

Bias Self-Assessment

**BETTER4U · AI FAIRNESS
WORKSHOP · 1 April 2026**

Date: 03.06.2026



Funded by
the European Union



UK Research
and Innovation

Project funded by



Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra
Swiss Confederation

Federal Department of Economic Affairs,
Education and Research EAER
State Secretariat for Education,
Research and Innovation SERI

Bias Self-Assessment

BETTER4U · AI FAIRNESS WORKSHOP · 1 April 2026

This Bias Self-Assessment is a formal record of the fairness risks identified in BETTER4U's AI-based causal model across four stages of development: training and validation data, model design and development, model testing and validation, and evaluation metrics. This report was prepared by WP2 based on the output of the AI Fairness Workshop facilitated by the WP2 leader (UNIVIE) on 1st April 2026. Work packages (WP1 – WP9) were represented by the respective BETTER4U partners who participated in the workshop.

All consortium partners are invited to review the content, correct or add to the risk descriptions, propose additional mitigations, and confirm or update the risk levels based on their knowledge of the project. The document will be updated if necessary.

Risk levels	
H	High — likely harm, not yet addressed
M	Medium — possible harm, partially addressed
L	Low — risk present but well managed

Domain	Risk identified/Risk (Low/Medium/High)	Level	Mitigating measure	Relevant WP(s)/Partner(s)
Training data	Representativeness. Certain groups, like younger children and patients taking obesity management medications are excluded from the RCT. Certain groups are underrepresented in the training data: - older adults (Low) - People of lower SES groups - datasets may generally be of higher SES (High)		Consider widening the scope of future AI model development to include currently excluded and underrepresented groups. Widen analyses to include underrepresented groups.	WP3, WP4

	• Migrants not speaking the local language underrepresented in training data due to lower healthcare system engagement, exclusion from original study cohorts, and fragmented health records across healthcare systems (High) • People with disabilities (High) • People living with higher obesity (Medium)		
Training data	Source integrity Training data reflects past biased or discriminatory decisions relating to people living with obesity affected by weight stigma, lower SES groups, migrants, women, people with poor prior healthcare access or other underrepresented groups (High)	Review datasets to identify variables or labels that may reflect weight stigma or past discriminatory clinical practices (e.g., where a patient's limited engagement with healthcare was recorded as low motivation rather than as a barrier to access). Check whether data gaps are distributed equally across BMI groups or whether patients living with obesity have	WP3, WP4

		systematically less complete records. Consider involving people with lived experience of obesity in reviewing whether the AI model's recommendations reflect their needs and priorities, and whether any labels or categories used feel stigmatising.	
Training data	Data timeliness & recency Current target users whose lifestyles, food environments, digital behaviors, and obesity risk drivers differ from earlier cohorts (Medium).	Use the most recent cohort and feasibility data available. Recalibrate models with recent RCT baseline data, especially digital behavior, diet, physical activity, and social determinants. Repeat fairness checks at interim points.	WP3, WP4
Training data	Relevance Some data collected may not be directly related to the purpose of the project, risking unfair or inaccurate outcomes, in particular, training data skewed toward people without overweight or obesity, normal weight users, not the people living with overweight or obesity (the target group). No ground truth exists to validate predictions (High).	Train and validate specifically in people living with overweight and obesity, not normal-weight dominant datasets. Establish clinically meaningful “ground truth” outcomes using behavior, metabolic health, adherence,	WP5

		and functional outcomes.	
Training data	<i>Problem formulation</i> The measure chosen to represent successful weight loss, namely BMI, might not change even if improvements are made in metabolic health after improved lifestyle habits. A larger BMI change is more beneficial in those living with obesity than in those with overweight. Depending on the population in the RCT, the recommendation for a decrease in BMI needs to be considered (High/Medium).	Consider stratifying BMI targets by baseline weight category (overweight vs obesity) to apply clinically appropriate thresholds for different subgroups, adding complementary monitoring indicators alongside BMI, flagging users for whom BMI is a clinically poor indicator, such as those with certain disabilities or comorbidities, for enhanced healthcare professional review.	WP5, WP7, WP9,
Model Design & Development	<i>Feature selection</i> Included or excluded features may disadvantage certain groups eg. variables that appear neutral may indirectly encode protected characteristics such as ethnicity or socioeconomic status. (Medium).	Seek data from a comprehensive set of vulnerable groups	WP3, WP4, WP5
Model Design & Development	<i>Variable coverage</i> Unclear whether all clinically relevant variables have been captured, including concurrent treatments, socioeconomic constraints, and patient-reported outcomes, which may affect AI model performance for patients	Conduct sensitivity analyses. Obtain datasets including treatments and outcomes.	WP3, WP4, WP5

	with these characteristics (Medium).		
Model Design & Development	Model interpretability The model may be too opaque to detect and address discriminatory patterns (Low).	Consider applying explainability methods to make the model's reasoning explainable to clinical users. The AI model uses causal graphs and simple tree-based prediction algorithms, which are easier to understand than, e.g., neural networks.	WP5
Model Design & Development	Missing data patterns The model may learn missing data patterns as if they were clinically meaningful signals, producing predictions that reflect data collection artefacts rather than genuine health characteristics. Not all model recommendations may lead to BMI reduction (Low).	Consider analyzing missing data patterns across cohorts to distinguish data artefacts from genuine clinical signals and account for these patterns explicitly in model training and validation.	WP3, WP4, WP5, WP6, WP7
Model Testing & Validation	Model transparency The model may be too complex to evaluate whether it produces discriminatory outcomes (Low).	Consider improving model documentation and transparency to enable fairness evaluation across user subgroups. The AI model uses causal graphs and simple tree-based prediction	WP3, WP4, WP5, WP6, WP7

		algorithms, which are more transparent and interpretable than complex approaches such as neural networks, allowing clinical users and researchers to evaluate whether the model produces discriminatory outcomes.	
Model Testing & Validation	<i>Procedural fairness</i> Rules and procedures may not be applied consistently and uniformly to patients who are unable to follow recommendations provided by the model (Low).	Healthcare professional review and validation of model recommendations before delivery to users.	WP5, WP6, WP7
Model Testing & Validation	<i>Non-actionable predictions</i> Model not evaluated on lower SES groups, where recommendations may be unactionable (Low).	Healthcare professional review and validation of model recommendations before delivery to users.	WP5, WP6, WP7
Evaluation Metrics	<i>Group fairness</i> The model performed inconsistently across demographic groups, such as ethnic minority groups or underrepresented older populations (Medium).	Healthcare professional review and validation of model recommendations before delivery to users.	WP3, WP4, WP5, WP7
Evaluation Metrics	<i>Data limitations</i> The retrospective datasets lack sufficient representation of people living with overweight and obesity, have an imbalanced set of dietary variables, and contain incomplete	Consider prioritizing data collection for underrepresented groups through the RCT, with a planned fairness re-	WP3, WP4, WP5, WP9

	demographic data, making it difficult to properly evaluate model fairness across relevant user groups (High).	evaluation as more representative data becomes available.	
--	---	--	--