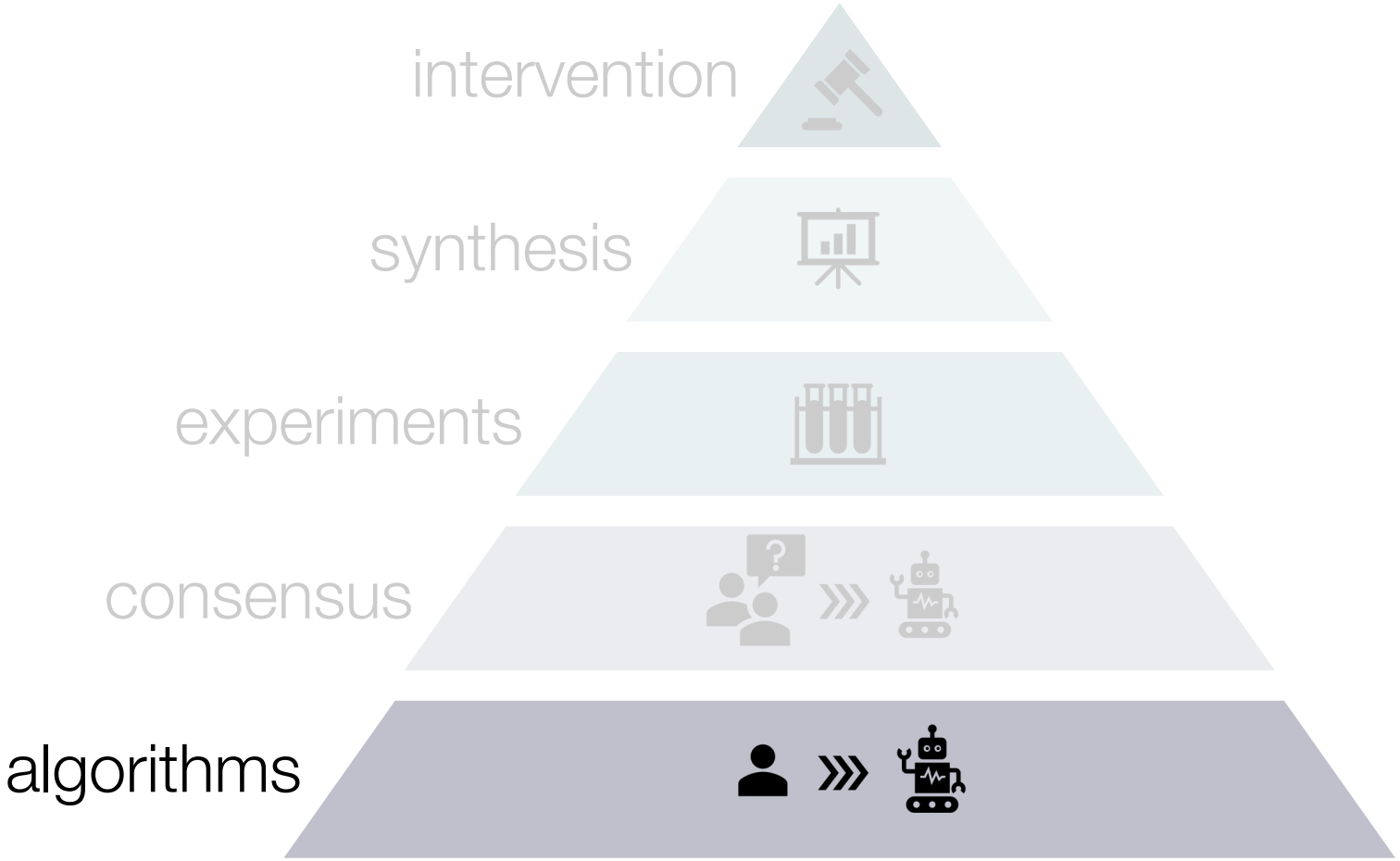
Evidence-based decision making

The power of algorithms

Rui Mata, FS 2026

Version: March 2, 2026

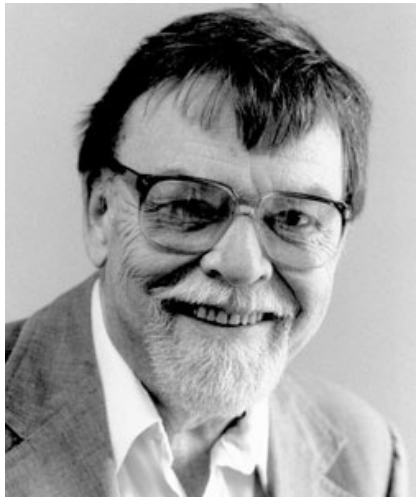

intervention
synthesis
experiments
consensus
algorithms

Goals for today

- distinguish clinical and statistical/actuarial judgment
- understand the lens model and learn associated terminology
- be aware of the sizeable increase in predictive accuracy of statistical/actuarial relative to clinical judgment
- Discuss the power of new technologies (LLMs) for scaling statistical/actuarial judgment

