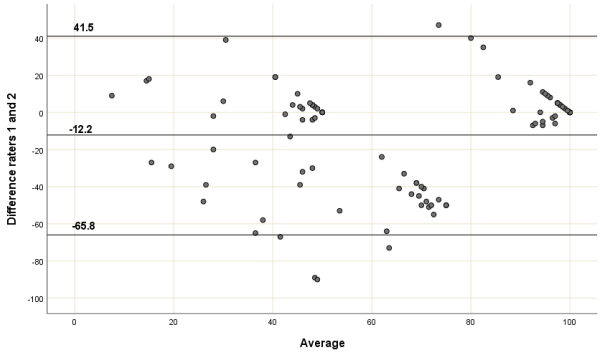
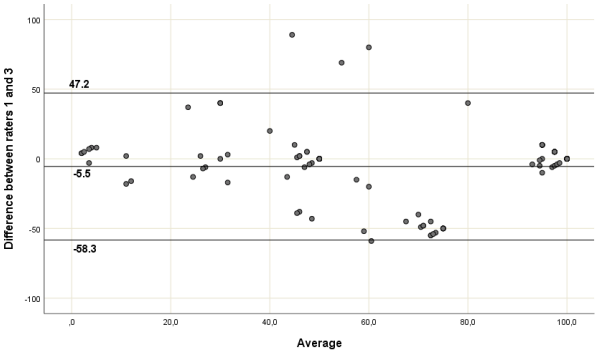
Challenges of Generative AI Human Validation Models in Health Scenarios

