Use of LLMs for Oversight in a Phase 3 Clinical Development Program

Todd Solomon, PhD
Senior Director | Clinical Development

Definium™ THERAPEUTICS

Disclosures

Full time Employee Definium Therapeutics
Prior Consulting Work: Alphasites, Signant, GLG, Decimal Health, EMA, Capvision

OUTLINE

What is Hammy

Model Development

Deploying Hammy into P3

Case Example

Challenges

What's the Problem LLMs Can Potentially Solve?

- In psychiatric drug development, clinician reported outcomes ClinROs obtained from participant interviews are considered gold standard primary efficacy endpoints
- ClinROs have significant limitations, including clinician bias, inter and intra-rater variability, poor sensitivity and the inherent subjectivity involved with psychological evaluation
- To mitigate these issues, sponsors have deployed methodologies such as rater training and certification, use of central raters, blinded data analytics and third-party review of endpoints to help reduce the risks associated with ClinRO assessments
- ISCTM is an organization exists to tackle these very issues

What is HAMMY

- Hammy is a Large Language Model (LLM) developed to provide oversight of the Hamilton Anxiety Scale (HAM-A)
- Mind Med created Hammy- a system of LLMs which transcribe ClinRO interviews, parse out individual ClinRO items, and provide associated scoring
- Hammy consists of multiple concurrent models that ingest audio recordings of ClinRO interviews and produce associated scores for each item of that interview

Hammy Pipeline

Fine-tuned from 1500 sessions from Phase 2b – 21,000 symptom ratings

Audio recording

HAM-A interviews between participants and central raters are recorded

Transcription

Whisper, an open-source transcription model, is prompt engineered to more accurately transcribe ClinRO interviews and is deployed on the recordings

Parse HAM-A items

Find the beginning of each of the 14 HAM-A items using regex or fine-tuned model from LLama 3.1 7B

Score items

Classify the severity of each item using a model fine-tuned from LLama 3.1 7B

Validating Hammy

1500 HAM-A interviews were audio recorded in P2

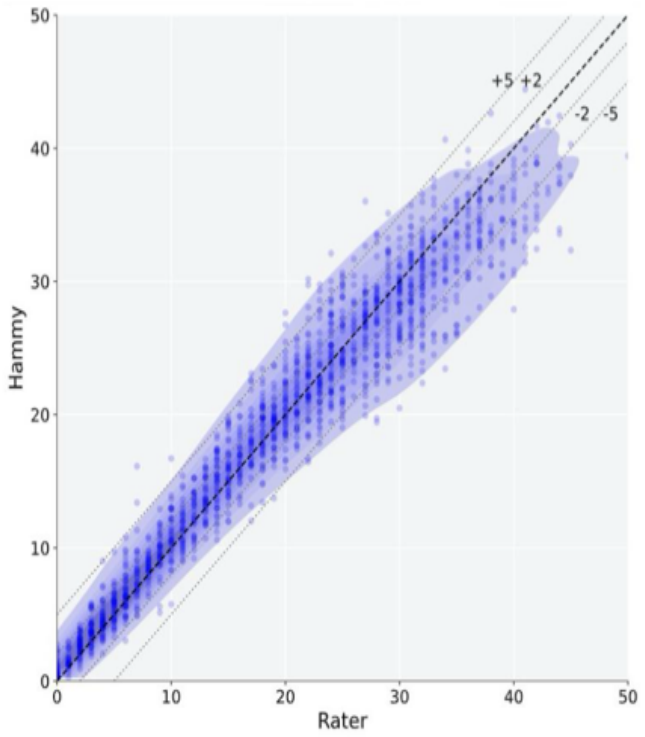
Cross-validation technique was used to create different versions of Hammy in order to appropriately test performance on the phase 2 dataset and approximate scoring confidence in other data sets

As a measure of performance, Hammy scored three training interviews used to certify human raters

Re-analyzed the results of the phase 2 study using Hammy scoring in place of the original central rater scoring.

When deployed on the phase 2 data, Hammy's scores differed on average 1.57 (+/- 1.39) points from the central raters scores, with Pearson r = 0.98)

Testing Hammy on Phase 2 Data and Training Interviews



Training Interview 1

Answer Key Score: 36-38
Hammy Score: 37

Training Interview 2

Answer Key Score: 8-11
Hammy Score: 8

Training Interview 3

Answer Key Score: 19-22
Hammy Score: 19

A graph with each dot representing a single HAM-A interview from the Phase 2b trial, plotted based on Hammy's total score (y-axis) and the central rater's total score (x-axis). The bold dashed line represents perfect agreement between the scores, with the grey dashed lines representing +/- 2 or +/- 5 points between the total scores.

Hammy achieved passing scores on all 3 training interviews as compared to an answer key created by rating experts

Kolar, Adam, et al. *Using Large Language Models for Endpoint Oversight*. (October, 2025). Poster presented at the International Society for CNS Clinical Trials and Methodology (ISCTM), Amsterdam (NE).

