

Statistical and Practical Considerations of a Large Academic Multicenter Validation Study of Digital Biomarkers in Multiple Sclerosis

BBS Hybrid Seminar, 27.08.2026

Andrea Wiencierz PhD, Study Statistician Validation Study 1

Jannis Müller MD, MSc, Co-Leader Workstream «Digital Future» RC2NB



Mobility



Balance



Dexterity



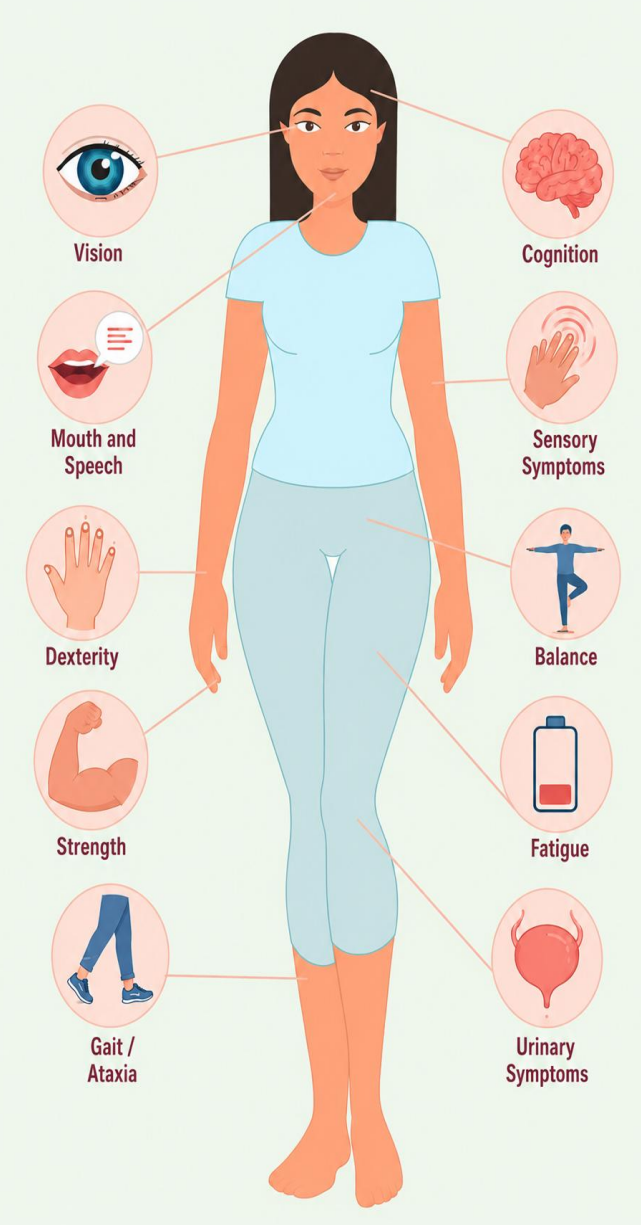
Vision



Cognition



Passive Monitoring

	Domain	Challenge	Duration (s)	Description	Function assessed
	Mobility	Short Promenade	30	Walk for 30 seconds	
		Musical Chairs	30	Sit-to-stand for 30 seconds	
		Spanish Stairs	≈60	Walk up and down stairs	
		Up & Go	60	Stand, walk, turn, return, sit	
		2-Min Walk	120	Walk for 2 minutes	
		U-Turn	60	Repeated 3m walk + U-turns	
	Balance	Butler	60	Hold phone with outstretched arms	
		Penguin	60	Stand still with phone in pocket	
		Walk-a-Line	60	Tandem walking	
	Dexterity	Catch-a-Cloud	50	Tap targets, avoid distractors	
		Confetti	60	Touch nose with screen	
		Draw-a-Shape	60	Trace shapes	
		Pinch Tomatoes	60	Pinch objects on screen	
	Vision	Fog	30	Identify direction of opening	
		Eagle Eye	30	Identify direction of opening	
	Cognition	Numbers	60	Tap numbers in ascending order	
		Mixer	60	Alternate numbers and letters	
		Match Maker	60	Compare current vs previous item	
		Think Fast	60	Sort items into categories	
		Treasure Hunt	≈60	Recall and reconstruct paths	
		eSDMT	120	Symbol-digit matching	

Clinical Studies

-  4 Studies¹⁻⁴, 7 centers
-  350+ PwMS, 160+ HC
-  750+ Neurostatus eEDSS
-  3100+ motor skill reference tests
-  500+ cognitive reference tests
-  200+ XSens acquisitions
-  550+ MRIs
-  235+ OCTs
-  490+ NfL samples
-  120+ other samples (hair, stool)
-  570+ years total follow-up time