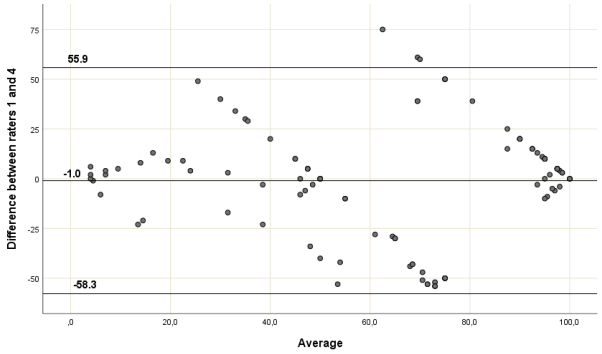
Supplementary files



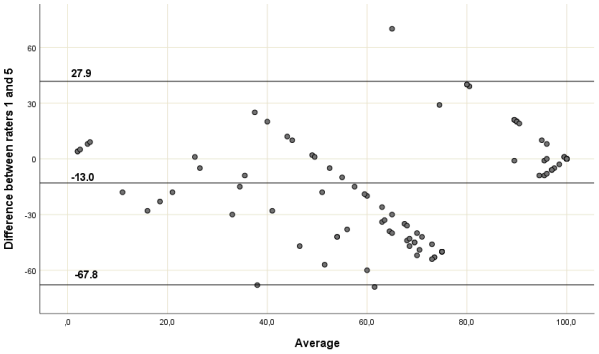
A. Rater 1 versus Rater 2



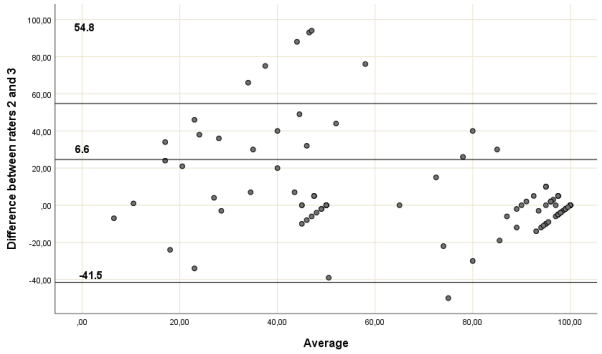
B. Rater 1 versus Rater 3



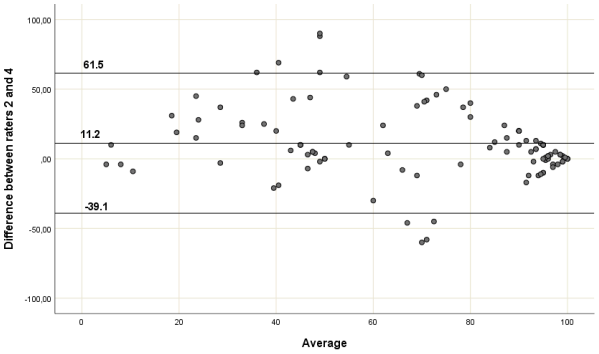
C. Rater 1 versus Rater 4



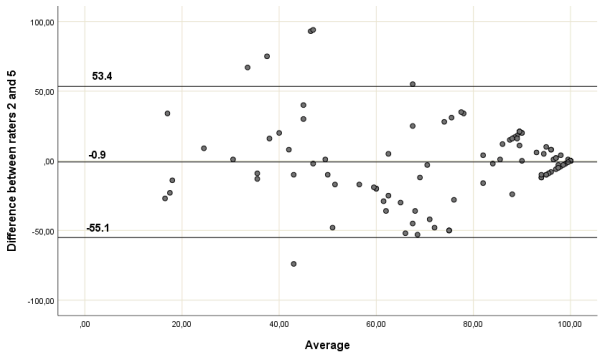
D. Rater 1 versus Rater 5



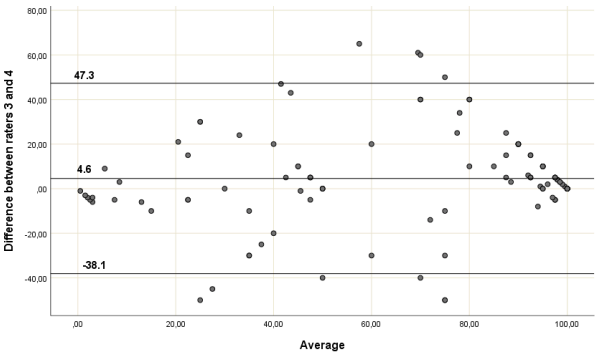
E. Rater 2 versus Rater 3



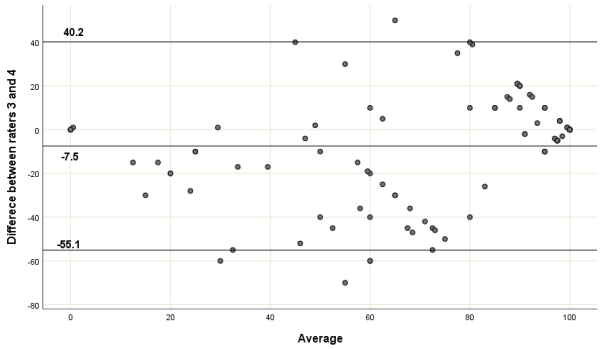
F. Rater 2 versus Rater 4



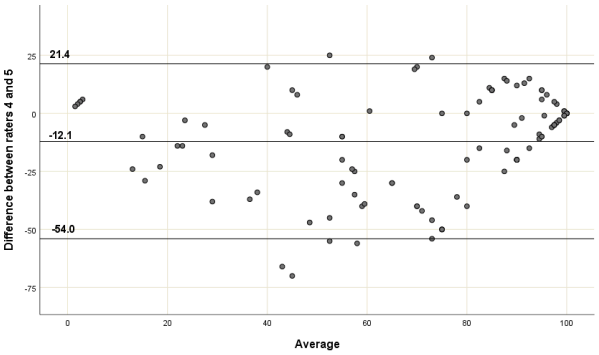
G. Rater 2 versus Rater 5



H. Rater 3 versus Rater 4



I. Rater 3 versus Rater 5



J. Rater 4 versus Rater 5

Supplementary figure. 1. Bland-Altman plots for agreement between two total scores assigned for pairs of raters

Supplementary table 1. Intra-rater variability in the sum of scored evaluations, % [95% CI]

	Rater 1 [n=94]	Rater 3 [n=96]	Total [n=190]
Rater 1 Average	5.23% [-3.4 to 13.9]		
Rater 3 Average	…	3.11% [-5.9 to 12.1]	
Total	…	…	4.2% [-2.0 to 10.4]

Supplementary table 2. Inter-rater variability in the sum of scored evaluations, % [95% CI]