What Does Hammy NOT Do

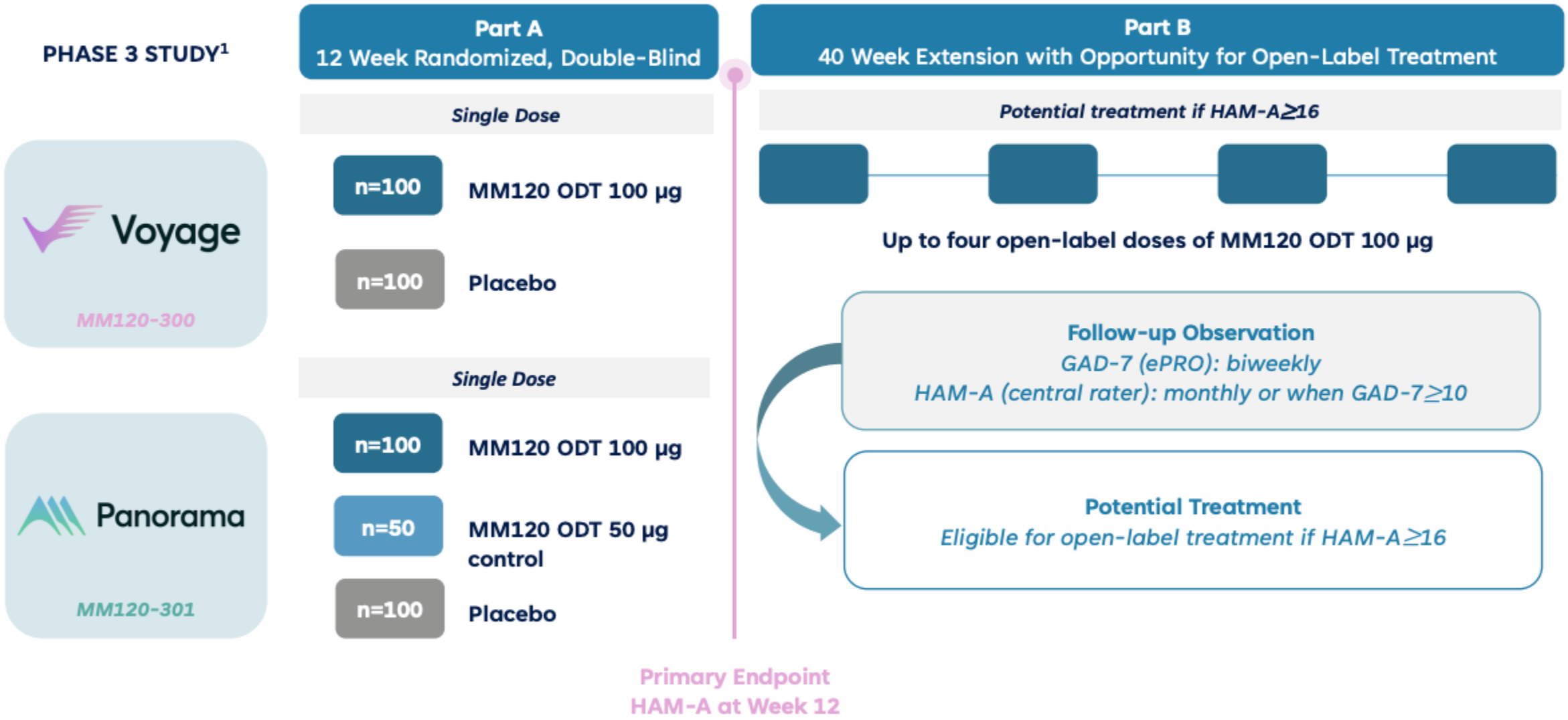
- Hammy is not replacing human raters
  - ……Yet

- Hammy does NOT produce study data
  - Scores are not part of the study data and CR scores are never changed or modified based on disagreement with Hammy

- Hammy does NOT impact eligibility
  - Scores that do not eligibility thresholds are never changed based on Hammy's rating alone

- Central or Site Raters are NOT stopped solely due to comparison to Hammy
  - Hammy provides an 'early detection' metric of discordance that can be confirmed via 3rd party review
  - With high fidelity of review, Hammy provides a consistent longitudinal metric to look at performance

MM120 for GAD | Two Complementary Pivotal Phase 3 Study Designs



1. Studies will employ an adaptive design with interim blinded sample size re-estimation based on nuisance parameters (e.g. patient retention rate, variability of primary outcome measure) to attempt to maintain statistical power. Clinical study designs subject to ongoing regulatory discussion and review, including of Phase 3 clinical trial protocols.

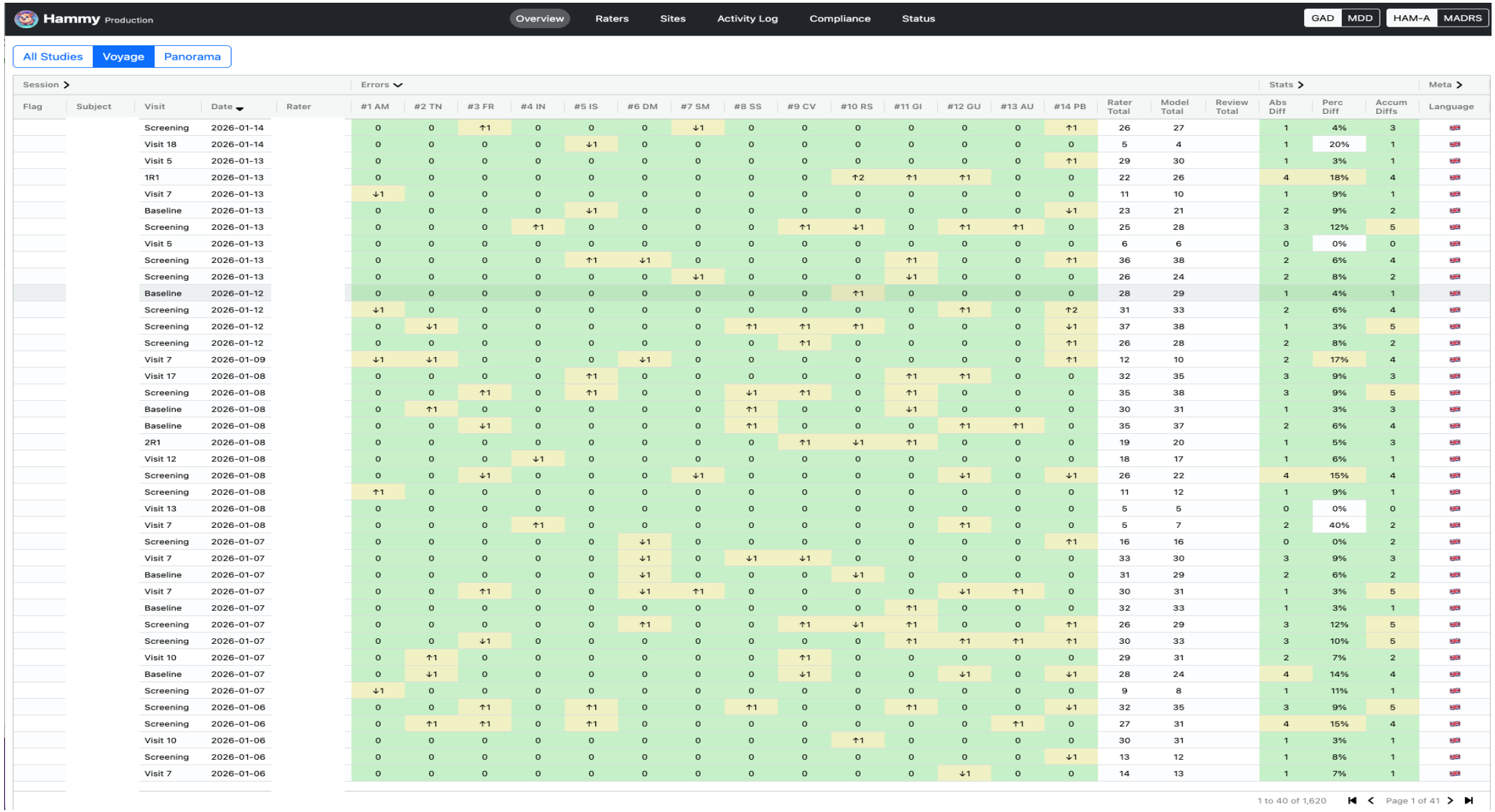
GAD: generalized anxiety disorder; GAD-7: diagnostic tool used to screen for and assess the severity of generalized anxiety disorder; HAM-A: Hamilton Anxiety Rating Scale; ODT: orally disintegrating tablet