RC2NB

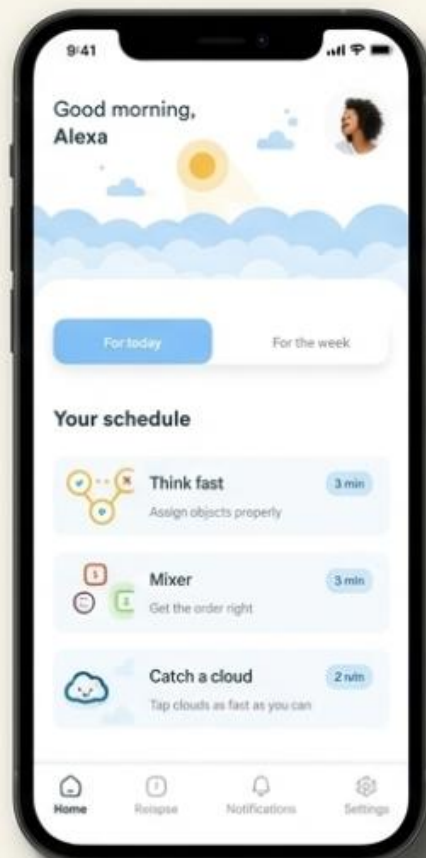
Technical Accomplishments

-  dreaMS App (Version 1.9.2)
-  Integration Floodlight/Konektom
-  4 Languages
-  21 Challenges, 13 PROs
-  1200+ features
-  1 Study Platform (Version 2.8)
-  1 Visualisation Tool
-  166'000+ dreaMS assessments
 - 8'100+ in clinic
 - 152'000+ remote
-  6860+ PROs via dreaMS
-  2240+ days of passive monitoring

RC2NB

 **Indivi**

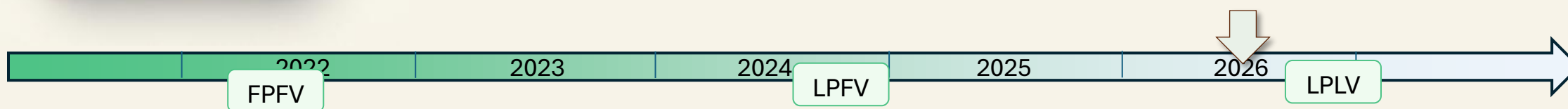
dreaMS Validation Study I



- Swiss-wide study, across 6 academic centers, nested within the SMSC
- March 2022 – December 2026
- 2 years per patient
- Regular schedule (monthly tests) and visit schedule (weekly tests, ± 2 weeks surrounding annual clinical visit)

VS1: Current State

- 🧑 275 pwMS, 50 healthy controls
- ✓ Approx. 240 pwMS concluded
- 💬 Liked by patients: 4.32/5 Likert
- 💓 Relevant for patients: 4.48/5 Likert

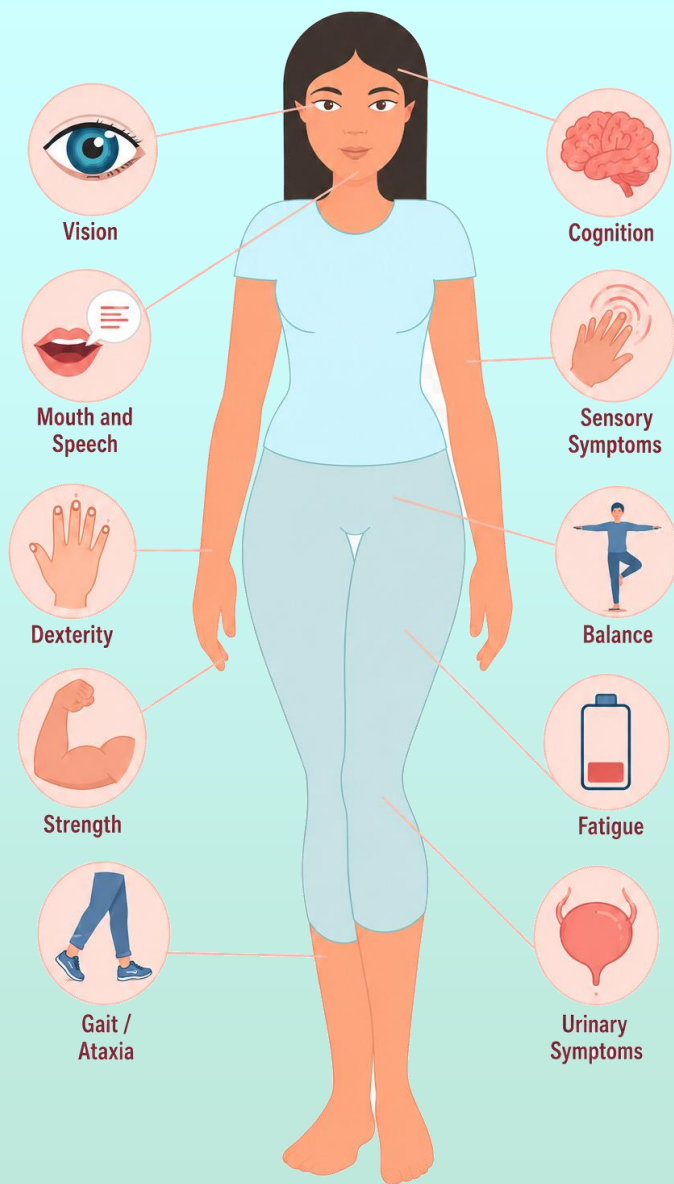


Considerations of General Interest

- Validation Criteria and Feature Selection
- Adherence and Data Completeness
- Context of Use

Considerations of General Interest

- **Validation Criteria and Feature Selection**
- Adherence and Data Completeness
- Context of Use



Domain	Challenge	Features, n
Movement	Short Promenade	247
	Musical Chairs	118
	Spanish Stairs	239
	Up & Go	239
	2-Min Walk	1*
	U-Turn	1*
Balance	Butler	61
	Penguin	46
	Walk-a-Line	46
Dexterity	Catch-a-Cloud	23
	Confetti	125
	Draw-a-Shape	1*
	Pinch Tomatoes	1*
Cognition	Numbers	17
	Mixer	17
	Match Maker	10
	Think Fast	3
	Treasure Hunt	14
	CPS (eSDMT)	2*
	Fog	6
Vision	Eagle eye	6