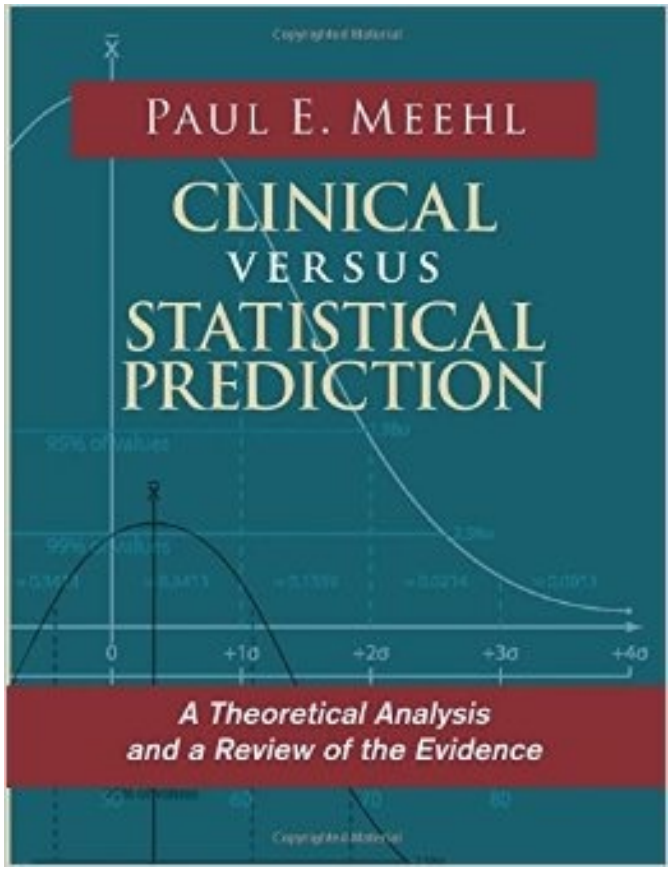
https://www.youtube.com/watch?v=-4QPVo0UIzc

Clinical vs. Statistical Prediction: "A disturbing little book"



Paul Meehl (1920-2003)

- clinician
- psychodynamic orientation
- familiar with projective tests



1954

Meehl, P. E. (1954). Clinical versus statistical prediction: A theoretical analysis and a review of the evidence. Echo Point Books & Media.

Clinical vs. Statistical Prediction: "A disturbing little book"

Meehl looked at ca. 20 studies (depending on inclusion criteria). In all but one case, predictions made by statistical/actuarial means were equal to or better than clinical methods

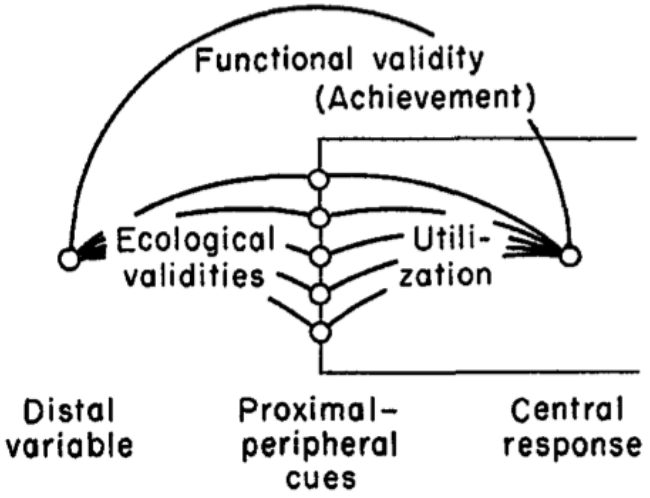
> "…it is clear that the dogmatic, complacent assertion sometimes heard from clinicians that 'naturally' clinical prediction, being based on 'real understanding' is superior, is simply not justified by the facts to date".

Meehl, P. E. (1954). Clinical versus statistical prediction: A theoretical analysis and a review of the evidence. Echo Point Books & Media.

The Lens Model and Policy Capturing

Egon Brunswick
(1903-1955)

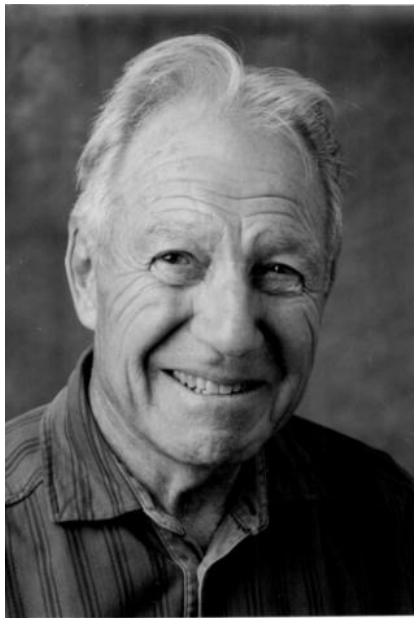
Brunswick proposed the **lens model** which describes the link between characteristics of the world and individual's perception of these characteristics.



Egon Brunswik (born in Budapest, studied in Vienna, later emigrated to USA) argued that psychology should give as much attention to the properties of the organism's environment as it does to the organism itself. He asserted that the environment with which the organism comes into contact is an uncertain, probabilistic one, however lawful it may be in terms of physical principles. Adaptation to a probabilistic world requires that the organism learn to employ probabilistic uncertain evidence (proximal cues) about the world (the distal object). His work has influenced psychology of perception (cf. Roger Shepard) and judgment and decision making (cf. Ken Hammond). His focus on the environment also led him to argue for the need to use representative designs in psychology (i.e., naturalistic sampling of stimuli)

Brunswick, E. (1955). Representative design and probabilistic theory in a functional psychology. ***Psychological Review***, ***62***(3), 193–217. http://doi.org/10.1037/h0046845

The Lens Model and Policy Capturing



Kenneth R. Hammond
(1917-2015)

