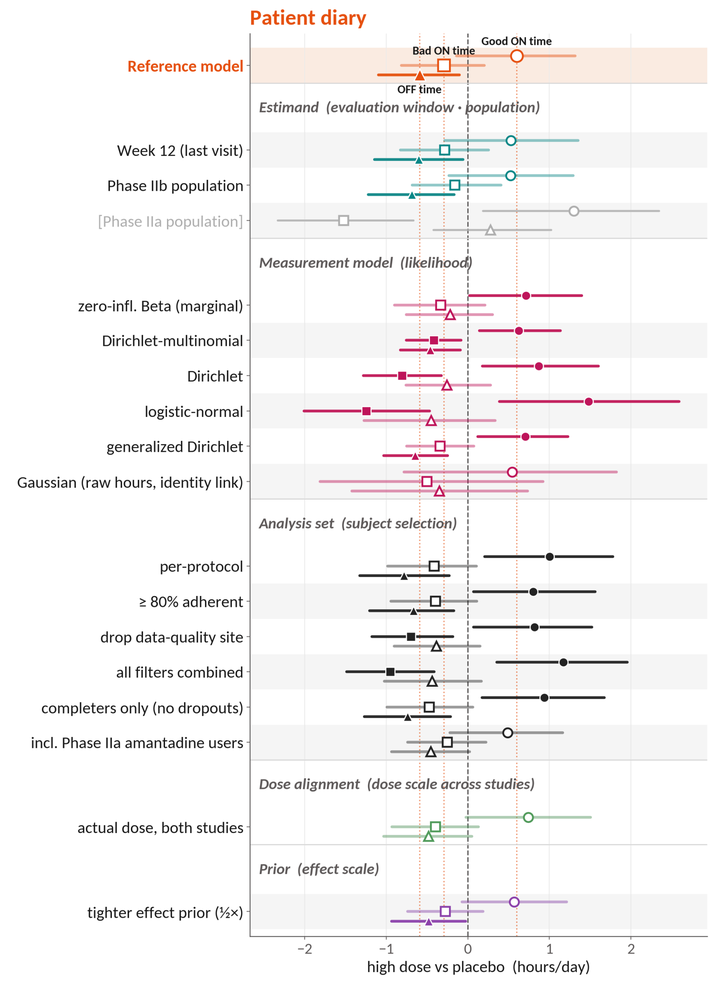
Estimand (window · population) · Measurement model · Analysis set · Dose alignment · Prior (incl. borrowing) · Reference model + 90% HDI · Reference mean (guide line) · 90% HDI includes no-effect · [bracketed] = excluded from ensemble (Phase IIa)



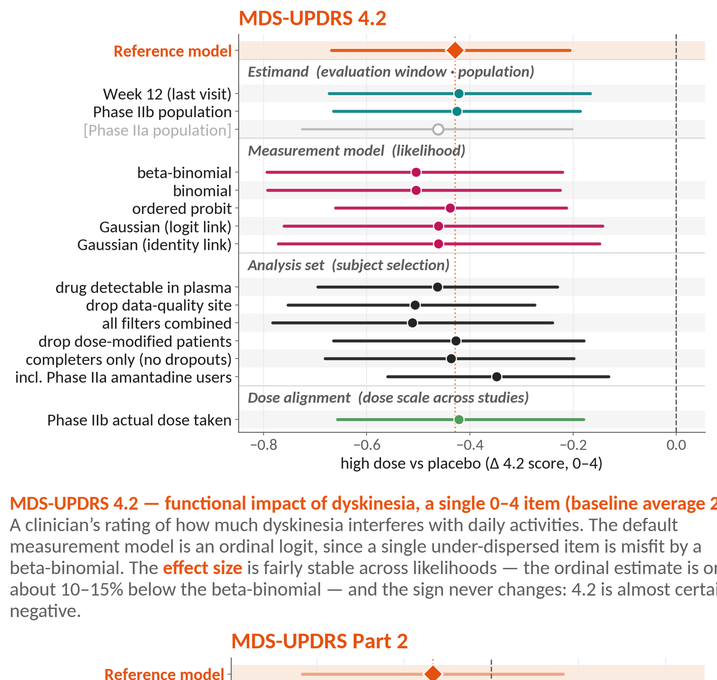
UDysRS dyskinesia, Parts 1+3+4 summed (0–88, baseline average 37.6) — 22 items scored 0–4. Part 1 is historical dyskinesia complications; Parts 3+4 are examiner-rated ON-state dyskinesia. The default is beta-binomial (N=88); a measurement variant drops unreliable Phase IIa Parts 3+4. Results are beneficial for all included models. Actual-dose variants, strict non-complier filtering and a wider effect prior strengthen the effect; including IIa amantadine users or a tighter prior attenuate it.