21 challenges, 1000+ features

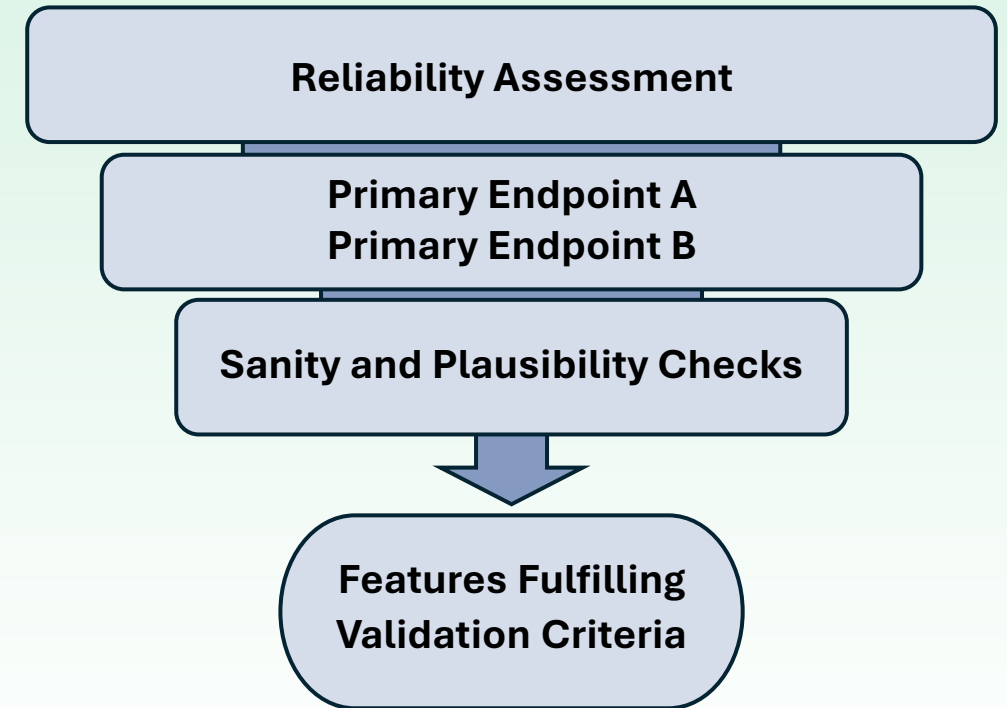
Established Tests
EDSS
EDSS Functional System Score
Timed 25 Foot Walk Test
9-Hole-Peg-Test
High contrast visual acuity
Low contrast letter acuity
SDMT
MRI-metrics
NfL
OCT
ROCF
TMT A & B
FSS
DSB
...

20+ clinical reference tests

Abbreviations: DSB Digit Span Backwards; EDSS Expanded Disability Status Score; FSS functional status score; OCT optical coherence tomography, ROCF Rey-Osterrieth Complex Figure Test, SDMT Symbol Digit Modalities Test, sNFL serum neurofilament light chain, T25FW Timed 25 Foot Walk; TMT A&B Trail Making Test A&B; 9-HPT Nine Hole Peg Test. *Floodlight challenges, with only the most promising feature(s) evaluated

Validation Criteria and Feature Selection

- A classical validation of one new measure against one gold standard is not possible.
- Approach: Evaluate each feature for pre-defined criteria, that we consider desirable:
 - Reliability (precondition)
 - Cross-sectional correlation (primary endpoint A)
 - Longitudinal discrimination (primary endpoint B)
 - Consistency (sanity checks)
- Features fulfilling these criteria are considered valid



Validation Criteria and Feature Selection

*Pre-defined cut-offs of **clinical relevance***

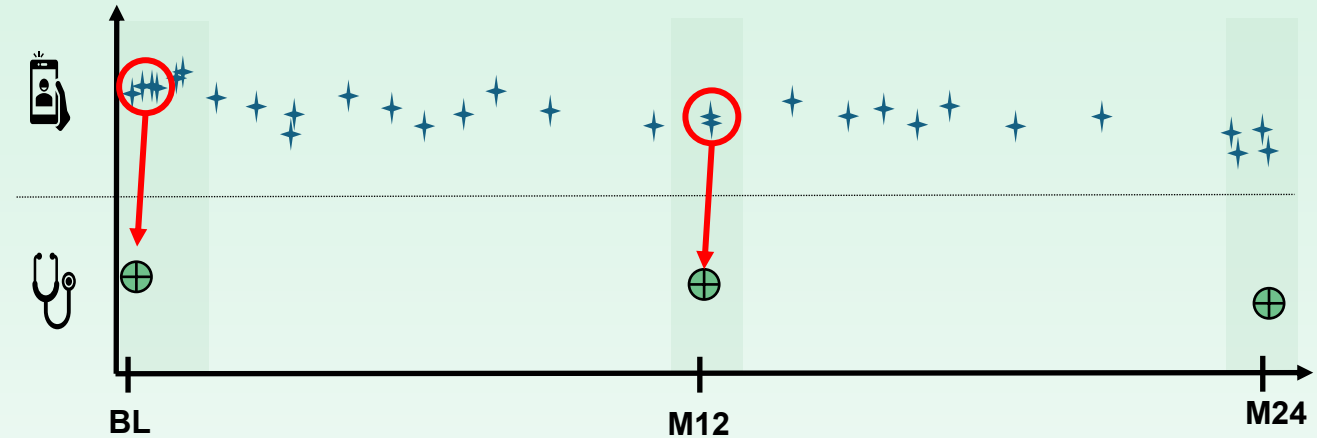
Reference Test	Clinically Meaningful Worsening
EDSS	1.5-point increase (if baseline EDSS = 0); 1-point (if EDSS 1.0-5.5); 0.5-point (if EDSS > 6.0)
FSS, Ambulation score	1-point increase
T25FW	20% increase
9HPT	20% increase
High contrast visual acuity	1-line decrease
Low contrast letter acuity	20% or 7-letter decrease
SDMT	4-point decrease
MRI-metrics	New or enlarging T2 lesions on follow-up scan; Brain atrophy rate < (-0.4%)
sNfL	Higher sNfL z-score on follow-up measurement
OCT	Decreased RNFL thickness on follow-up measurement
ROCF	0.5 lower z-score on follow-up measurement
TMT A & B	0.5 higher z-score on follow-up measurement
DSB	0.5 lower z-score on follow-up measurement