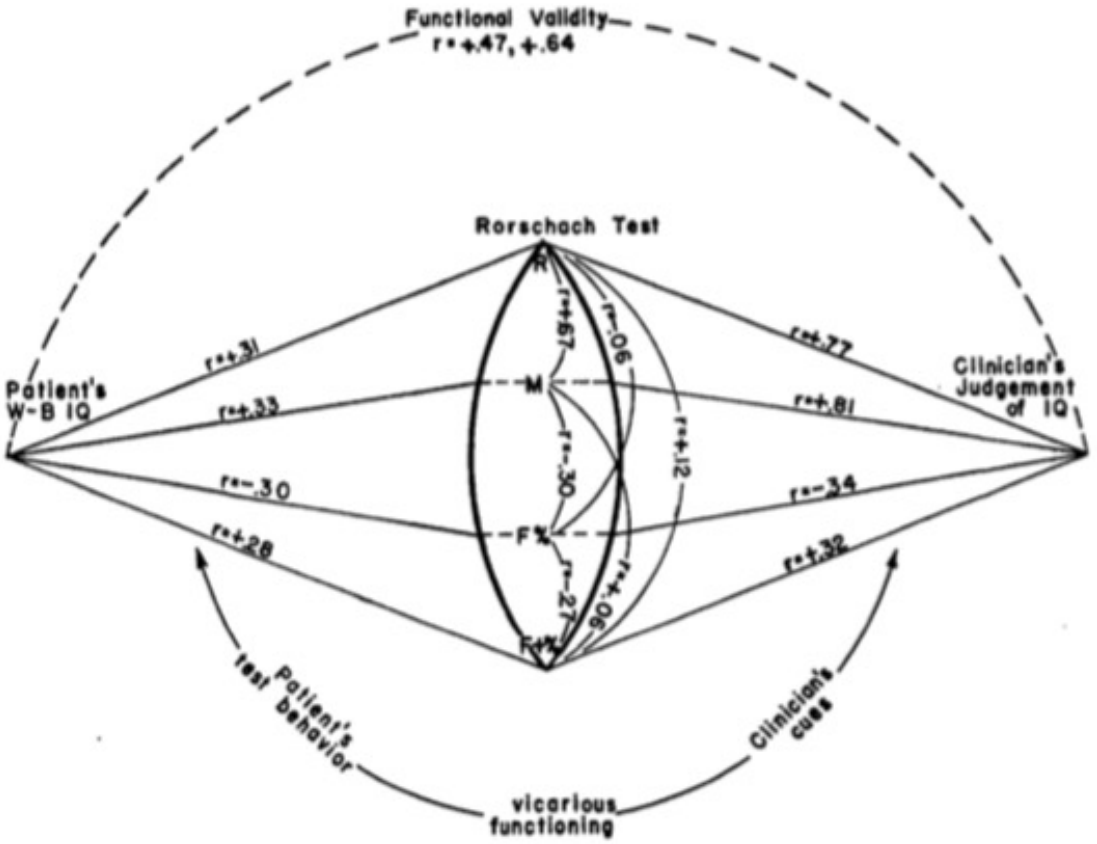


FIG. 1. Functional validity and mediating factors in clinicians' judgments of IQ from the Rorschach test.

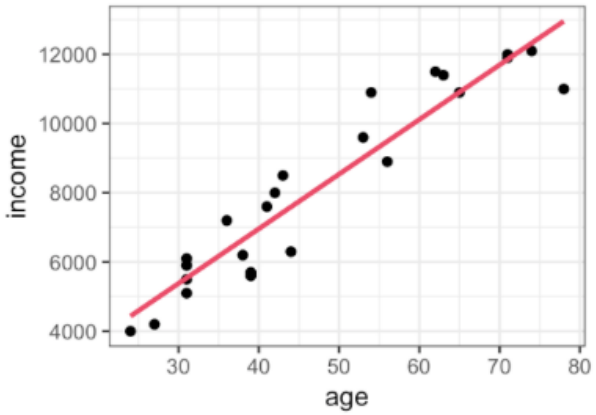
Hammond, K. R. (1955). Probabilistic functioning and the clinical method. ***Psychological Review, 62***(4), 255–262.

Dhami, M.K, & Mumpower, J.L. (2018). Kenneth R. Hammond's contributions to the study of judgment and decision making. (2018). ***Judgment and Decision Making, 13***(1), 1–22.

Simple Linear Regression

Definition: Simple linear regression is a linear model with one predictor x, and where the error term ϵ is Normally distributed.

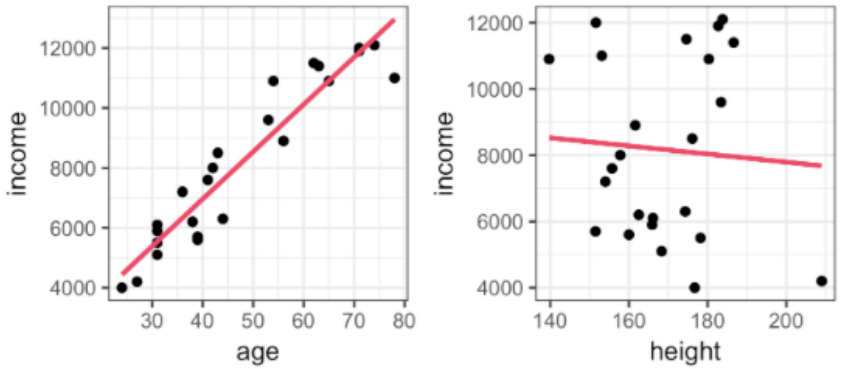
$$y = \beta_0 + \beta_1 x + \epsilon$$



Multiple Linear Regression

Definition: Multiple linear regression is a linear model with many predictors $x_1, x_2, \ldots . x_n$, and where the error term ϵ is Normally distributed.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \epsilon$$



Parameter	Description	In words
β_0	Intercept	When all x values are 0, what is the predicted value for y?
$\beta_1, \beta_2, \ldots$	Coefficient for $x_1, x_2, \ldots$	For every increase of 1 in coefficient for $x_1, x_2, \ldots$ how does y change?

