Table 2. Overall performance on MH-1813 test data, performance without 1-1-2 training data, and performance on data from 2021 without diagnostic categories as well as performance on MH-1813 based on demographic subgroups (age/sex) [mean (95% CI)].

	F1-score [%] ↑	Sensitivity [%] ↑	PPV [%] ↑	FOR [%] ↓ (1 - specificity)	FPR [%] ↓ (1 - NPV)
Overall					
Call takers	25.8 (23.7–27.9)	52.7 (49.2–56.4)	17.1 (15.5–18.6)	0.105 (0.094–0.116)	0.565 (0.539–0.590)
Model	35.7 (35.0–36.4)	63.0 (62.0–64.1)	24.9 (24.3–25.5)	0.082 (0.079–0.085)	0.419 (0.413–0.426)
Without 1-1-2 training data					
Model	32.4 (31.8–33.1)	60.4 (59.3–61.4)	22.2 (21.6–22.7)	0.088 (0.085–0.091)	0.467 (0.460–0.474)
On MH-1813 data without diagnostic category					
Model	32.6 (31.9–33.4)	48.3 (47.2–49.4)	24.7 (23.9–25.3)	0.153 (0.148–0.158)	0.435 (0.427–0.443)
18-64 years					
Call takers	15.9 (13.1–18.5)	50.5 (43.6–57.2)	9.40 (7.61–11.18)	0.036 (0.028–0.043)	0.353 (0.331–0.375)
Model	22.9 (21.8–24.0)	54.1 (52.1–56.3)	14.5 (13.8–15.3)	0.033 (0.031–0.035)	0.231 (0.226–0.236)
65+ years					
Call takers	32.9 (30.1–35.7)	53.5 (49.4–57.6)	23.7 (21.4–26.0)	0.401 (0.352–0.449)	1.467 (1.373–1.560)
Model	42.8 (41.9–43.7)	66.3 (65.1–67.5)	31.6 (30.8–32.4)	0.290 (0.278–0.303)	1.224 (1.198–1.249)
Male					
Call takers	30.2 (27.2–33.3)	53.9 (49.1–58.9)	21.0 (18.5–23.5)	0.124 (0.105–0.141)	0.542 (0.506–0.580)
Model	39.0 (38.0–40.1)	63.7 (62.3–65.2)	28.1 (27.3–29.0)	0.097 (0.093–0.102)	0.435 (0.425–0.445)
Female					
Call takers	21.9 (19.1–24.6)	51.3 (46.0–56.6)	13.9 (12.0–15.8)	0.090 (0.076–0.103)	0.582 (0.547–0.616)
Model	32.4 (31.4–33.4)	62.3 (60.7–63.8)	21.9 (21.1–22.7)	0.069 (0.066–0.073)	0.407 (0.399–0.416)

Abbreviations: NPV, negative predictive value; PPV, positive predictive value; FOR, false omission rate; CI, confidence interval