Abbreviations: DSB Digit Span Backwards; EDSS Expanded Disability Status Score; FSS functional status score; OCT optical coherence tomography, ROCF Rey-Osterrieth Complex Figure TestSDMT Symbol Digit Modalities Test, sNfL serum neurofilament light chain, T25FW Timed 25 Foot Walk; TMT A&B Trail Making Test A&B; 9-HPT Nine Hole Peg Test.

Validation Criteria and Feature Selection: Primary Endpoints

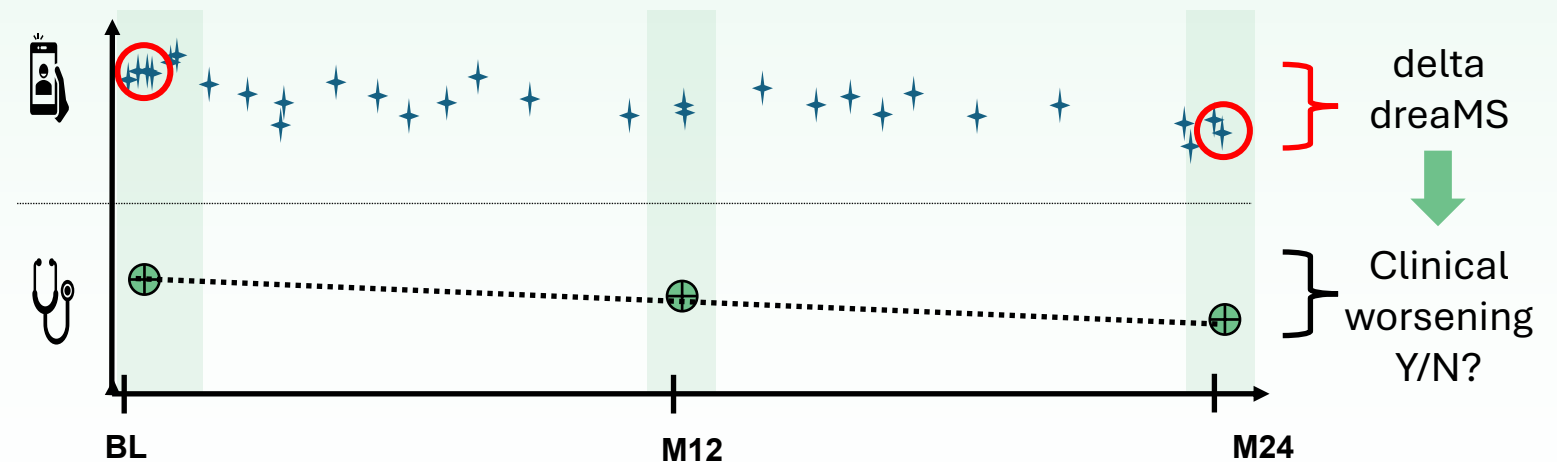
Primary Endpoint A: cross-sectional validity

Spearman correlation at BL and M12.
Validation Criterion: Lower bound of 95% CI > 0.4 (if positive association) or < -0.4 (if negative association) at either BL or M12



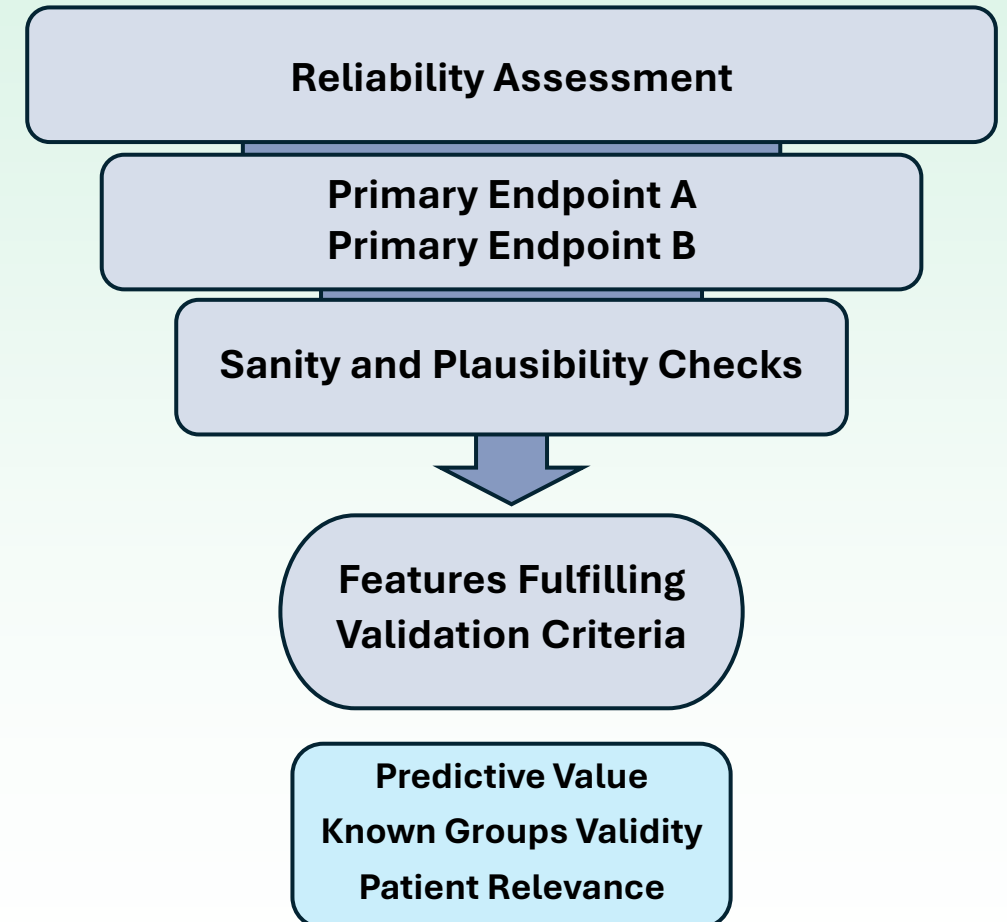
Primary Endpoint B: longitudinal validity