How has Hammy been integrated into our studies

- Works very much like ‘traditional’ oversight programs which used a combination of a priori criteria to choose what assessments to review (First 2 per rater; SCN, BL, Primary; Large changes in scores, etc.)
  - 100% Fidelity: Every visit is reviewed (N= 3000)

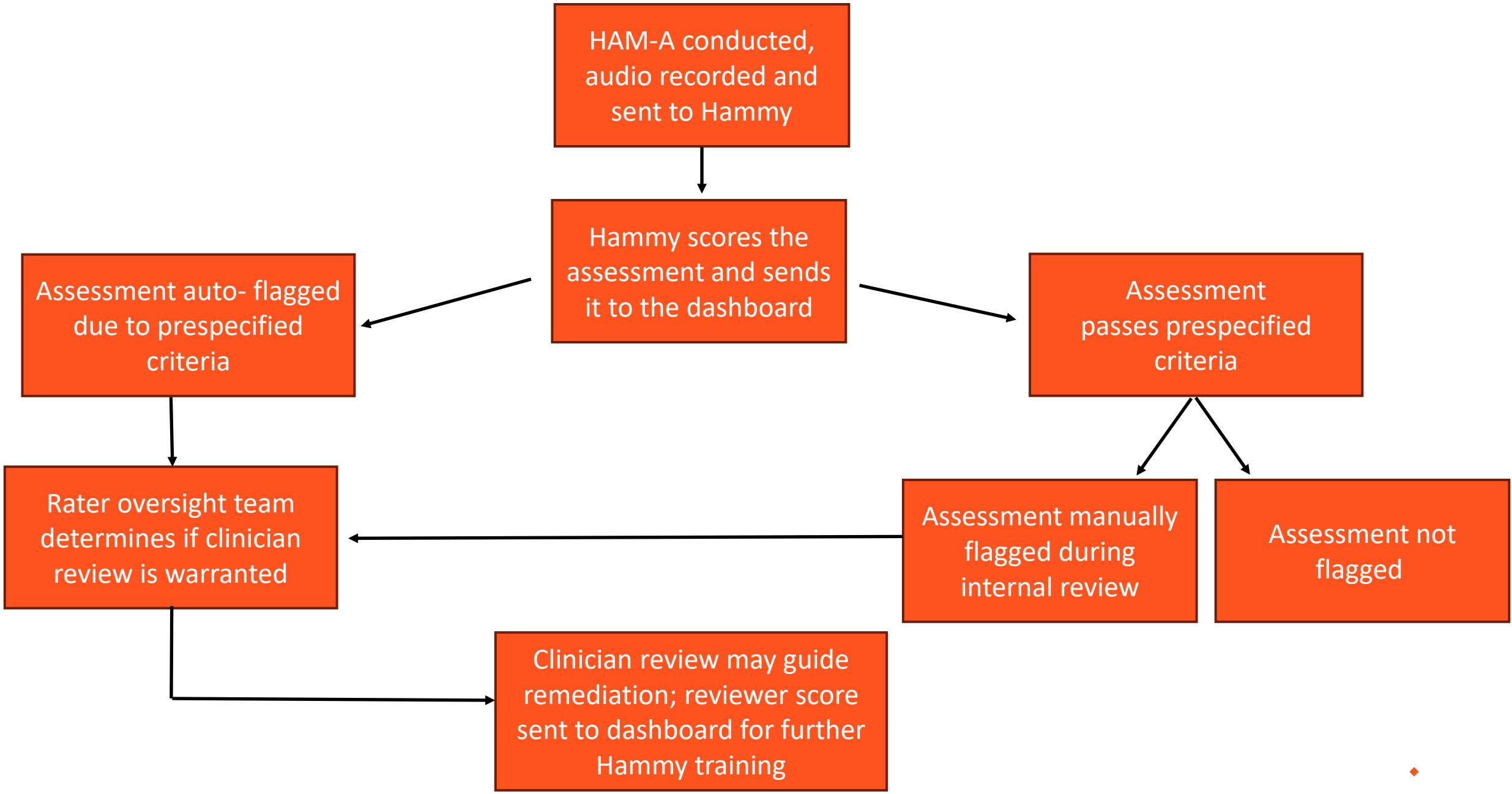
- Rater Total vs. Model Total
  - **Absolute Difference** = (Hammy Total Score – Rater Total Score)
  - **Percentage Difference** = (Absolute Difference / Rater Total Score) X 100
  - **Accumulated Difference** = Sum of absolute differences between Hammy and Rater **per symptom**
  - **Number Different** = Same as Accumulated Difference, but instead of summing magnitudes, we *count symptoms with any non-zero difference*, regardless of size
  - **Exclusion Agreement** = TRUE (if Hammy and Rater agree on IE criteria at SCN & BL)
  - **Max Error** = Maximum absolute error of any *individual symptom* within the session
  - **Audio Quality** = This is a proxy for cases where the participant is not audible and the transcript consists mostly of rater questions without responses
  - **Language Match** = Checks whether the actual transcript language matches the language code provided

