

Dr. Soliman Fakeeh Hospital — Jeddah

Value-Based Healthcare & Population Health Management Committee

The question we had not been asking

Case-mix adjustment of the PHQ-9 depression score — measuring low mood across the hospital, fairly

2,763 patients • 2,764 questionnaires •



Why a general hospital screens for depression

Depression is not only a psychiatric diagnosis. It is a condition that changes how a patient sleeps, eats, takes medication, attends follow-up and recovers from surgery. It sits underneath almost every outcome we already measure — and until now we were not looking for it outside the psychiatry clinic.

1 — It is usually invisible

Nobody asks, so nobody knows

A patient does not raise low mood in a ten-minute dermatology appointment. A standard form asks the question for us, the same way every time.

2 — This is what value means

Outcomes that matter to patients

Value-based healthcare defines value as the outcomes that matter to patients for the resources we use. A treated body and an untreated mind is not value.

3 — A number we can act on

A trigger, not a statistic

The PHQ-9 has a published threshold and a safety question built into it. Both point to a next step. The measurement only has value if it changes the next consultation.

The PROM we use: the PHQ-9

Over the last two weeks, how often have you been bothered by each of these? **0 = not at all** to **3 = nearly every day**.

What we ask about	Any	Half the days or more
Feeling tired or having little energy	72.5%	38.5%
Trouble sleeping, or sleeping too much	61.2%	33.4%
Little interest or pleasure in doing things	56.6%	23.6%
Feeling down, depressed or hopeless	56.6%	23.8%
Poor appetite or overeating	51.7%	23.3%
Trouble concentrating	41.0%	17.7%
Feeling bad about yourself, or a failure	39.6%	18.2%
Moving or speaking slowly, or being restless	36.7%	16.1%
Thoughts of being better off dead or of hurting yourself	13.9%	5.8%

How it is scored

Add the nine answers. The total runs from **0** to **27**. Higher means more depressive symptoms.

The line that matters

10 or more

The published threshold for clinically significant depression. At or above it, the patient needs assessment, not just reassurance.

Our hospital average

7.1 out of 27, which sits in the mild band — and, as the next slides show, tells you almost nothing on its own.

The clinical picture we place beside it

Every questionnaire was joined, patient by patient, to the clinical and demographic values already held in the hospital system — so what the patient told us about their mood and what the record holds about their body are read together.

MEASURED VALUES FROM THE RECORD

- **Body mass index**
- **Systolic blood pressure**
- **Diastolic blood pressure**

Recorded by clinicians, not asked of the patient.

PATIENT AND CONDITION FACTORS

- **Age**
- **Sex**
- **Diabetes** recorded
- **Cancer** recorded
- **Autoimmune** condition recorded

TWO THINGS WE DELIBERATELY LEFT OUT

Psychiatry visits and **antidepressant use** are in the export. We kept them out of the model.

They describe treatment, not the patient we started with. Adjusting for them would have cancelled out the very thing we are trying to measure — and the test on a later slide shows it would have moved a department by three whole points.

HbA1c was also left out: missing for 2,176 of 2,764 questionnaires.

Where the numbers come from

Everything exported for DSFH Jeddah

2,764 questionnaires · 2,763 patients · 56 departments · ages 18 to 88

All nine questions answered

2,764 questionnaires — not one was incomplete

Clinically plausible measurements only

2,757 questionnaires — 7 removed

Complete clinical record — the cohort we analysed

2,714 questionnaires · 2,713 patients — **98% of everything exported**

What we lost, and why

43 missing BMI

20 missing blood pressure

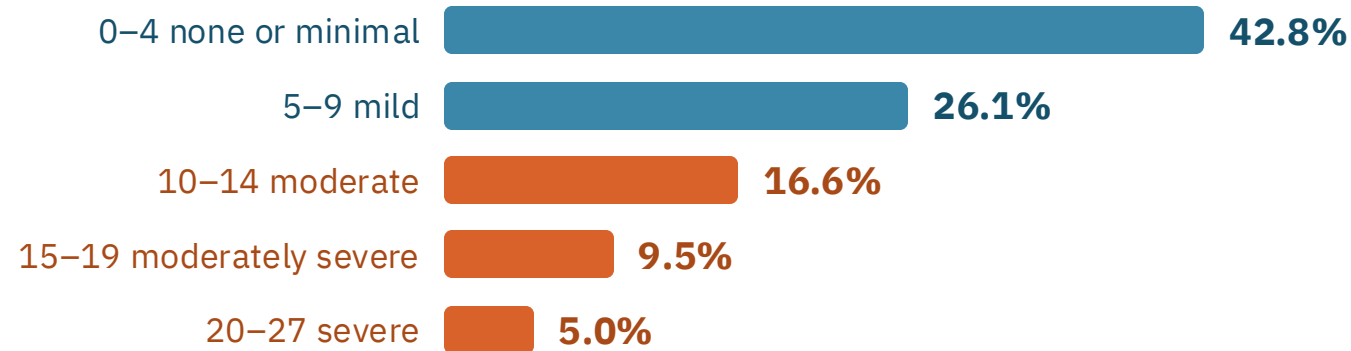
7 implausible measurements

The cleanest of our three PROM datasets by a wide margin.