Logistic regression using delta in digital biomarker as predictor, determine AUROC with bootstrap-based CI.
Validation Criterion: Lower bound of 95% CI > 0.6 .



Validation Criteria and Feature Selection

- A classical validation of one new measure against one gold standard is not possible.
- Approach: Evaluate each feature for pre-defined criteria, that we consider desirable:
 - Reliability (precondition)
 - Cross-sectional correlation (primary endpoint A)
 - Longitudinal discrimination (primary endpoint B)
 - Consistency (sanity checks)
- Features fulfilling these criteria are considered valid
- In addition: Evaluate secondary desirable properties, to further refine the set of valid digital biomarkers



Considerations of General Interest

- **Validation Criteria and Feature Selection**
- Adherence and Data Completeness
- Context of Use

Considerations of General Interest

- Validation Criteria and Feature Selection
- **Adherence and Data Completeness**
- Context of Use

Data Completeness

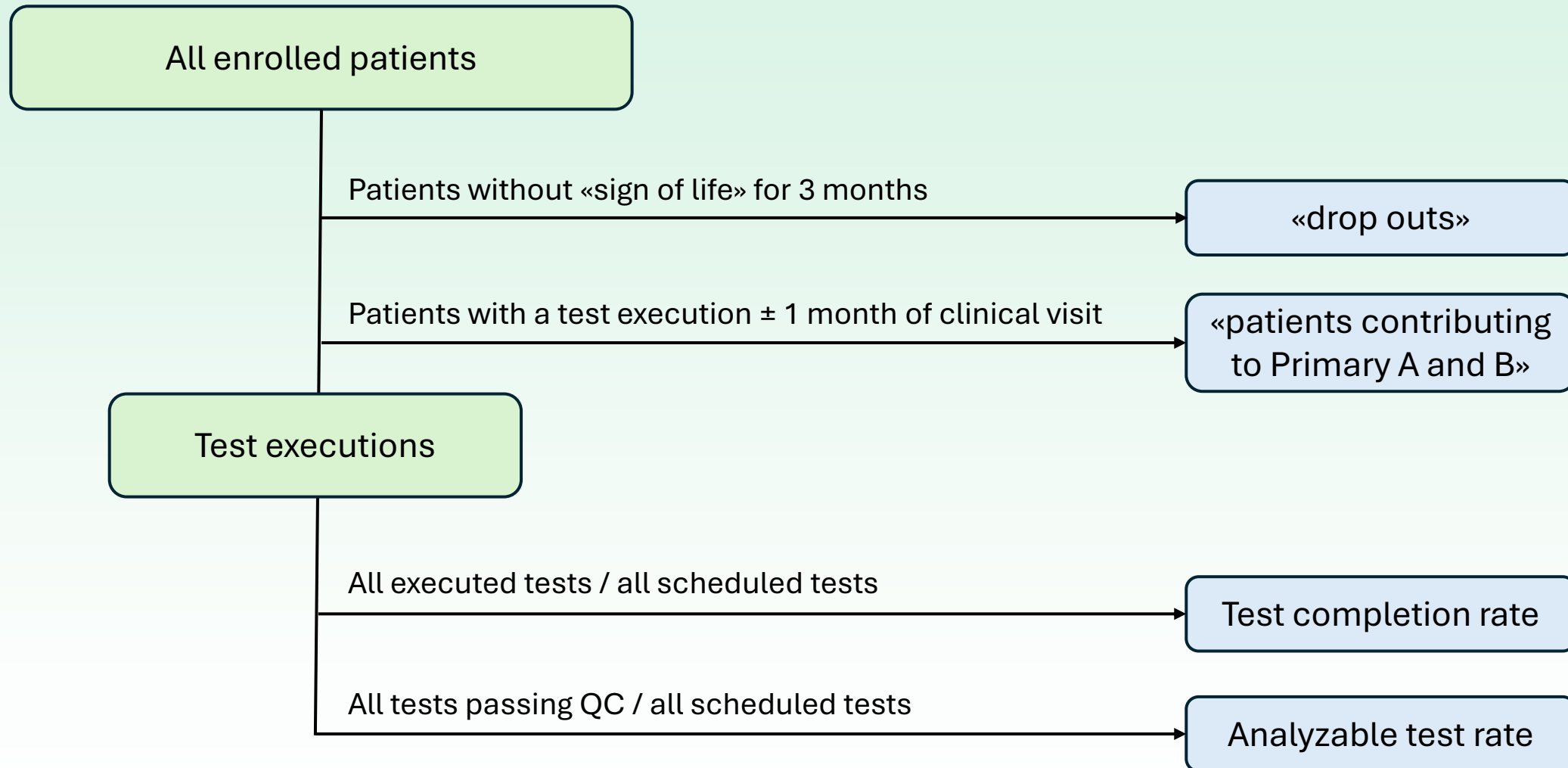
- To fully contribute to the evaluation of primary endpoints, all dreaMS tests need to be performed at least once within ± 1 month around each clinical visit (BL, M12, M24), therefore, dreaMS tests are scheduled more often around clinical visits.
- In practice, data are incomplete for various reasons: missed scheduled dreaMS tests, specific dreaMS tests not done, technical problems of the app, some clinical assessment missing, drop-out etc.
- Data incompleteness $\longleftrightarrow$ lack of adherence ?