Hammy Dashboard



Flag	Subject	Visit	Date	Rater	#1 AM	#2 TN	#3 FR	#4 IN	#5 IS	#6 DM	#7 SM	#8 SS	#9 CV	#10 RS	#11 GI	#12 GU	#13 AU	#14 PB	Rater Total	Model Total	Review Total	Abs Diff	Perc Diff	Accum Diffs	Language
		Screening	2026-01-14		0	0	↑1	0	0	0	↓1	0	0	0	0	0	0	↑1	26	27		1	4%	3	
		Visit 18	2026-01-14		0	0	0	0	↓1	0	0	0	0	0	0	0	0	0	5	4		1	20%	1	
		Visit 5	2026-01-13		0	0	0	0	0	0	0	0	0	0	0	0	0	↑1	29	30		1	3%	1	
		1R1	2026-01-13		0	0	0	0	0	0	0	0	0	↑2	↑1	↑1	0	0	22	26		4	18%	4	
		Visit 7	2026-01-13		↓1	0	0	0	0	0	0	0	0	0	0	0	0	0	11	10		1	9%	1	
		Baseline	2026-01-13		0	0	0	0	↓1	0	0	0	0	0	0	0	0	↓1	23	21		2	9%	2	
		Screening	2026-01-13		0	0	0	↑1	0	0	0	0	↑1	↓1	0	↑1	↑1	0	25	28		3	12%	5	
		Visit 5	2026-01-13		0	0	0	0	0	0	0	0	0	0	0	0	0	0	6	6		0	0%	0	
		Screening	2026-01-13		0	0	0	0	↑1	↓1	0	0	0	0	↑1	0	0	↑1	36	38		2	6%	4	
		Screening	2026-01-13		0	0	0	0	0	0	↓1	0	0	0	↓1	0	0	0	26	24		2	8%	2	
		Baseline	2026-01-12		0	0	0	0	0	0	0	0	0	↑1	0	0	0	0	28	29		1	4%	1	
		Screening	2026-01-12		↓1	0	0	0	0	0	0	0	0	0	0	↑1	0	↑2	31	33		2	6%	4	
		Screening	2026-01-12		0	↓1	0	0	0	0	0	↑1	↑1	↑1	0	0	0	↓1	37	38		1	3%	5	
		Screening	2026-01-12		0	0	0	0	0	0	0	0	↑1	0	0	0	0	↑1	26	28		2	8%	2	
		Visit 7	2026-01-09		↓1	↓1	0	0	0	↓1	0	0	0	0	0	0	0	↑1	12	10		2	17%	4	
		Visit 17	2026-01-08		0	0	0	0	↑1	0	0	0	0	0	↑1	↑1	0	0	32	35		3	9%	3	
		Screening	2026-01-08		0	0	↑1	0	↑1	0	0	↓1	↑1	0	↑1	0	0	0	35	38		3	9%	5	
		Baseline	2026-01-08		0	↑1	0	0	0	0	0	↑1	0	0	↓1	0	0	0	30	31		1	3%	3	
		Baseline	2026-01-08		0	0	↓1	0	0	0	0	↑1	0	0	0	↑1	↑1	0	35	37		2	6%	4	
		2R1	2026-01-08		0	0	0	0	0	0	0	0	↑1	↓1	↑1	0	0	0	19	20		1	5%	3	
		Visit 12	2026-01-08		0	0	0	↓1	0	0	0	0	0	0	0	0	0	0	18	17		1	6%	1	
		Screening	2026-01-08		0	0	↓1	0	0	0	↓1	0	0	0	0	↓1	0	↓1	26	22		4	15%	4	
		Screening	2026-01-08		↑1	0	0	0	0	0	0	0	0	0	0	0	0	0	11	12		1	9%	1	
		Visit 13	2026-01-08		0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	5		0	0%	0	
		Visit 7	2026-01-08		0	0	0	↑1	0	0	0	0	0	0	0	↑1	0	0	5	7		2	40%	2	
		Screening	2026-01-07		0	0	0	0	0	↓1	0	0	0	0	0	0	0	↑1	16	16		0	0%	2	
		Visit 7	2026-01-07		0	0	0	0	0	↓1	0	↓1	↓1	0	0	0	0	0	33	30		3	9%	3	
		Baseline	2026-01-07		0	0	0	0	0	↓1	0	0	0	↓1	0	0	0	0	31	29		2	6%	2	
		Visit 7	2026-01-07		0	0	↑1	0	0	↓1	↑1	0	0	0	0	↓1	↑1	0	30	31		1	3%	5	
		Baseline	2026-01-07		0	0	0	0	0	0	0	0	0	0	↑1	0	0	0	32	33		1	3%	1	
		Screening	2026-01-07		0	0	0	0	0	↑1	0	0	↑1	↓1	↑1	0	0	↑1	26	29		3	12%	5	
		Screening	2026-01-07		0	0	↓1	0	0	0	0	0	0	0	↑1	↑1	↑1	↑1	30	33		3	10%	5	
		Visit 10	2026-01-07		0	↑1	0	0	0	0	0	0	↑1	0	0	0	0	0	29	31		2	7%	2	
		Baseline	2026-01-07		0	↓1	0	0	0	0	0	0	↓1	0	0	↓1	0	↓1	28	24		4	14%	4	
		Screening	2026-01-07		↓1	0	0	0	0	0	0	0	0	0	0	0	0	0	9	8		1	11%	1	
		Screening	2026-01-06		0	0	↑1	0	↑1	0	0	↑1	0	0	↑1	0	0	↓1	32	35		3	9%	5	
		Screening	2026-01-06		0	↑1	↑1	0	↑1	0	0	0	0	0	0	0	↑1	0	27	31		4	15%	4	
		Visit 10	2026-01-06		0	0	0	0	0	0	0	0	0	↑1	0	0	0	0	30	31		1	3%	1	
		Screening	2026-01-06		0	0	0	0	0	0	0	0	0	0	0	0	0	↓1	13	12		1	8%	1	
		Visit 7	2026-01-06		0	0	0	0	0	0	0	0	0	0	0	↓1	0	0	14	13		1	7%	1	