Who answered

Patients in **56 different departments** — from obstetrics and orthopaedics to dermatology and the emergency department. Psychiatry accounts for 66 of 2,764.

What our patients told us



Bar length is the share of patients in each severity band. The three orange bands are at or above the published threshold of 10 — the patients for whom the questionnaire says assessment is warranted.

31.1%

scored **10 or more** — clinically significant depressive symptoms. That is **843 of our outpatients**, in 56 ordinary clinics.

Why the average misleads

The mean of **7.1** sits in the mild band and sounds unremarkable. One patient in twenty scored 20 or more. Both facts come from the same number.

Question 9 — the one that needs a response

The ninth question asks whether the patient has had thoughts that they would be better off dead, or of hurting themselves. It is the only item on any of our PROMs that is a safety question, and it is answered by patients in every clinic.

13.9%

of patients answered anything other than **"not at all"**
376 questionnaires

5.8%

answered **more than half the days** or **nearly every day**
158 questionnaires

The finding that changes the rule

60 of those 376 patients scored under 10 on the total.

A screening programme that only looks at the total score would have passed over every one of them. Question 9 has to be read on its own, not as one ninth of a sum.

What this slide is not

It is not a count of attempts or of risk. The PHQ-9 is a screening questionnaire; a positive answer means a clinician should ask more, not that a diagnosis has been made.

What it asks of us

A written rule for what happens when a patient answers anything but "not at all" — who assesses, how soon, and how it is recorded. We do not have one today.

Why a raw score is not a fair score

Two factors push the PHQ-9 up more than anything else in our data: **being younger and being a woman**. A clinic whose patients are mostly young women will score higher than a clinic of older men, whatever the care is like. Our obstetrics service is exactly that clinic — and on raw figures it looked like a problem.

If we don't adjust

We blame the wrong service

A team seeing young women would be told its patients are unusually depressed, when the age and sex of its list explains it. An improvement plan gets written for a problem that was never theirs.

If we don't adjust

We miss a real problem

A clinic of older men can look calm while genuinely missing depression in its patients. A favourable case mix hides it, and nobody goes looking.

If we don't adjust

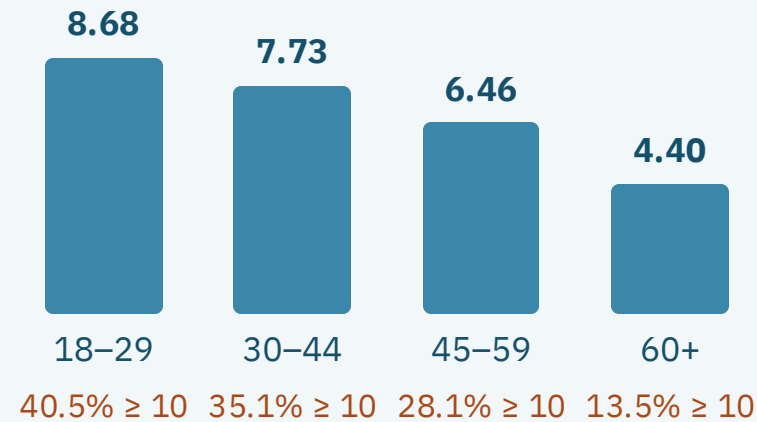
We punish taking hard cases

The most damaging outcome. If a demanding caseload lowers your score, the quiet incentive is to avoid it. Adjustment removes that incentive.

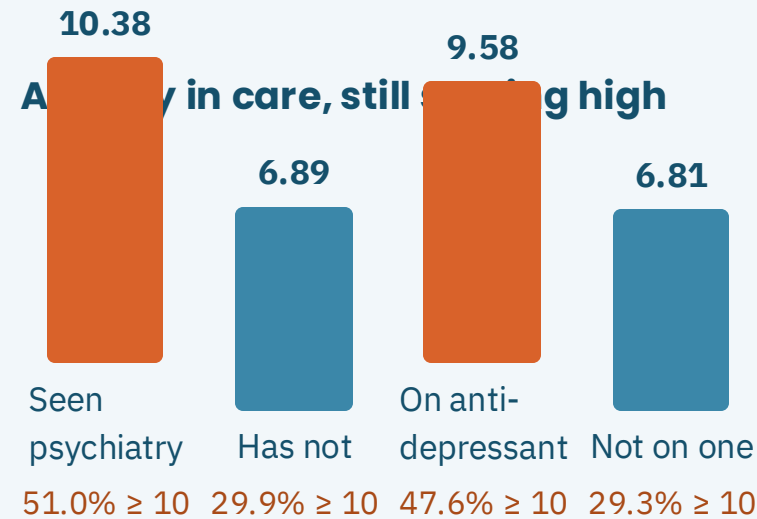
Case-mix adjustment answers one question: **given the patients this service actually sees, is the result better or worse than we would expect?**