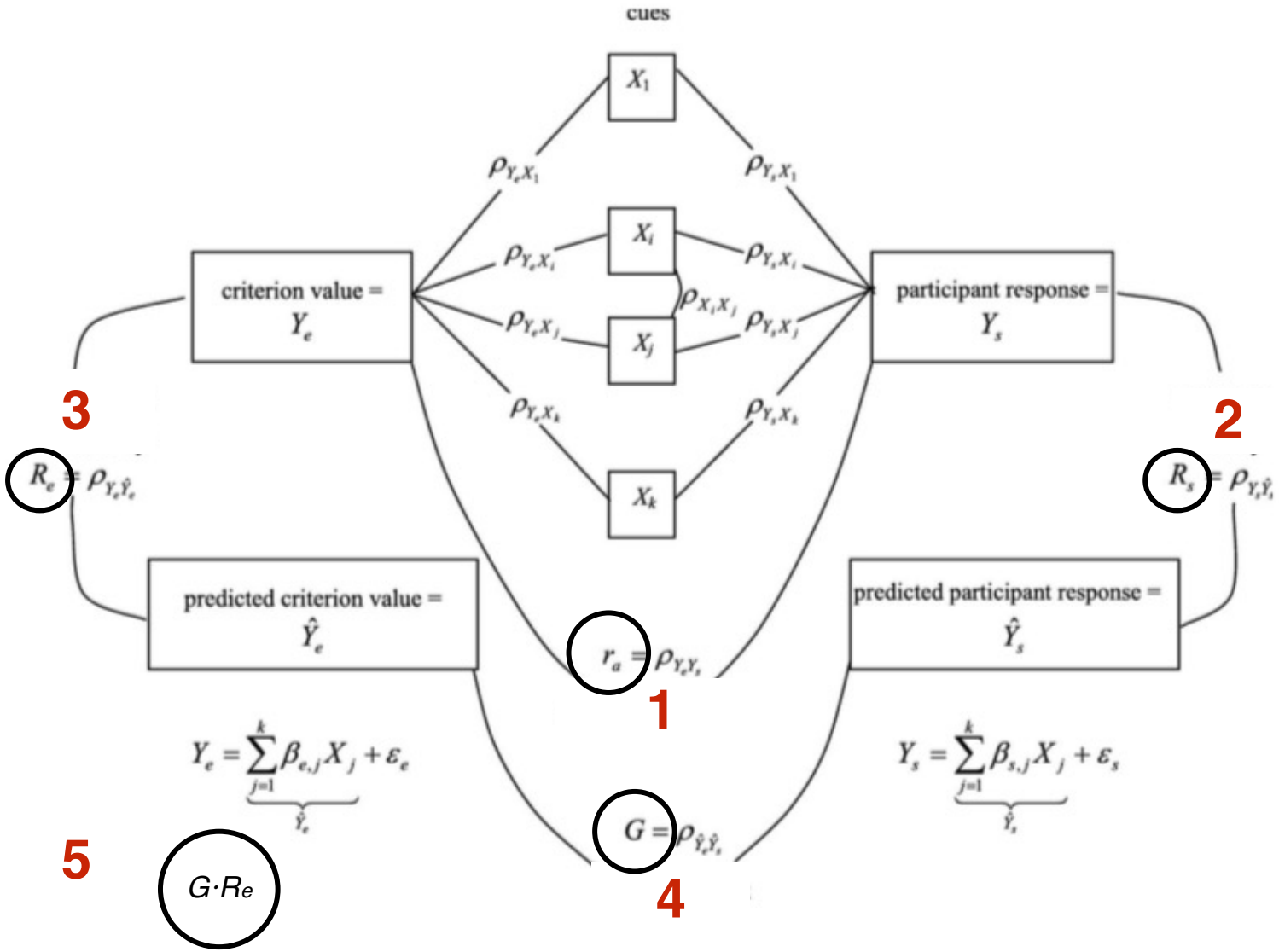
Formula

$$income = 1628 + 147 \times age - 4.1 \times height + \epsilon$$

Coefficients

$$\beta_0 = 1628, \beta_{age} = 147, \beta_{weight} = -4.1$$

The Lens Model and Policy Capturing



Karelaia, N., & Hogarth, R. M. (2008). Determinants of linear judgment: A meta-analysis of lens model studies. *Psychological Bulletin, 134*(3), 404–426. http://doi.org/10.1037/0033-2909.134.3.404

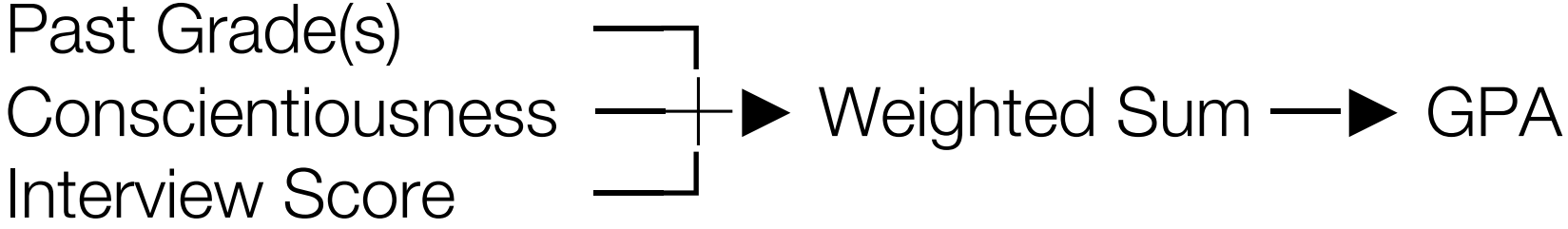
Clinical Prediction: Proper and improper linear models

"Proper linear models are those in which predictor variables are given weights in such a way that the resulting linear composite optimally predicts some criterion of interest; examples of proper linear models are standard regression (…). Research summarized in Paul Meehl's book on clinical versus statistical prediction—and a plethora of research stimulated in part by that book—all indicates that when a numerical criterion variable (e.g., graduate grade point average) is to be predicted from numerical predictor variables, proper linear models outperform clinical intuition. Improper linear models are those in which the weights of the predictor variables are obtained by some nonoptimal method; for example, they may be obtained on the basis of intuition, derived from simulating a clinical judge's predictions, or set to be equal.

Dawes, R. (1979). The robust beauty of improper linear models in decision making. ***American Psychologist, 34***(7), 571–582.