Adherence and Data Completeness

- Describe how consistently participants complete what is expected
- Ambiguity in terminology and operationalization: Aherence, retention, persistence, engagement, completion...

Study / program	Concept / unit	Core assessments & schedule	Cohort	Operational definition / metric
Floodlight <i>Midaglia 2019</i>	Adherence / week	Pinching, Draw-a-Shape, balance, U-turn, 2MWT, SDMT; daily to fortnightly	76 pwMS + 25 controls	Active-test adherence = study weeks with ≥ 3 days of completed testing / total study weeks $\times 100$.
MS PATHS <i>Mowry 2020</i>	Completion / module	Processing speed, contrast sensitivity, dexterity, walking	14,643 patients with ≥ 1 assessment	Per-module completion = completed / assigned $\times 100$.
elevateMS <i>Pratap 2020</i>	Compliance / week; retention	Finger tapping, walk/balance, voice DSST, finger-to-nose; weekly	495 pwMS + 134 controls	Minimal weekly compliance = ≥ 1 of 4 scheduled sensor-based active tests completed/week.
CONSONANCE <i>Pointeau 2022</i>	Adherence / rolling week	Hand function, balance and gait daily; cognition weekly	327 of 633 enrolled progressive pwMS	Weekly adherence = compliant days / 7, rowing calculation (Compliance(t) = Compliant days in [t-6, t] / 7 $\times 100$)
Konectom <i>Scaramozza 2024</i>	Adherence / scheduled test	CPS, dexterity, balance/gait; initially daily, then optional	88 pwMS + 70 controls	For CPS: Adherence rate = performed CPS tests / planned CPS tests.
Floodlight Open <i>Oh 2024</i>	Persistence / week	Smartphone active tests + passive monitoring; unsupervised BYOD	1,350 self-declared MS + 1,133 non-MS	Participant persistent in ≥ 1 test per week, reported as e.g. “% remaining active after X days”

Adherence and Data Completeness: Reporting in VS1



Adherence and Data Completeness

Dreams challenge	Test completion rate	Generally liked
short-promenade		
2-min-walk		
musical-chairs		
butler		
penguin		
spanish-stairs		
u-turn		
walk-a-line		
up-and-go		
catch-a-cloud		
confetti		
draw-shape		
pinch-tomato		
eagle-eye		
Fog		
eSDMT		
numbers		
mixer		
Think-fast		
treasure-hunt		
match-maker		

- «What would motivate you to use this App for multiple years»

	Trend	Key Points	Typical User Response (Condensed)
	Enjoyment, Gamification	Game-like exercises increase motivation; gamification desired	"Exercises as games make it fun and engaging."
	Progress Tracking	Demand for feedback on performance and disease status	"I want regular feedback to see if I'm improving"
	Variety and Difficulty	Exercises at times too easy; more cognitive/coordination tasks	"Exercises are repetitive; more variety needed."
	Motivational Systems	Points and badges could boost engagement	"A point system would make me want to keep exercising."
	Perceived Usefulness	Motivation linked to symptom presence	"I do this for research but would be more motivated if it helped me personally"

Considerations of General Interest

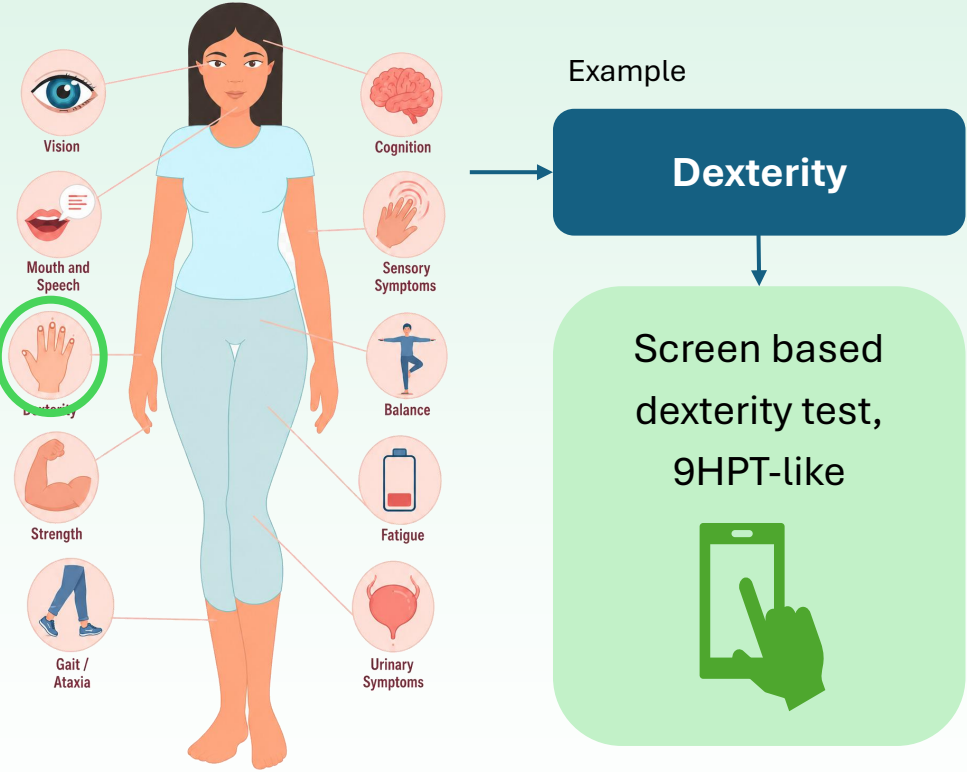
- Validation Criteria and Feature Selection
- **Adherence and Data Completeness**
- Context of Use