	Rater 1 [n=27]	Rater 2 [n=27]	Rater 3 [n=27]	Rater 4 [n=27]	Rater 5 [n=27]
	Kappa Cohen [95% CI][1] p-value[2]	Kappa Cohen [95% CI][1] p-value[2]	Kappa Cohen [95% CI][1] p-value[2]	Kappa Cohen [95% CI][1] p-value[2]	Kappa Cohen [95% CI][1] p-value[2]
Error type 1: instruction capable of leading to incorrect use of the medication					
Rater 1 [n=27]	1	0.404 [0.106 to 0.727] 0.016	0.587 [0.228 to 0.886] 0.063	0.512 [0.129 to 0.833] 0.219	0.675 [0.372 to 0.919] 0.125
Rater 2 [n=27]	…	1	0.514 [-0.059 to 0.886] 0.625	0.216 [-0.161 to 0.617] 0.453	0.440 [-0.063 to 0.809] 0.375
Rater 3 [n=27]	…	…	1	0.494 [0.036 to 0.867] 1.000	0.494 [0.069 to 0.836] 1.000
Rater 4 [n=27]	…	…	…	1	0.036 [-0.312 to 0.419] 1.000
Error type 2: vague or incorrect instruction					
Rater 1 [n=27]	1	0.502 [0.063 to 0.842] 0.375	0.400 [-0.008 to 0.727] 1.000	0.303 [-0.096 to 0.658] 0.453	0.502 [0.069 to 0.824] 0.375

		Rater 1 [n=27]	Rater 2 [n=27]	Rater 3 [n=27]	Rater 4 [n=27]	Rater 5 [n=27]
Rater 2 [n=27]	…		1	0.625 [0.237 to 0.908] 0.123	0.755 [0.260 to 1.000] 1.000	0.755 [0.292 to 1.000] 1.000
Rater 3 [n=27]	…		…	1	0.625 [0.270 to 0.908] 0.125	0.625 [0.276 to 0.899] 0.125
Rater 4 [n=27]	…		…	…	1	0.509 [-0.052 to 0.886] 1.000
Error type 3: essential information missing						
Rater 1 [n=27]		1	0.274 [-0.035 to 0.562] 0.021	0.349 [0.106 to 0.634] 0.004	0.274 [-0.005 to 0.571] 0.021	0.416 [0.124 to 0.695] 0.079
Rater 2 [n=27]	…		1	-0.025 [-0.295 to 0.377] 1.000	0.571 [0.129 to 0.899] 1.000	0.617 [0.201 to 0.914] 0.625
Rater 3 [n=27]	…		…	1	-0.025 [-0.301 to 0.362] 1.000	0.502 [0.121 to 0.866] 0.375
Rater 4 [n=27]	…		…	…	1	0.426 [-0.38 to 0.787] 0.687
Error type 4: factual, non-medical error.						
Rater 1 [n=27]		1	0.341 [-0.098 to 0.926] 1.000	-0.125 [-0.216 to 0.000] 1.000	-0.059 [-0.140 to 0.000] 0.625	0.000 [0.000 to 0.000] …
Rater 2 [n=27]	…		1	-0.098 [-0.197 to 0.000] 1.000	-0.052 [-0.110 to 0.000] 1.000	0.000 [0.000 to 0.000] …
Rater 3 [n=27]	…		…	1	-0.059 [-0.145 to 0.000] 0.625	0.000 [0.000 to 0.000] …
Rater 4 [n=27]	…		…	…	1	0.000 [0.000 to 0.000] …