Clinical Prediction: Proper and improper linear models

An example (admissions officer predicting student performance)



Model	How are weights chosen?
Human judge	Chosen intuitively (varies case to case)
Judge model (bootstrapping)	Regression used to predict the judge's ratings
Random model (correct sign)	Random weights, but direction correct
Equal weighting	All weights = 1
Optimal liner model	Regression predicts true GPA

Clinical Prediction: Paramorphic and simple models beat experts!

equal weighting model: models in which weights are 1 but sign is consistent with those from linear model

optimal linear model: linear regression of attributes on criterion

random model: models in which weights were randomly chosen except for sign (which is obtained from linear model)

bootstrapping model: build a paramorphic model of the judge's judgements by linking attributes to the judge's estimated criterion

Correlations Between Predictions and Criterion Values

Example	Average validity of judge	Average validity of judge model	Average validity of random model	Validity of equal weighting model	Validity of optimal linear model
Prediction of neurosis vs. psychosis	.28	.31	.30	.34	.46
Illinois students' predictions of GPA	.33	.50	.51	.60	.69
Oregon students' predictions of GPA	.37	.43	.51	.60	.69
Prediction of later faculty ratings at Oregon	.19	.25	.39	.48	.54
Yntema & Torgerson's (1961) experiment	.84	.89	.84	.97	.97

Note. GPA = grade point average.

"This article presents evidence that even such improper linear models are superior to clinical intuition when predicting a numerical criterion from numerical predictors."

Dawes, R. (1979). The robust beauty of improper linear models in decision making. ***American Psychologist, 34***(7), 571–582.

Clinical Prediction: Experts are inconsistent

Surely we all know that the human brain is poor at weighting and computing. When you check out at a supermarket, you don't eyeball the heap of purchases and say to the clerk, "Well it looks to me as if it's about $17.00 worth; what do you think?" The clerk adds it up. There are no strong arguments . . . from empirical studies . . . for believing that human beings can assign optimal weights in equations subjectively or that they apply their own weights consistently.

It might be objected that this analogy, offered not probatively but pedagogically, presupposes an additive model that a proponent of configural judgment will not accept. Suppose instead that the supermarket pricing rule were, "Whenever both beef and fresh vegetables are involved, multiply the logarithm of 0.78 of the meat price by the square root of twice the vegetable price"; would the clerk and customer eyeball that any better? Worse, almost certainly. When human judges perform poorly at estimating and applying the parameters of a simple or component mathematical function, they should not be expected to do better when required to weight a complex composite of these variables.

Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. ***Science, 243***(4899), 1668–1674.

Clinical Prediction: Experience helps (but not much)

Effects of experience on judgment accuracy in clinical judgment (***d*** ≈ .15)

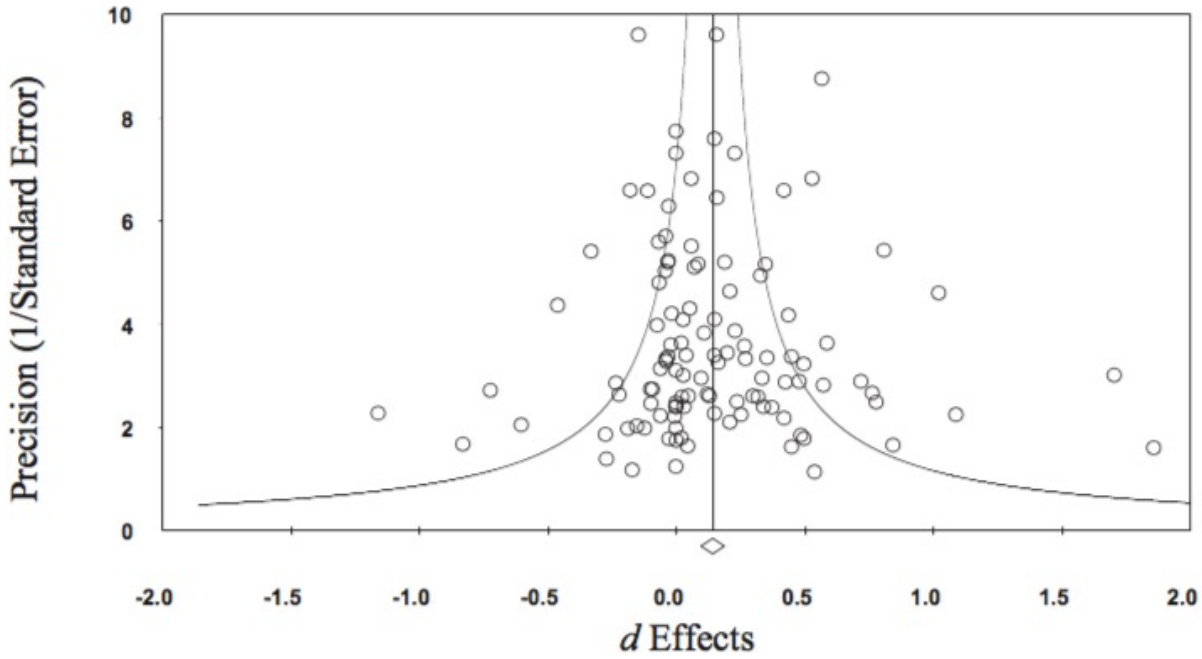


Figure 2. Funnel plot of precision by *d* effects with removal of Garcia (1993), $d = 3.08$.

discussion: heterogeneity across studies in defining "experience" is problematic; overall, effect of experience is small; crucially, moderator analysis that consider extreme groups (e.g., experts/novices) don't find significant differences!

