How can Major Depressive Disorder be both well-detected by questionaries and have low inter-rater agreement?

The answer may depend on the analyzed cohort, suggesting that PHQ-9 is better as a screen than as a diagnostic tool, and that PHQ-9 based research should take this into account.

The Mystery of PHQ-9 Questionnaires' Predictive Accuracy

Background: The Patient Health Questionnaire (PHQ-9) has been consistently validated as an effective screen for depression, yet field trials of the DSM-5 suggest only moderate inter-rater agreement for Major Depressive Disorder (MDD). Importantly, validation studies focus on general population, whereas the field trials were conducted in psychiatric settings.

Distribution of PHQ-9 scores

Predictive accuracy

AUC for predicting MDD and other types of depression
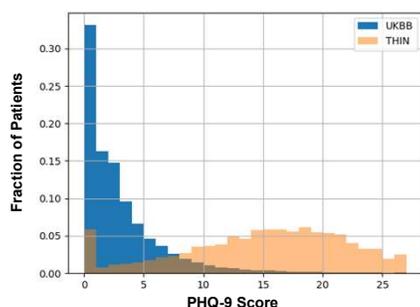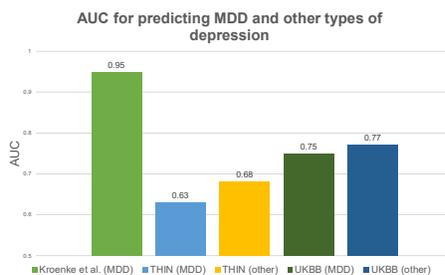
Kroenke K, Spitzer RL, Williams JB. **The PHQ-9: validity of a brief depression severity measure**. J Gen Intern Med. 2001

Number of visits, 90 days before

PHQ-9 range	Kroenke (GP)	Kroenke (OB GYN)	THIN	UKBB
0 - 4	1.0	0.9	3.82	0.29
5 – 9	1.8	0.9	3.56	0.37
10 - 14	2.0	1.3	3.29	0.40
15 – 19	2.4	2.3	3.23	0.52
20 - 27	3.7	2.3	3.37	0.53

It is thought that frequent GP visits are an indication for MDD, typically preceding diagnosis. This is evident in Kroenke et al. and in UKBB, but not in THIN.

Methods

	Cohort	Diagnosis
The PHQ-9 (Validity of a brief depression severity measure)		Survey at GP/OB-GYN clinics. Diagnosis of a sample by phone (allegiance bias?)
THIN		Patients to whom GP decided to administer PHQ-9. Depression deduced from electronic health record.
biobank		Survey of all participants (Healthy, middle aged) Depression deduced from electronic health record.

Limitation: Deducing depressive state from electronic health records is heuristic. UKBB participants are aged 40-70 at recruitment. Bias in THIN diagnosis intuitively strengthens these results (diagnosis depends on score), but perhaps when diagnosis is clear PHQ-9 is not administered.

OHDSI

Yonatan Bilu & Tal El Hay, The KI Research Institute