Considerations of General Interest

- Validation Criteria and Feature Selection
- Adherence and Data Completeness
- **Context of Use**

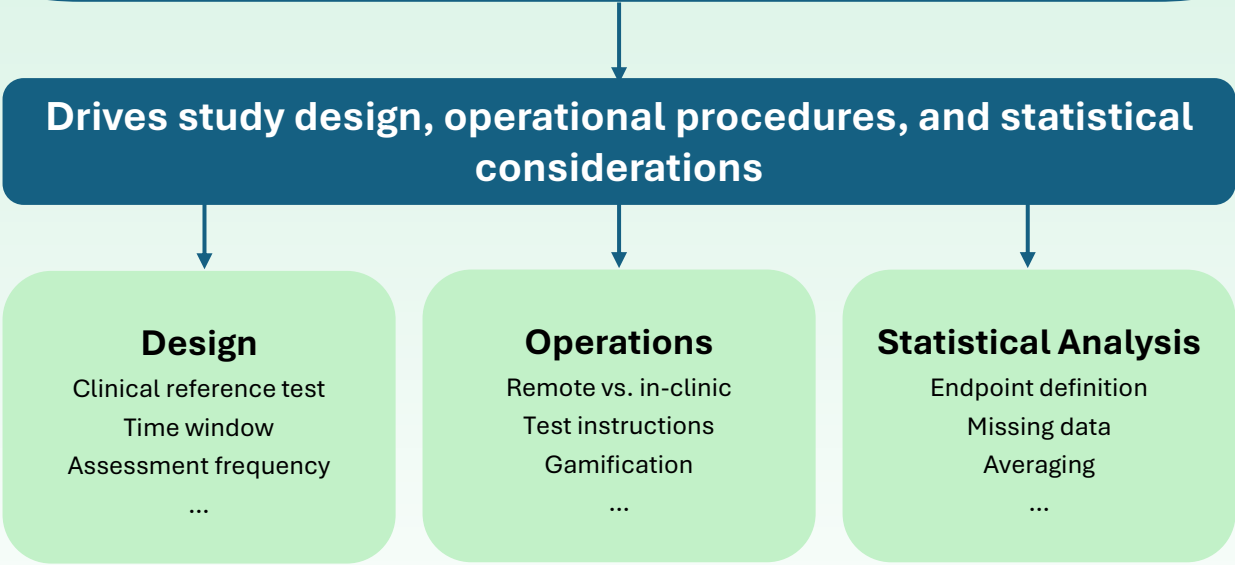
Context of Use

Concept of Interest: The functional characteristic the digital biomarker is intended to quantify



Abbreviations: 9HPT 9-hole peg test.

Context of Use: The setting, population, purpose, and conditions under which a digital biomarker will be used.



Digital Biomarker strengths are context-dependent

Accessibility | Ecological Validity | Sensor Precision | Standardization
Prioritize the strengths that match the intended context of use

Context of Use

		Advantages	Disadvantages
	High-frequency testing	• High temporal resolution • Sensitive to subtle change	• Learning effects • Higher burden
	Low-frequency testing	• Lower burden • Fewer practice effects	• Lower granularity • Changes may be missed
	Supervised testing	• Correct test execution • Standardized conditions	• Resource-intensive • Lower ecological validity
	Remote testing	• Real-world assessment • Scalable and convenient	• Less control • More invalid or missing data
	Active testing	• Standardized tests • Tests measure specific skills	• Tests need to be actively performed
	Passive monitoring	• Lower burden	• Unspecific measurements • Device needs to be worn

- For the purpose of continuous disease monitoring of pwMS, data need to be obtained at a high frequency, but low burden (e.g. use passive monitoring for motor skills instead of active tests). Tests should be suitable for remote monitoring and gamification may be needed to keep monitored subjects engaged.

Averaging to reduce noise is possible and may be useful.

- If the purpose of digital biomarkers is to (in part) replace clinical assessments in routine visits or to be used as endpoints in clinical studies, reliability and correlations with corresponding reference tests need to be very high. Moreover, a supervised setting with standardized conditions might be preferred.

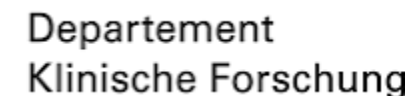
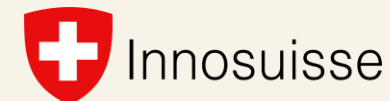
Single measurements to reflect current state.