Spengler, P. M., & Pilipis, L. A. (2015). A comprehensive meta-reanalysis of the robustness of the experience-accuracy effect in clinical judgment. ***Journal of Counseling Psychology, 62***(3), 360–378. http://doi.org/10.1037/cou0000065

Clinical Prediction: Experts are not perfectly calibrated

Correlation between confidence and judgment accuracy in clinical judgment ($r \approx .15$)

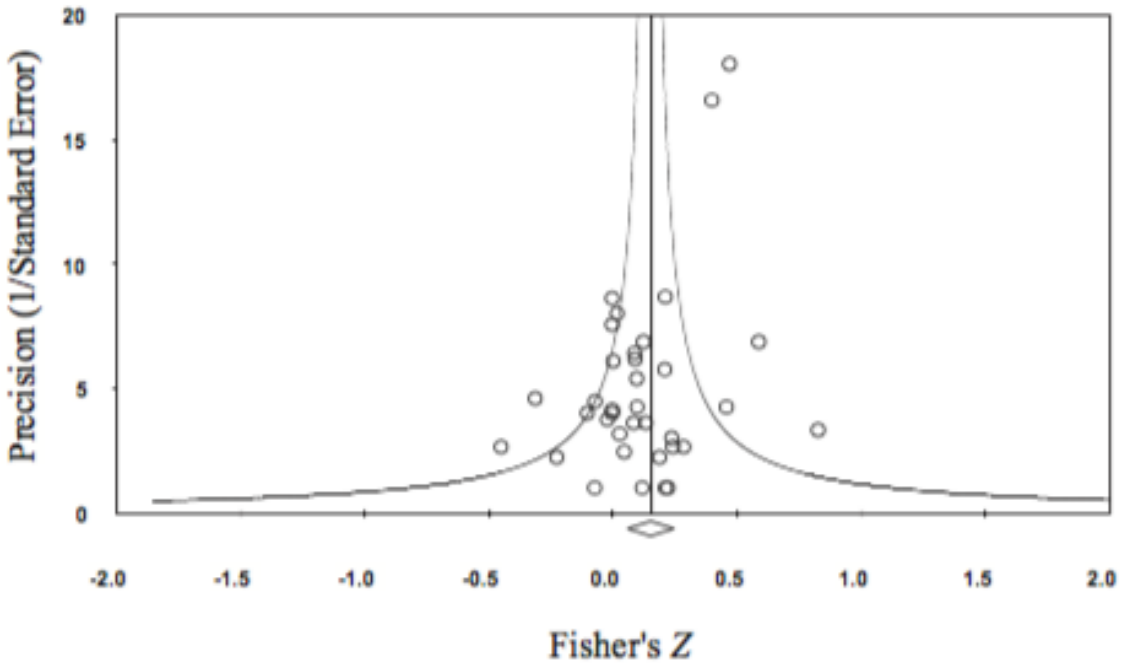


Figure 3. Funnel plot of precision by Fisher's Z.

"(…) an *r* of .15 reflects that confidence accounts for **2%** of variance in judgment accuracy (r^2 = .0225), which by any standard seems inconsequential. If counseling and other psychologists do in reality have the ability to appropriately gauge the accuracy of their own judgments, one would expect the aggregated effect size to be much larger"

Miller, D. J., Spengler, E. S., & Spengler, P. M. (2015). A meta-analysis of confidence and judgment accuracy in clinical decision making. ***Journal of Counseling Psychology, 62***(4), 553–567. http://doi.org/10.1037/cou0000105

Clinical vs. Statistical Prediction: Meta-Analyses

Domain	Improvement of Statistical vs Clinical	Reference
Experimental and field studies	10%	Karelaia, N., & Hogarth, R. M. (2008). Determinants of linear judgment: A meta-analysis of lens model studies. Psychological Bulletin, 134(3), 404–426.
"Human health and behaviour" (e.g., psychology, medicine, forensics, finance)	10%	Grove, W.M., Zald, D.H., Hallberg, A.M., Lebow, B., Snitz, E., & Nelson, C. (2000). Clinical versus mechanical prediction: A meta-analysis. ***Psychological Assessment, 12***, 19–30.
"Counseling psychology" (e.g., diagnosis, prognostic in therapy)	13%	Ægisdóttir, S., White, M. J., & Spengler, P. M. (2006). The meta-analysis of clinical judgment project: Fifty-six years of accumulated research on clinical versus statistical prediction. ***The Counseling Psychologist, 34***, 341-382.
"Employee selection and academic admission"	20% (median)	Kuncel, N. R., Klieger, D. M., Connelly, B. S., & Ones, D. S. (2013). Mechanical versus clinical data combination in selection and admissions decisions: A meta-analysis. ***Journal of Applied Psychology, 98***(6), 1060–1072.

Advantages of Statistical/Actuarial Prediction

Immunity from fatigue and other limitations (forgetfulness, over-confidence)