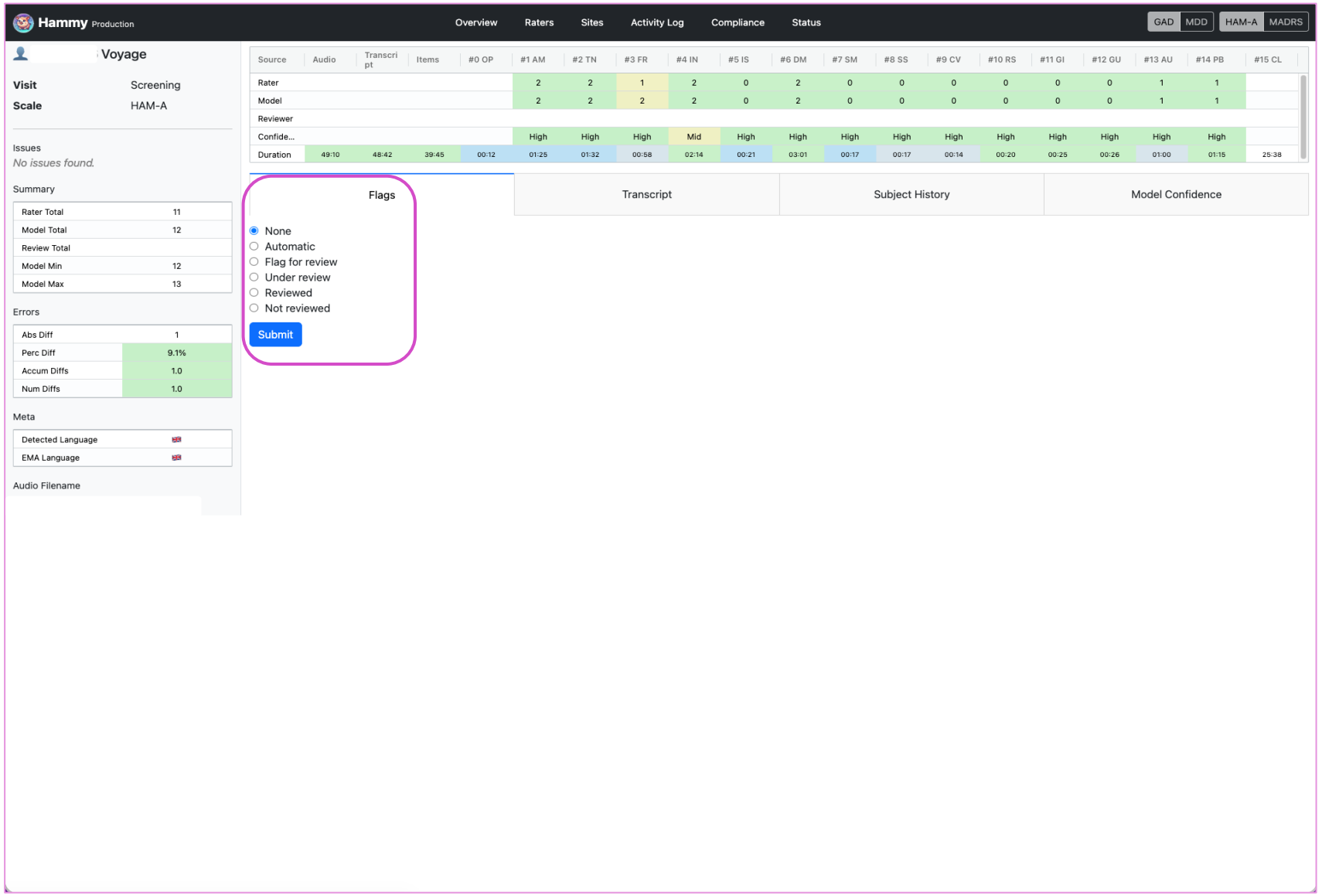
Operationalizing Hammy in ongoing trials



How do Assessments get “Flagged”

Manual Process:

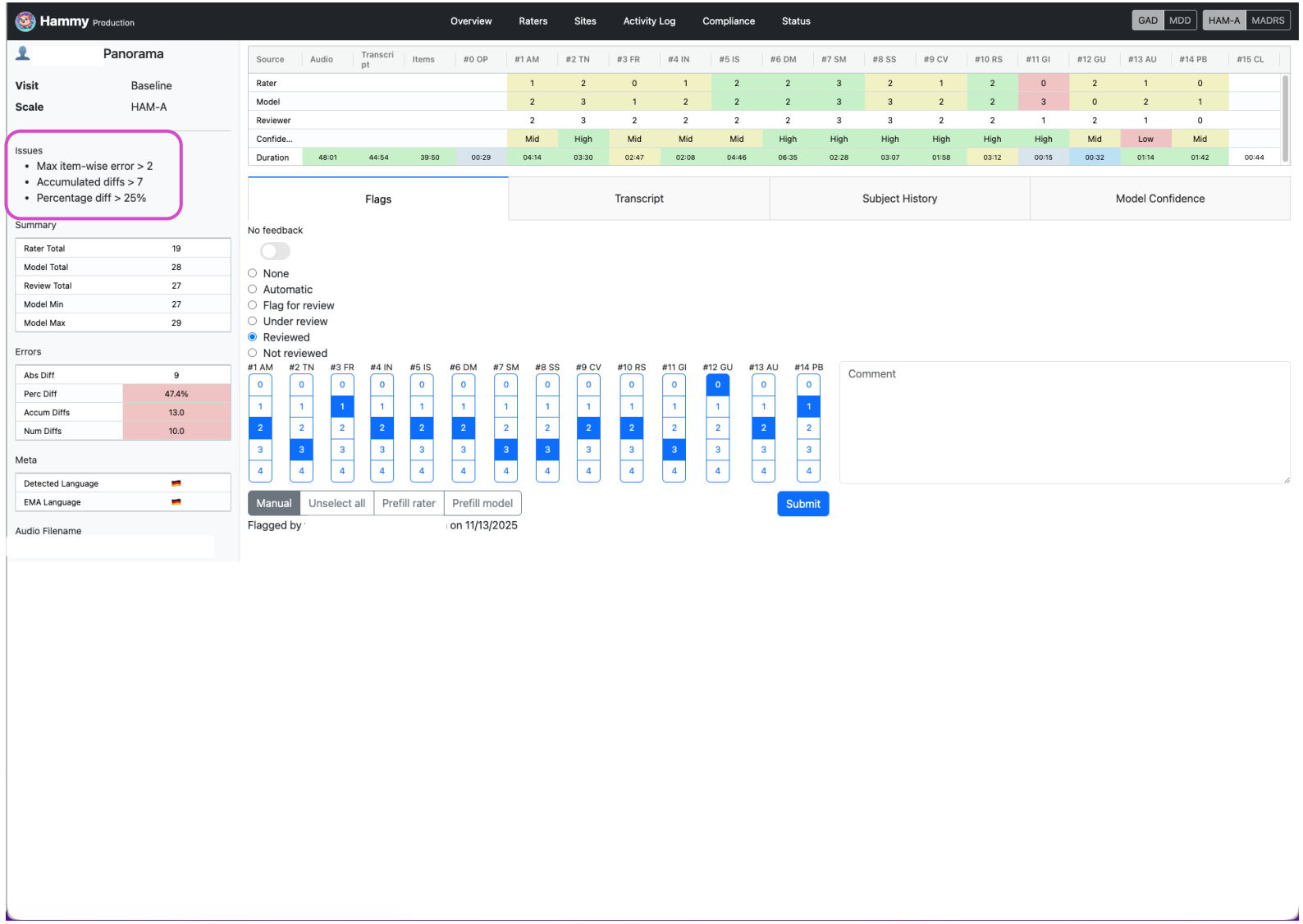
- Central Rater Performance Review
- Data Analytics
- Site Request



