“The Limits of Algorithmic Accuracy”

Presented at College of Liberal Arts and Sciences | CU Denver (Beijing Campus)

13th October, 2025

Charles Rahal

How accurately can you predict the result of a flipped coin?

What about a spin on the roulette wheel?

Now, consider the following quote…

“We may regard the present state of the universe as the effect of its past and the cause of its future. An intellect which at a certain moment would know all forces that set nature in motion, and all positions of all items of which nature is composed, if this intellect were also vast enough to submit these data to analysis, … nothing would be uncertain.”

– Laplace (1814), ‘Essai philosophique sur les probabilites’

“We may regard the present state of the universe as the effect of its past and the cause of its future. An intellect which at a certain moment would know all forces that set nature in motion, and all positions of all items of which nature is composed, if this intellect were also vast enough to submit these data to analysis, … nothing would be uncertain.”

– Laplace (1814), ‘Essai philosophique sur les probabilites’

Do you think that Laplace’s Demon could predict the value of a flipped coin?

Imagine you have a prediction task…

For example, let's imagine we are predicting GPA, or one-year all-cause mortality.

Pick your favourite model evaluation metric. This could be accuracy, R2, or ROC-AUC.

Here is your model performance:

R^2 = 0.85

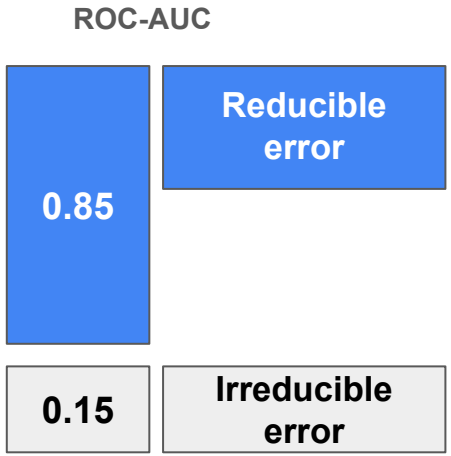
The best possible R^2=1, where we have perfectly predicted everything.

Then…

1. **What's the 0.15, the gap between your current prediction and the best possible prediction?**

2. **Are we ever able to achieve the 'perfect' score as defined by any common metric?**

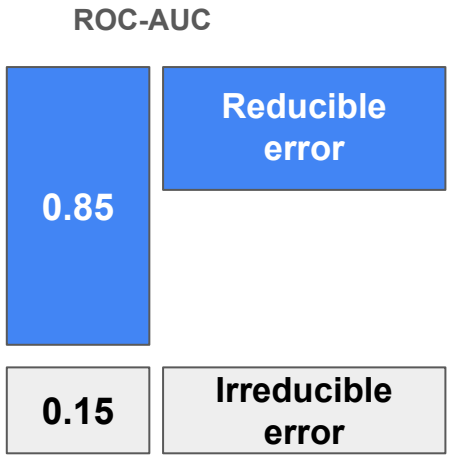
Existing error decomposition (e.g., Hastie et al. 2001)



$$\begin{aligned} \mathrm{EPE}_k(x_0) &= \mathrm{E}[(Y - \hat{f}_k(x_0))^2 | X = x_0] \\ &= \sigma^2 + [\mathrm{Bias}^2(\hat{f}_k(x_0)) + \mathrm{Var}_{\mathcal{T}}(\hat{f}_k(x_0))] \qquad (2.46) \\ &= \sigma^2 + \left[f(x_0) - \frac{1}{k} \sum_{\ell=1}^{k} f(x_{(\ell)}) \right]^2 + \frac{\sigma^2}{k}. \qquad (2.47) \end{aligned}$$

How can we better systematically think about prediction errors?

Existing error decomposition (e.g., Hastie et al. 2001)



From ▭ to ▭ :

What determines accuracy?

- Algorithms
- Variable exclusion/inclusion.
- Construction of the target feature

What is 'irreducible' today may not be irreducible tomorrow!

Everything is EVOLVING!

Including data, algorithms and the construction of the target feature itself

How can we better systematically think about prediction errors?

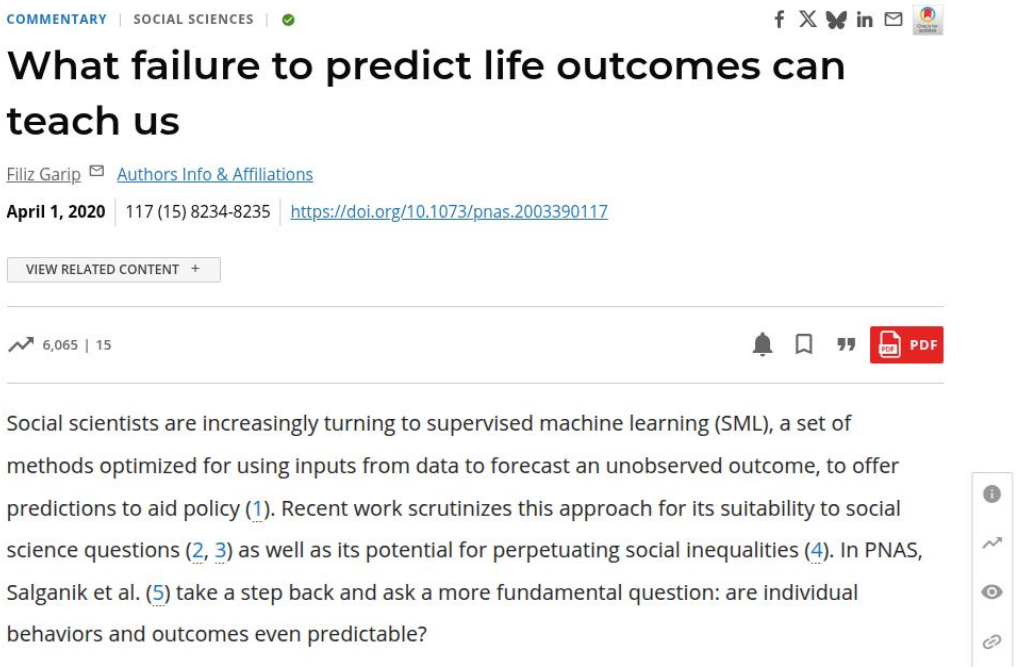
This is increasingly important as health data scientists are increasingly making statements about how ‘predictable’ things are.