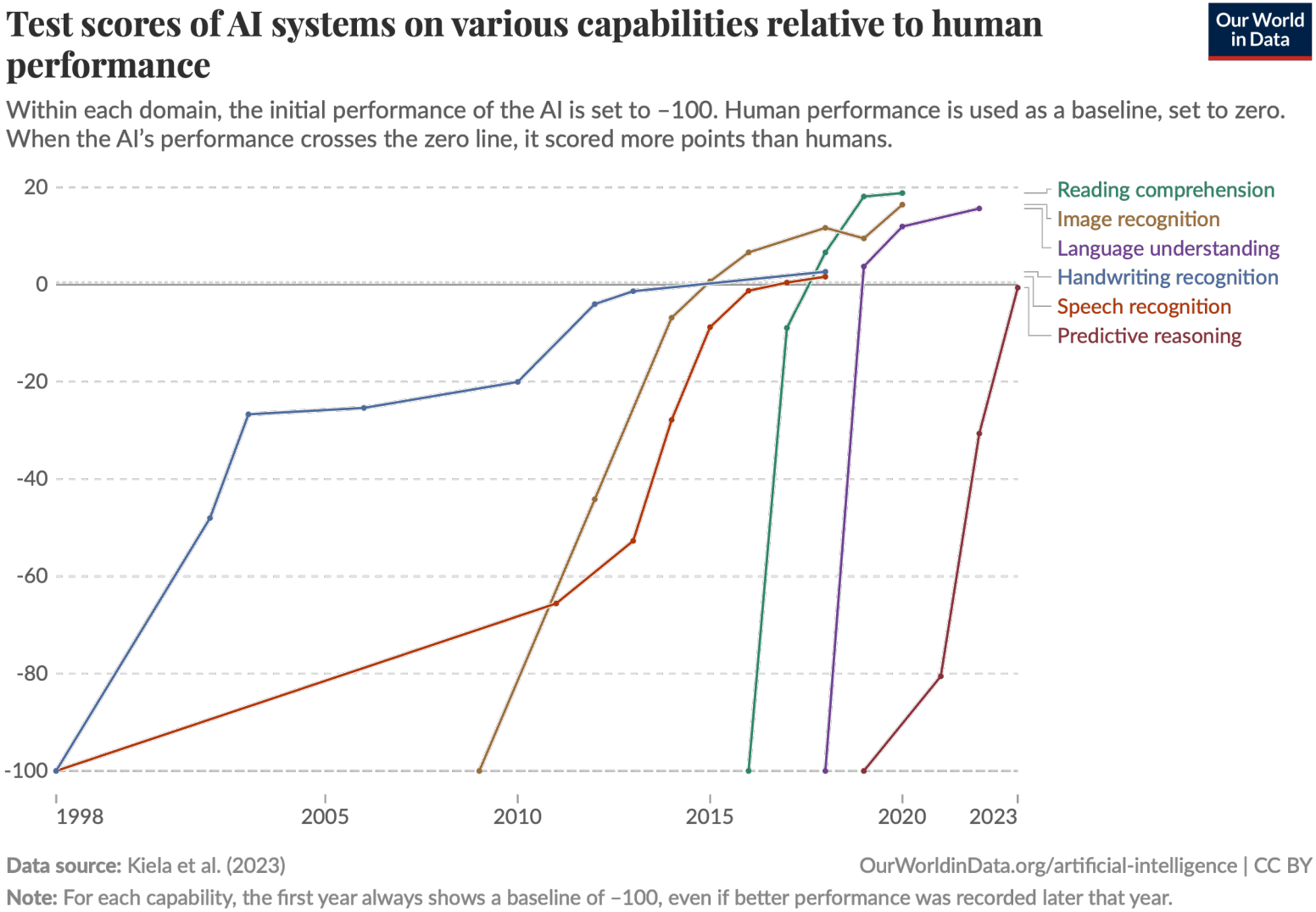
Consistency and proper weighting (variables are weighted the same way every time, according to their importance)

Feedback & base-rates 'built-in' to the system (clinicians rarely get immediate feedback and have imperfect memory)

Not overly sensitive to optimal weightings (simple linear weightings often do well)

Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. ***Science, 243***(4899), 1668–1674.

Scaling up actuarial judgment…



https://ourworldindata.org/artificial-intelligence

Performance of LLMs in diagnostic prediction

"We used a zero-shot approach in that we did not prompt or finetune the chatbots with the goal of improving their performance. The three AI chatbots were asked to provide the three most likely medical diagnoses, rank their choice in order of likelihood, and offer clinical reasoning behind their diagnoses"

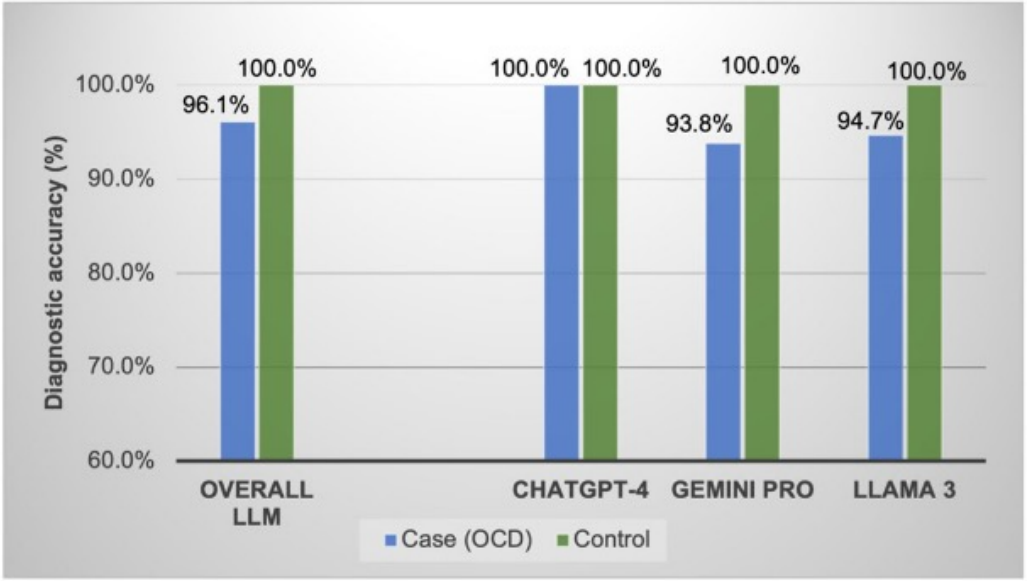


Fig. 1 | Diagnostic accuracy of LLMs by case (OCD) and control (other psychiatric disorders). Overall LLM performance: Case ($N = 49/51$) and control ($N = 21/21$). ChatGPT-4 and Gemini Pro (Case: 16 OCD vignettes) and Llama 3 (Case: 19 OCD vignettes). All LLMs had the same control group comprised of seven psychiatric disorders (major depressive disorder, generalized anxiety disorder, post-traumatic stress disorder, uni or bipolar depression, depression among adolescents, social anxiety disorder, and panic disorder).