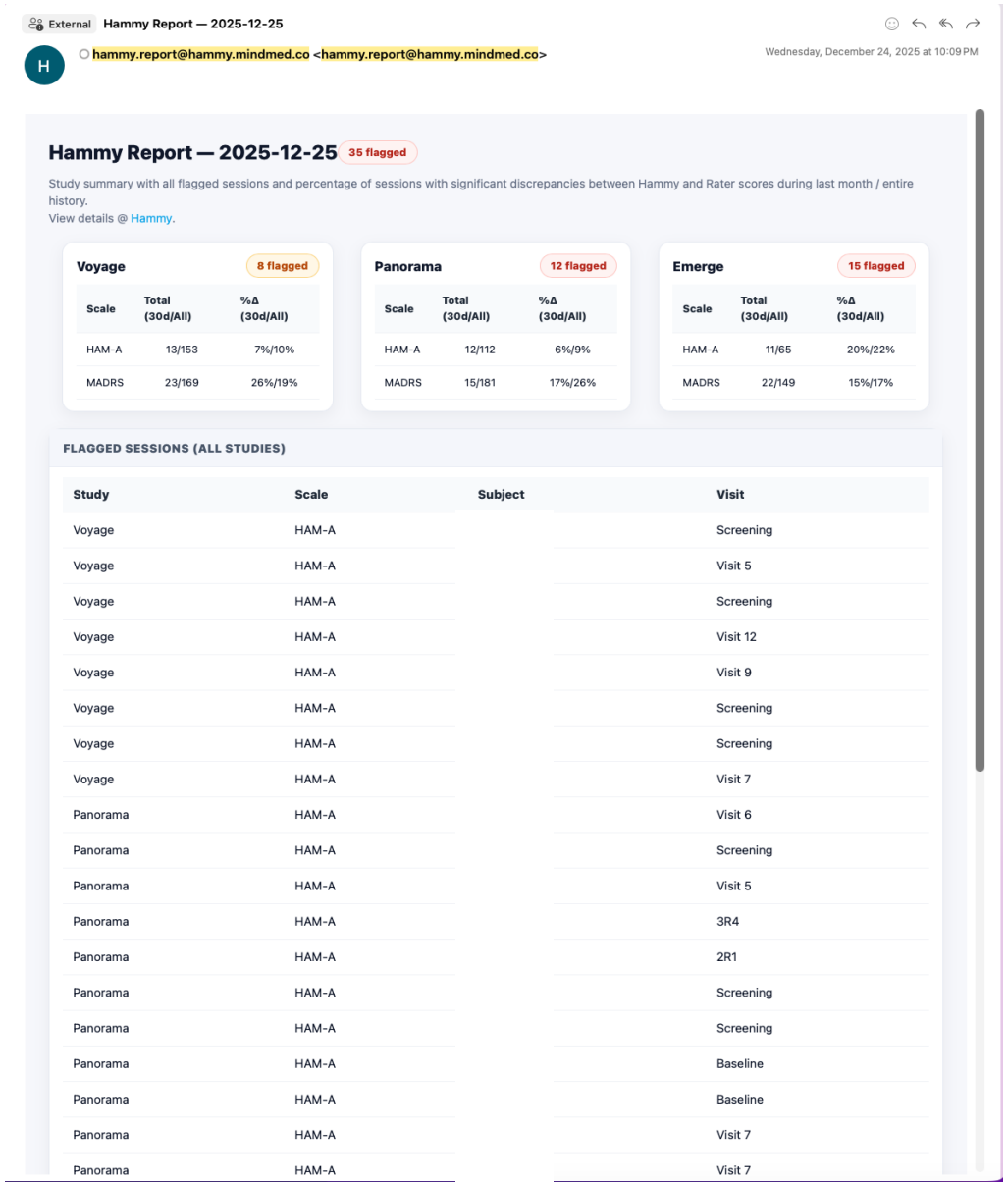
How do Assessments get “Flagged”

Automatic Flagging:

- *A Priori* Conditions Met
- Resulted in 3rd Party Review
- Allows for Triangulation of data points



How do we Monitor?



External **Hammy Report — 2025-12-25**

H hammy.report@hammy.mindmed.co <hammy.report@hammy.mindmed.co>

Wednesday, December 24, 2025 at 10:09 PM

Hammy Report — 2025-12-25 35 flagged

Study summary with all flagged sessions and percentage of sessions with significant discrepancies between Hammy and Rater scores during last month / entire history.
View details @ Hammy.

Voyage 8 flagged

Scale	Total (30d/All)	%Δ (30d/All)
HAM-A	13/153	7%/10%
MADRS	23/169	26%/19%

Panorama 12 flagged

Scale	Total (30d/All)	%Δ (30d/All)
HAM-A	12/112	6%/9%
MADRS	15/181	17%/26%

Emerge 15 flagged

Scale	Total (30d/All)	%Δ (30d/All)
HAM-A	11/65	20%/22%
MADRS	22/149	15%/17%

FLAGGED SESSIONS (ALL STUDIES)

Study	Scale	Subject	Visit
Voyage	HAM-A		Screening
Voyage	HAM-A		Visit 5
Voyage	HAM-A		Screening
Voyage	HAM-A		Visit 12
Voyage	HAM-A		Visit 9
Voyage	HAM-A		Screening
Voyage	HAM-A		Screening
Voyage	HAM-A		Visit 7
Panorama	HAM-A		Visit 6
Panorama	HAM-A		Screening
Panorama	HAM-A		Visit 5
Panorama	HAM-A		3R4
Panorama	HAM-A		2R1
Panorama	HAM-A		Screening
Panorama	HAM-A		Screening
Panorama	HAM-A		Baseline
Panorama	HAM-A		Baseline
Panorama	HAM-A		Visit 7
Panorama	HAM-A		Visit 7