Conclusion

- DreaMS Validation Study 1 results are expected in 2027
- Establishing clinical validity requires rigorous statistical planning
- Adherence should be reported transparently using complementary measures
- Statistical and practical requirements depend on the intended context of use

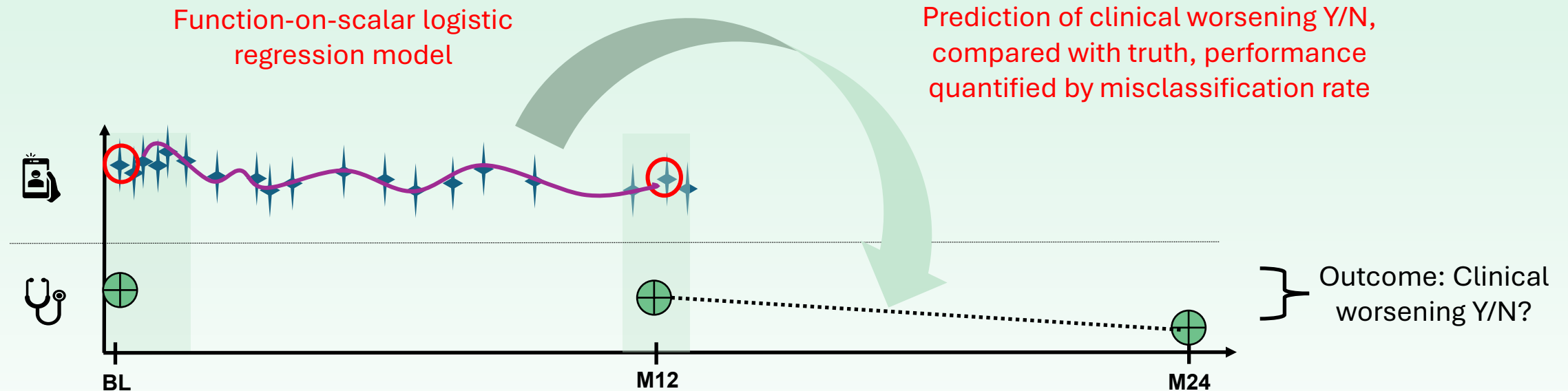
Thanks to all patients and their families.

andrea.wiencierz@usb.ch; jannis.mueller@usb.ch



Example: Predictive Value

Can changes in digital biomarkers during the first year forecast worsening in reference tests in the second year?



- Function-on-scalar logistic regression model using functional principle component analysis of the dreaMS trajectory between BL and M12 to predict worsening in reference test between M12 and M24
- Rationale: Compare model prediction (clinical worsening in future, Y/N?) with clinical truth
- Performance evaluated with misclassification rate (MCR), e.g. $MCR \leq 0.1$ means a correct prediction in $\geq 90\%$ of cases

Validation Criteria and Feature Selection

Test-specific clinical reference tests

Functional Domain	Challenges	Reference Test
Movement	Short Promenade	T25FW, EDSS, ambulation score, pyramidal and cerebellar FSS, MSWS-12
	Musical Chairs	T25FW, EDSS, pyramidal and cerebellar FSS, MSIS
	Spanish Stairs	T25FW, EDSS, ambulation scores, pyramidal and cerebellar FSS, MSWS-12
	Up & Go	T25FW, EDSS, ambulation score, pyramidal and cerebellar FSS, MSWS-12
	2-Min Walk	T25FW, EDSS, ambulation score, pyramidal and cerebellar FSS, MSWS
	U-Turn	T25FW, EDSS, ambulation score, pyramidal and cerebellar FSS, MSWS
Balance	Butler	Pyramidal FSS, EDSS, position test (Romberg), MSIS physical
	Penguin	Pyramidal FSS, EDSS, position test (Romberg), MSIS physical
	Walk-a-Line	T25FW, EDSS, ambulation score, pyramidal and cerebellar FSS, MSWS-12
Dexterity	Catch-a-Cloud	9-HPT, EDSS, MSIS physical
	Confetti	9-HPT, EDSS, cerebellar subscore, upper limb ataxia subscore, MSIS physical
	Draw-a-Shape	9-HPT, EDSS, cerebellar subscore, upper limb ataxia subscore, MSIS physical
	Pinch Tomatoes	9-HPT, EDSS, cerebellar subscore, upper limb ataxia subscore, MSIS physical
Cognition	Numbers	TMT A, SDMT, EDSS, MSIS psychological, FSS
	Mixer	TMT B, EDSS, MSIS psychological, FSS
	Match Maker	DSB, SDMT, EDSS, MSIS psychological, FSS
	Think Fast	SDMT, TMT A, EDSS, FSS, MSIS psychological
	Treasure Hunt	ROCF copy and recall, VLMT, EDSS, MSIS psychological, FSS
	CPS (eSDMT)	SDMT, EDSS, MSIS psychological, FSS
Vision	Fog	Pelli-Robson low contrast chart, EDSS
	Eagle eye	3m visual acuity chart, EDSS

Abbreviations: DSB Digit Span Backwards; EDSS Expanded Disability Status Scale; MSIS Multiple Sclerosis Impact Scale; MSWS-12 Twelve Item MS Walking Scale; ROCF Rey-Osterrieth Complex Figure Test; SDMT Symbol Digit Modalities Test; TMT A&B Trail Making Test A&B; T25FW Timed 25 Foot Walk; 9-HPT Nine Hole Peg Test; VLMT verbaler Lern- und Merkfähigkeitstest.