	Rater 1 [n=27]	Rater 2 [n=27]	Rater 3 [n=27]	Rater 4 [n=27]	Rater 5 [n=27]
Error type 5: text unsupported by scientific evidence					
Rater 1 [n=27]	1	0.000 [0.000 to 0.000] …	-0.059 [-0.145 to 0.000] 0.625	0.000 [0.000 to 0.000] …	0.471 [0.000 to 1.000] 0.500
Rater 2 [n=27]	…	1	0.000 [0.000 to 0.000] …	0.000 [0.000 to 0.000] …	0.000 [0.000 to 0.000] …
Rater 3 [n=27]	…	…	1	0.000 [0.000 to 0.000] …	-0.038 [-0.098 to 0.000] 1.000
Rater 4 [n=27]	…	…	…	1	0.000 [0.000 to 0.000] …
Error type 6: text failing to address the requested task					
Rater 1 [n=27]	1	-0.052 [-0.110 to 0.000] 1.000	0.780 [0.000 to 1.000] 1.000	-0.080 [-0.161 to 0.000] 1.000	0.000 [0.000 to 0.000] …
Rater 2 [n=27]	…	1	-0.059 [-0.145 to 0.000] 0.625	-0.052 [-0.125 to 0.000] 1.000	0.000 [0.000 to 0.000] …
Rater 3 [n=27]	…	…	1	0.341 [-0.098 to 0.867] 1.000	0.000 [0.000 to 0.000] …
Rater 4 [n=27]	…	…	…	1	0.000 [0.000 to 0.000] …
Error type 7: incorrect information and hallucinations					
Rater 1 [n=27]	1	0.289 [0.000 to 0.710] 0.125	0.867 [0.471 to 1.000] 1.000	0.710 [0.000 to 1.000] 0.500	0.000 [0.000 to 0.000] …
Rater 2 [n=27]	…	1	0.362 [0.000 to 1.000] 0.250	0.471 [0.000 to 1.000] 0.500	0.000 [0.000 to 0.000] …

	Rater 1 [n=27]	Rater 2 [n=27]	Rater 3 [n=27]	Rater 4 [n=27]	Rater 5 [n=27]
Rater 3 [n=27]	…	…	1	0.836 [0.362 to 1.000] 1.000	0.000 [0.000 to 0.000] …
Rater 4 [n=27]	…	…	…	1	0.000 [0.000 to 0.000] …

[1]95% Confidence Interval using 1000 bootstrap sampling. [2]McNemar-Bowker Test.

Supplementary table 3. Intra-rater variability in the sum of scored evaluations, % [95% CI]

	Rater 1 - first assessment (n=48)	Rater 3 - first assessment (n=46)
	Kappa Cohen [95% CI][1] p-value[2]	Kappa Cohen [95% CI][1] p-value[2]
Error type 1: instruction capable of leading to incorrect use of the medication		
Rater 1 - second assessment (n=48)	0.341 [0.130 to 0.600] 0.004	…
Rater 3 - second assessment (n=46)	…	0.623 [0.000 to 1.000] 0.500
Error type 2: vague or incorrect instruction		
Rater 1 - second assessment (n=48)	0.396 [0.158 to 0.673] 0.008	…
Rater 3 - second assessment (n=46)	…	0.465 [0.023 to 0.824] 1.000
Error type 3: essential information missing		
Rater 1 - second assessment (n=48)	0.895 [0.625 to 1.000] 1.000	…
Rater 3 - second assessment (n=46)	…	0.489 [-0.062 to 0.881] 1.000
Error type 4: factual, non-medical error.		
Rater 1 - second assessment (n=48)	1.000	…
Rater 3 - second assessment (n=46)	…	1.000 [1.000 to 1.000] 1.000
Error type 5: text unsupported by scientific evidence		
Rater 1 - second assessment (n=48)	0.500 [-0.071 to 1.000] 1.000	…
Rater 3 - second assessment (n=46)	…	…
Error type 6: text failing to address the requested task		
Rater 1 - second assessment (n=48)	0.455 [-0.091 to 1.000] 1.000	…
Rater 3 - second assessment	…	0.496

	Rater 1 - first assessment (n=48)	Rater 3 - first assessment (n=46)
(n=46)		[-0.095 to 1.000] 1.000
Error type 7: incorrect information and hallucinations		
Rater 1 - second assessment [n=48]	0.407 [-0.081 to 0.833] 0.625	…
Rater 3 - second assessment [n=46]	…	1.000 [1.000 to 1.000] 1.000

[1]95% Confidence Interval using 1000 bootstrap sampling. [2]McNemar-Bowker Test.

A kappa ranging between 0.21-0.40 was considered ‘fair’ agreement, between 0.41-0.60 was considered ‘moderate’ agreement, between 0.61-0.80 be considered ‘substantial’ agreement, and >0.81 be considered ‘almost perfect’ agreement [1].

1. Wan, T. A. N. G., et al: Kappa coefficient: a popular measure of rater agreement. Shanghai archives of psychiatry 27.1: 62 [2015].