- **Automatic Reports** – emailed at set schedule
  - Tells us where to look based on auto flagging
- **Regular Review of Raters Performance**
  - Bi-Weekly
- **Study Metrics Review**
  - Look at ‘study gestalt’ ratings are a part of the story

Rater vs. Hammy Aggregate Performance

Hammy Production

Overview | Raters | Sites | Activity Log | Compliance | Status

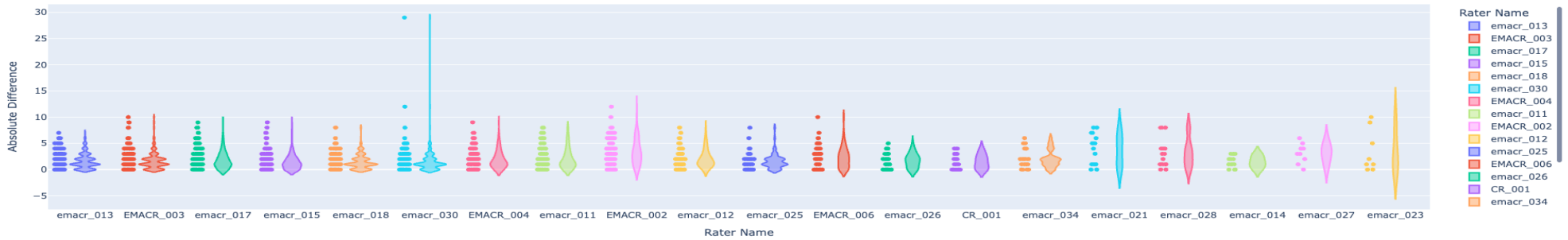
GAD | MDD | HAM-A | MADRS

Start Date → End Date

Rater error averages: coloring based on t-test p-values comparing with 008 data (min 5 sessions) <.05 <.01

Rater Name	Language	Count	Diff	Abs Diff	Accum Diffs	Num Diffs	#1 AM	#2 TN	#3 FR	#4 IN	#5 IS	#6 DM	#7 SM	#8 SS	#9 CV	#10 RS	#11 GI	#12 GU	#13 AU	#14 PB	Exclusion Agreement
		336	0.89	1.64	2.77	2.67	−0.01	0.11	0.09	0.03	0.04	0.01	−0.05	0.08	0.13	0.13	0.05	−0.03	0.12	0.17	0.93
		317	−0.09	1.67	2.66	2.62	−0.04	−0.03	0.03	−0.01	0.03	−0.15	−0.03	0.03	0.08	0.04	0.02	0.01	0.06	−0.13	0.93
		309	1.10	1.82	2.92	2.85	−0.05	0.16	0.08	0.26	0.06	−0.01	0.03	0.11	0.16	0.05	0.11	0.02	0.12	0.00	0.94
		289	0.72	1.39	2.06	2.03	−0.05	−0.02	0.07	0.20	0.00	0.00	−0.02	0.06	0.06	0.05	0.10	0.05	0.09	0.12	0.98
		277	0.28	1.50	2.87	2.83	0.10	0.14	0.07	0.16	0.14	−0.13	0.02	0.03	0.09	−0.15	0.04	0.04	0.08	−0.35	0.94
		275	−0.89	1.55	2.41	2.31	−0.15	−0.15	−0.07	0.01	−0.03	−0.07	−0.10	−0.11	−0.03	−0.04	−0.01	−0.06	0.00	−0.08	0.95
		168	0.53	1.91	2.92	2.86	−0.10	−0.13	0.06	0.10	0.03	−0.15	0.01	0.05	0.08	0.03	0.05	0.10	0.14	0.26	0.94
		160	1.36	2.04	3.42	3.23	0.06	0.06	0.17	0.12	0.16	−0.01	0.13	0.13	0.03	0.01	0.08	0.22	0.10	0.11	0.87
		119	−3.00	3.33	4.49	4.39	−0.29	−0.24	−0.14	−0.18	−0.05	−0.12	−0.32	−0.26	−0.34	−0.32	−0.26	−0.12	−0.28	−0.08	0.87
		88	−0.63	2.08	3.06	2.98	0.02	−0.07	0.07	0.11	−0.07	−0.15	−0.14	−0.01	−0.03	−0.13	−0.05	−0.05	0.08	−0.23	0.90
		73	−0.15	1.55	2.53	2.42	0.07	0.01	0.03	0.07	0.08	−0.11	0.03	−0.10	−0.01	−0.05	−0.03	−0.11	−0.03	0.00	1.00
		72	1.07	2.29	3.40	3.18	−0.04	0.03	0.04	0.35	0.01	0.07	0.15	0.03	0.01	0.03	0.21	0.11	0.14	−0.07	0.94
		39	0.26	1.69	3.74	3.62	−0.10	−0.13	0.03	−0.08	−0.05	0.08	−0.03	0.13	0.13	0.08	0.00	−0.05	0.10	0.15	1.00
		34	−1.21	1.76	3.24	3.18	0.00	−0.06	−0.26	0.12	0.03	−0.41	−0.21	−0.15	0.06	−0.03	−0.06	0.12	−0.06	−0.29	0.93
		26	0.19	2.04	3.50	3.35	−0.08	−0.04	0.00	0.12	0.00	−0.31	0.12	0.19	0.15	0.08	−0.15	0.08	0.12	−0.08	0.92
		17	2.12	3.65	5.29	5.18	0.18	0.18	0.06	0.35	0.12	0.41	0.24	0.00	0.06	0.12	0.18	0.00	−0.12	0.35	0.93
		16	3.13	3.25	4.50	4.44	0.13	0.06	0.31	0.06	0.19	0.31	0.25	0.31	0.38	0.25	0.25	0.25	0.31	0.06	0.93
		13	−0.08	1.46	3.00	3.00	−0.15	−0.15	0.00	−0.08	−0.31	0.15	−0.08	−0.08	−0.08	0.08	0.15	0.08	0.00	0.38	1.00
		9	2.89	3.11	4.22	4.22	0.22	0.22	0.11	0.11	0.22	0.11	0.22	0.33	0.22	0.33	0.22	0.11	0.11	0.33	0.83
		8	2.75	3.50	5.25	4.50	0.38	0.25	0.50	0.25	0.25	0.13	0.13	0.00	0.38	0.25	0.50	−0.38	0.13	0.00	0.83
		2	1.50	1.50	2.50	2.50	0.00	0.50	0.00	0.00	0.00	0.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.50	0.00
		2	2.00	2.00	4.00	4.00	−0.50	0.00	0.50	0.00	0.00	0.50	0.00	−0.50	0.00	0.00	0.50	0.50	0.50	0.50	0.00
		1	2.00	2.00	2.00	2.00	0.00	1.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00
		1	6.00	6.00	6.00	5.00	1.00	0.00	1.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	1.00	2.00	0.00	0.00	0.00