Who scores higher

Younger patients, higher scores



Bar height is the average PHQ-9 score. Our youngest patients score twice what our oldest do, and are three times as likely to screen positive.



Same scale as the chart on the left. Half of those already in psychiatric care still screen positive — yet they are a small minority: **84% of everyone who screens positive is not recorded as being in any mental-health care.**

What the model confirms

Each 10 years younger

+0.81 points

Women compared with men

+1.55 points

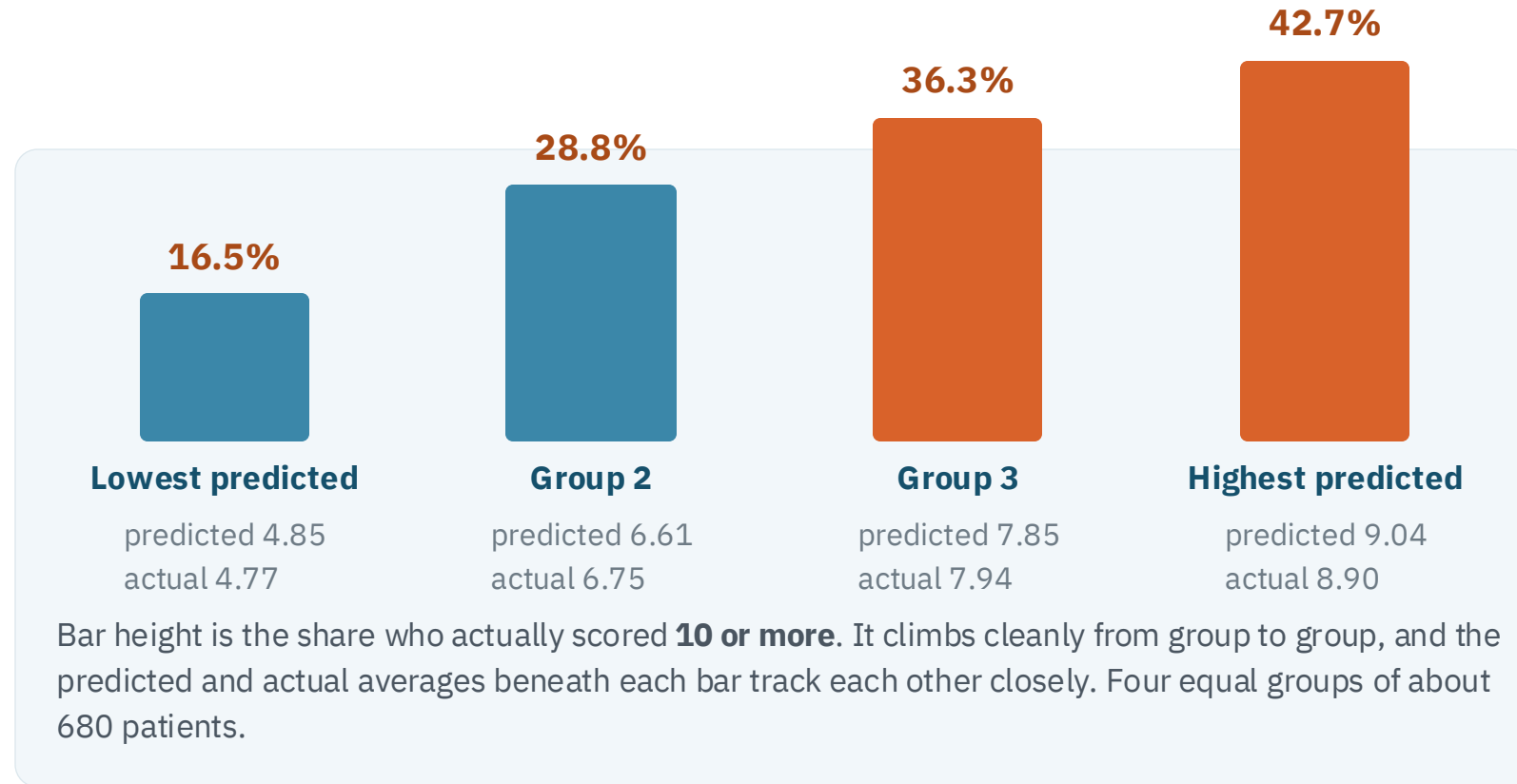
Autoimmune condition

+0.95 points

Weight, blood pressure, diabetes and recorded cancer showed no clear effect. Bars start at zero.