Kim, J., Leonte, K. G., Chen, M. L., Torous, J. B., Linos, E., Pinto, A., & Rodriguez, C. I. (2024). Large language models outperform mental and medical health care professionals in identifying obsessive-compulsive disorder. Npj Digital Medicine, 7(1), 193. https://doi.org/10.1038/s41746-024-01181-x

LLMs outperform mental health care professionals

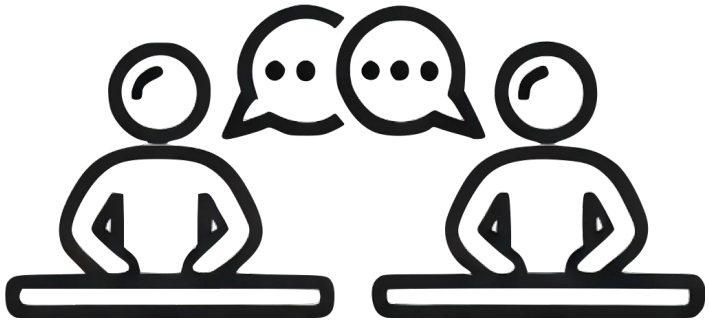
Table 1 | OCD identification rate of LLMs and medical and mental health professionals

Content of OCD vignette	LLM			Group 1. APA members[a] (*N* = 360)	Group 2. Primary care physicians[b] (*N* = 208)	Group 3. Doctoral trainees in psychology[c] (*N* = 130)	Group 4. Medical providers in Guam[d] (*N* = 105)	Group 5. Clergy members in Guam (*N* = 110)
	ChatGPT-4 (*N* = 16)	Gemini Pro (*N* = 16)	Llama 3 (*N* = 19)					
Across all vignettes	96.1% (*N* = 49/51)[e]			61.1%	49.5%	81.5%	41.9%	35.5%
	100% (*n* = 16/16)	93.8% (*n* = 15/16)	94.7% (*n* = 18/19)					
Vignette 1. Harm obsessions	100% (3/3 trials)	100% (3/3 trials)	100% (3/3 trials)	68.5%	20.0%	77.8%	18.2%	27.8%
Vignette 2. Sexual orientation obsessions	100% (3/3 trials)	100% (2/2 trials)[g]	66.7% (2/3 trials)	23.0%	15.4%	66.7%	10.0%	6.7%
Vignette 3. Sexual attraction to children obsessions	No response[f]	100% (1/1 trial)[h]	100% (3/3 trails)	57.1%	29.2%	77.8%	15.0%	11.1%
Vignette 4. Religious obsessions	100% (3/3 trials)	100% (3/3 trials)	100% (3/3 trials)	71.2%	62.5%	80.0%	72.7%	21.7%
Vignette 5. Contamination obsessions	100% (3/3 trials)	100% (3/3 trials)	100% (3/3 trials)	84.2%	67.7%	93.7%	83.3%	73.3%
Vignette 6. Blurting out offensive language obsessions	100% (1/1 trial)	100% (1/1 trial)	100% (1/1 trial)	N/A	26.1%	74.5%	N/A	N/A
Vignette 7. Somatic obsessions	100% (1/1 trial)	0% (0/1 trial)	100% (1/1 trial)	N/A	60.0%	82.4%	N/A	N/A
Vignette 8. Symmetry obsessions	100% (2/2 trials)	100% (2/2 trials)	100% (2/2 trials)	N/A	96.3%	100.0%	85.0%	71.4%

[a]The top five degrees/licenses of the American Psychological Association (APA) members were PhD (67.6%), MA/MS (31.5%), PsyD (14.2%), EdD/EdS/EdM (6.8%), MSW/LMSW (1.7%). Currently licensed was 81.3%.
[b]The areas of specialty included Internal Medicine (35.4%), Pediatrics (32.3%), Family Medicine (22.2%), other specialties (10.6%), and General Medicine (4.5%).
[c]The degrees include Clinical Psychology with Health Emphasis PhD, School-Clinical PsyD, School-Clinical PhD, Clinical Psychology PsyD, and Clinical Psychology PhD in 7 APA-accredited doctoral programs in the Greater New York area.
[d]Group 4 includes medical doctors, nurse practitioners, physician assistants, and doctors of Osteopathic Medicine. The areas of specialty included Internal Medicine (33.3%), Family Medicine (26.7%), Pediatrics (16.2%), Obstetrics and Gynecology (6.7%), Emergency Medicine (4.8%), and other (12.6%).
[e]The sample size differs between LLMs and mental health and health care professionals due to study design (LLM: responses from three LLM models; Human participants: responses from the wide distribution of the vignette studies).
[f]LLM (ChatGPT-4) did not respond to all three vignette trials due to content violation.
[g]LLM (Gemini Pro) did not respond to one vignette trial due to content violation.
[h]LLM (Gemini Pro) did not respond to two vignette trials due to content violation.

Kim, J., Leonte, K. G., Chen, M. L., Torous, J. B., Linos, E., Pinto, A., & Rodriguez, C. I. (2024). Large language models outperform mental and medical health care professionals in identifying obsessive-compulsive disorder. Npj Digital Medicine, 7(1), 193. https://doi.org/10.1038/s41746-024-01181-x

HOW CAN OR SHOULD AI REPLACE HUMAN PSYCHOLOGISTS?



Summary

- **Clinical vs. actuarial judgment:** clinical judgment as the integration of data in the head; actuarial judgment as the integration of data through an algorithm; problem of integration NOT data availability…
- **Lens model and policy capturing:** Lens model as a general depiction of data integration; supplies framework and terminology to help assess the relative benefits of clinical and actuarial judgment; formalisation of judgment process using an algorithm (e.g., regression model)
- **Why humans fail**: increased experience may not be strongly related to improved performance (lack of immediate/appropriate feedback?), (over)confidence, incorrect or inconsistent weighting
- **Empirical evidence**: Meta-analyses of field studies in several domains (e.g., academic, mental health) suggest that actuarial judgment can outperform clinical judgment by ca. 20%; recent studies suggest new models (LLMs) can already outperform clinical judgement (zero-shot performance) promising actuarial judgment at scale…