Diff | Abs Diff | Accum Diffs | Num Diffs | #1 AM | #2 TN | #3 FR | #4 IN | #5 IS | #6 DM | #7 SM | #8 SS | #9 CV | #10 RS | #11 GI | #12 GU | #13 AU | #14 PB | Excl Ag



Hammy Case Example: Baseline Exclusion

Panorama

Visit	Baseline
Scale	HAM-A

Issues

No issues found.

Summary

Rater Total	19
Model Total	21
Review Total	20
Model Min	20
Model Max	22

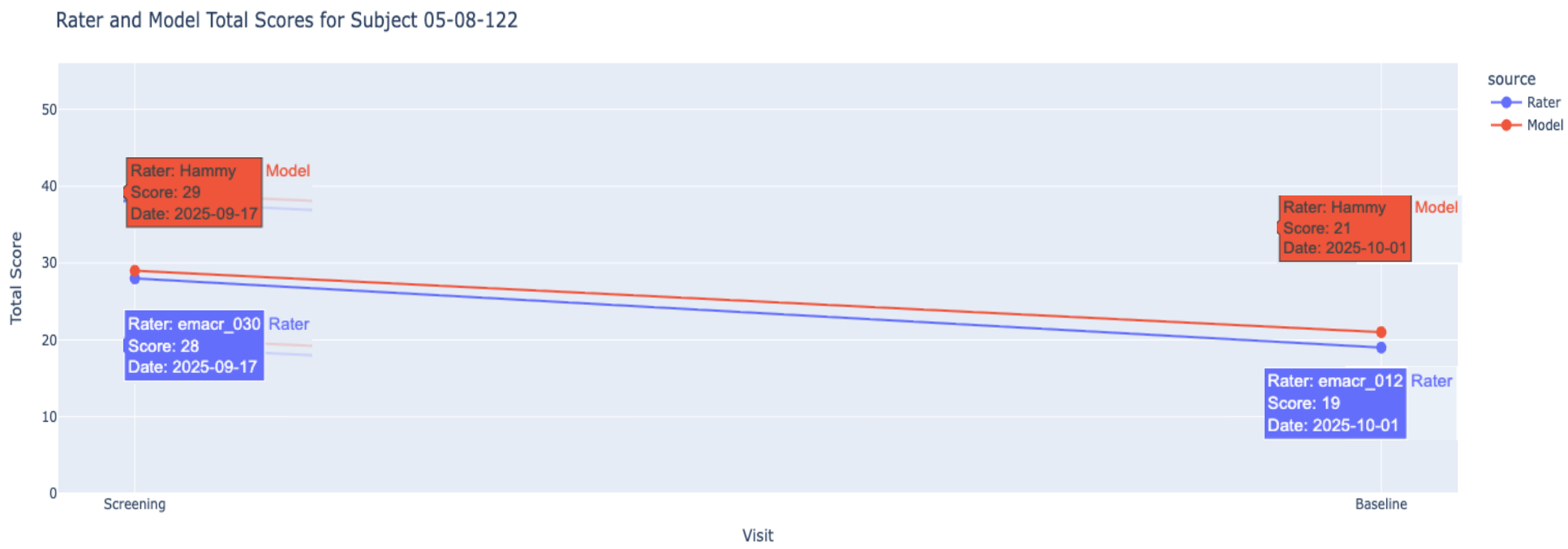
Errors

Abs Diff	2
Perc Diff	10.5%
Accum Diffs	2.0
Num Diffs	2.0

Meta

Detected Language	
EMA Language	

Audio Filename



Hammy Case Example:

Baseline CR: 19
Baseline Hammy: 21
Baseline Reviewer: 20

Source	Audio	Transcript	Items	#0 OP	#1 AM	#2 TN	#3 FR	#4 IN	#5 IS	#6 DM	#7 SM	#8 SS	#9 CV	#10 RS	#11 GI	#12 GU	#13 AU	#14 PB	#15 CL
Rater					3	3	2	2	3	0	2	0	1	1	0	0	0	2	
Model					3	3	3	3	3	0	2	0	1	1	0	0	0	2	
Reviewer					3	3	2	2	3	1	2	0	1	1	0	0	0	2	
Confide...					High	High	Low	High	High	Mid	High	High	Mid	High	High	Mid	High	Mid	
Duration	36:24	36:08	32:18	00:32	04:00	02:21	02:32	05:09	03:54	01:51	02:45	00:29	02:14	00:54	00:39	01:02	00:56	01:13	01:39

Reviewer Feedback:
There were not concerns about the quality of the baseline HAM-A interview that was conducted by CR012. To be thorough and comprehensive, I did look at our recent Central Ratings Data to see if there were any HAM-A overall or individual item scoring trends related to CR012. CR012 has trended at times to rate HAM-A individual items relatively a bit higher compared to other central raters and overall study means. This is some additional data to have when considering the case (because if we saw CR012 typically trending to score lower; then this would lead to some different thoughts regarding the HAM-A baseline/v2 scoring).

Decision: Subject Excluded

Challenges To Implementation:

- Technical Infrastructure
  - How are assessments captured (telehealth, local, etc.)
  - Movement of data/information between interested parties
    - eCOA Provider (audio/data) → Sponsor (flagging/dashboard) → Rater Oversight (decision making)
  - Data Privacy and Access

- Technical knowledge
  - Building and Training Models
  - Structuring and Integrating Data

- Clinical Knowledge
  - Clinical Methodology
  - Appropriate Documentation

- Models are not a Panacea
  - Can be impacted by audio quality
  - Data privacy based on country of origin
  - Models as ground truth → Only as good as the data they are trained on

Acknowledgements

Adam Kolar
Miguel Amvael Pinheiro
Alex Deschamps
Cara Pendergrass
Dan DeBonis
Stephen Saber

Thank You