Risk stratification: does the model find the right patients?

We asked the model to predict a score for every patient from age, sex, weight, blood pressure and recorded conditions alone — **without seeing the answers**. Then we sorted patients into four equal groups by that prediction and looked at what they actually reported.



What this means

A patient in the highest-predicted quarter is **two and a half times** as likely to screen positive as one in the lowest — 42.7% against 16.5%.

Endorsement of question 9 follows the same climb: 8.0% in the lowest group, 17.1% in the highest.

And what it does not mean

It cannot tell you which individual patient is depressed. Even in the lowest group, one in six screens positive. It points the effort; the questionnaire still does the

How the adjustment works

OBSERVED

What this service's patients actually reported

—

EXPECTED

What patients like theirs report across the hospital

+

HOSPITAL AVERAGE

7.09 — the common yardstick for everyone

=

ADJUSTED

The fair, like-for-like number we compare

Worked example — Obstetrics & Gynaecology

Step	PHQ-9
What its 307 patients reported	7.47
What patients like theirs report	– 8.48
Hospital average added back	+ 7.09
Adjusted score	6.08

Its patients — almost all women, mostly young — are predicted to score the **highest of any department**. Measured against that, Obstetrics moves from 6th highest on raw figures to **22nd of 24**.

The arithmetic always balances

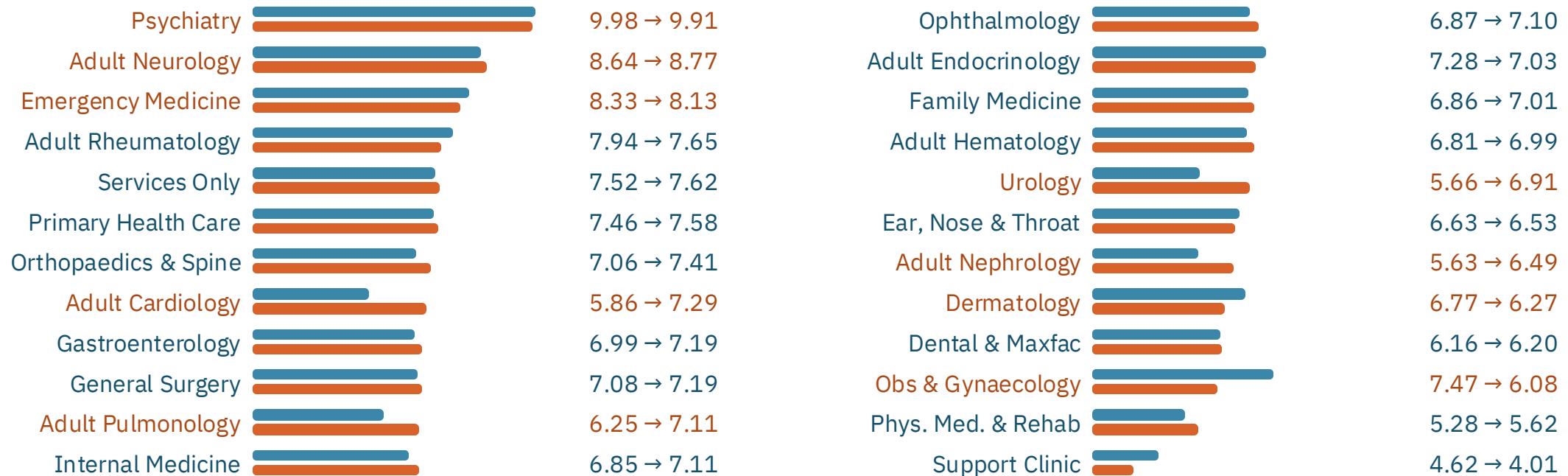
Across the whole hospital, observed, expected and adjusted are identical at 7.09. Adjustment moves symptoms **between** services; it never invents or removes any.

How sure are we?

Every department's range was built by re-drawing the patient list 1,000 times and refitting the model each time, so the uncertainty shown includes the model's own.

Twenty-four departments, before and after adjustment

Observed Adjusted — like-for-like Hospital average 7.09 · shorter bar = fewer symptoms · scale begins at 3.0



The orange labels are the departments that move by four ranks or more. **Obstetrics falls from 6th to 22nd** once the age and sex of its list is allowed for; **Cardiology rises from 20th to 8th**, **Urology from 21st to 17th** and **Nephrology from 22nd to 19th** — older, mostly male services that looked calm on raw figures. Each department shown has at least 30 patients; some names are shortened and the full list of 24 is in the workbook.