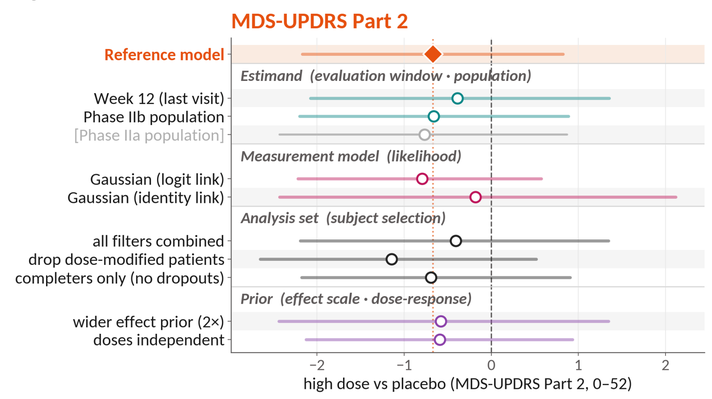
UDysRS Part 1 — historical disability (0–44, baseline average 19.4) — 11 historical dyskinesia complication items. Default: beta-binomial (N=44). Variations shown: dose-shape and looser prior; subject-selection filters (drug-detectable in plasma, drop the quality-issue site and their combination); IIa amantadine users included, completers-only and Week-12 readout. The effect is negative throughout; analysis-set variants move it most, with measurement-model, dose-shape and prior rows contributing more modestly — the identity-link Gaussian widens the interval enough to reach zero.



Patient diary — Good ON time (baseline average 6.7), Bad ON time (5.6) & OFF time (3.9) hours per day, fit jointly. Patients log each half-hour of the day as Good ON time, Bad ON time, OFF time, or asleep; the four sum to 24 h, forming a daily composition (we report the three waking states). The default measurement model is a generalized Dirichlet-multinomial on that composition, with sleep as the reference state. The results are quite sensitive to both the measurement model used, and the analysis set (stricter compliance checks have stronger effects). Most changes to the prior have almost no effect, the joint model and the fact that diary endpoints are measured over multiple days yields a more informative likelihood than for the other endpoints.



MDS-UPDRS 4.2 — functional impact of dyskinesia, a single 0–4 item (baseline average 2.2). A clinician's rating of how much dyskinesia interferes with daily activities. The default measurement model is an ordinal logit, since a single under-dispersed item is misfit by a beta-binomial. The **effect size** is fairly stable across likelihoods — the ordinal estimate is only about 10–15% below the beta-binomial — and the sign never changes: 4.2 is almost certainly negative.



MDS-UPDRS Part 2 — motor experiences of daily living, 13 items (0–52, baseline average 16.3). In this case a **safety** rather than a dyskinesia measure: does this D_3 antagonist worsen general motor function? The default measurement model is a beta-binomial with N=52. The effect is small (≈ −0.7 of 52) and its 90% HDI straddles zero — **no sign of motor worsening**.

CONCLUSIONS

For the range of instruments used to assess dyskinesia in the two trials, mesdopetam has a clear anti-dyskinetic effect, robust across the included model specifications. The comprehensive, standard dyskinesia assessment scale, UDysRS, is clearly improved vs placebo under most included specifications, both in terms of the total sum of parts 1, 3, and 4 assessing dyskinesia severity and disability, and part 1 assessing disability over the past week. Likewise, UPDRS 4.2, a single item capturing functional disability, is with high certainty improved vs placebo, with only a modest dependence of the effect size on the measurement model. Motor diary endpoints are very sensitive to both the measurement model and data-quality filtering, but both Good ON time and OFF time show a benefit under many different specifications, corroborating the findings on the other measurement scales applied. Importantly, MDS-UPDRS Part 2 (general motor function), included as a safety check to pick up adverse effects on motor function, shows no sign of worsening under any included specification.

Taken together, these analyses, combining data from two Phase 2 trials, applying somewhat different designs, and each with some technical challenges mainly related to the inherent difficulty to reliably quantify levodopa induced dyskinesias, indicate consistent, antidyskinetic effects both in terms of objective and subjective assessments. Furthermore, this effects does not occur at the cost of general motor impairment, ie the reduction of dyskinesia is not paired with any attenuation of the therapeutic effects of levodopa. Hence, mesdopetam could be a meaningful therapeutic option as add-on to levodopa for patients with advanced Parkinson's disease and disabling dyskinesia. From a modelling perspective, one take-away is that the measurement model often has a large impact, and a linear Gaussian model significantly increases the uncertainty of the effect estimate.

Contact: Erik Werner · erik.werner@irlab.se