This is increasingly important as health data scientists are increasingly making statements about how ‘predictable’ things are.

insurance has the advantage. When an individual considers the medical insurance policy in a financial plan, it can reduce the trouble. The safety assurance against rising healthcare charges is a remarkable benefit of medical insurance.

- ***Insurance coverage for healthcare emergencies*** - health emergencies are unpredictable and can occur at random. With a good medical insurance policy, one need not worry about the expenses, but rather concentrate and full recovery from the illness.
- ***Improvements in medical technology and capacity*** - according to a study by Weisbrod (1991), the expansion of health insurance coverage over the years has created incentives to

This is increasingly important as health data scientists are increasingly making statements about how ‘predictable’ things are.



COMMENTARY | SOCIAL SCIENCES |

What failure to predict life outcomes can teach us

Filiz Garip | Authors Info & Affiliations

April 1, 2020 | 117 (15) 8234-8235 | https://doi.org/10.1073/pnas.2003390117

VIEW RELATED CONTENT +

6,065 | 15

PDF

Social scientists are increasingly turning to supervised machine learning (SML), a set of methods optimized for using inputs from data to forecast an unobserved outcome, to offer predictions to aid policy (1). Recent work scrutinizes this approach for its suitability to social science questions (2, 3) as well as its potential for perpetuating social inequalities (4). In PNAS, Salganik et al. (5) take a step back and ask a more fundamental question: are individual behaviors and outcomes even predictable?

This is increasingly important as health data scientists are increasingly making statements about how ‘predictable’ things are.

We argue that they should not be doing this!

nature computational science

Correspondence | Published: 04 March 2025

On the unknowable limits to prediction

Jiani Yan ✉ & Charles Rahal ✉

Nature Computational Science **5**, 188–190 (2025) | Cite this article

Conceptualise error differently/more reflexively:

Aleatoric error

Inherent randomness can never be modelled
(genuine stochasticity)

Epistemic error

In theory eliminable with human progress
(resulting from a lack of knowledge)

A typical prediction task workflow

1. **Form a research question:** e.g., *'How predictable is GPA/one-year all-cause mortality?'*
2. **Read relevant paper & theories, identify covariates:** *e.g., social determinants of health.*
3. **Find appropriate/ collect data for the research question:** *e.g., biobanks, surveys.*
4. **Use proper algorithm(s) to run the prediction task:** *e.g., a regression, or an LLM.*
5. **Analyse outcomes:** *e.g., any of your favourite model evaluation metrics.*

Once data, research object and covariates are specified, the research design is confirmed; The aleatoric error of the research object target feature becomes fixed.

Table 2: Evaluation of Model Performance in Death Prediction over Ten Seeds

	HRS			SHARE			ELSA		
Metrics	SL	LightGBM	LR	SL	LightGBM	LR	SL	LightGBM	LR
IMV	0.192	0.195	0.111	0.073	0.077	-0.055	0.056	0.060	0.068
ROC-AUC	0.814	0.816	0.823	0.789	0.795	0.453	0.823	0.828	0.837
PR-AUC	0.687	0.691	0.702	0.508	0.522	0.172	0.491	0.499	0.521
Pseudo R^2	0.275	0.280	0.295	0.195	0.207	-0.041	0.208	0.221	0.243
FFC R^2	0.602	0.605	0.613	0.443	0.451	0.279	0.394	0.404	0.421
Test IP	0.317	0.317	0.317	0.202	0.202	0.202	0.155	0.155	0.155

Is 0.823 the predictive ceiling? Can I ever improve it?

Example taken from Yan, J. (2025), '*Revisiting the social determinants of health with explainable AI: a cross-country perspective*', American Journal of Epidemiology

$$
\begin{aligned}
y_{\text{true}} &= \underbrace{f^*(\mathbf{x}_{\text{true}})}_{\substack{\text{Predictive}\\ \text{ceiling}}} + \underbrace{\varepsilon}_{\substack{\text{Aleatoric}\\ \text{error}}} \\
&= \underbrace{[f^*(\mathbf{x}_{\text{true}}) - f(\mathbf{x}_{\text{true}}|y_{\text{true}})]}_{\substack{\text{Model approximation gain}\\ \text{(Epistemic)}}} \\
&+ \underbrace{[f(\mathbf{x}_{\text{true}}|y_{\text{true}}) - f(\mathbf{x}_{\text{true}}|y_{\text{observed}})]}_{\substack{\text{Measurement gain from } y\\ \text{(Epistemic)}}} \\
&+ \underbrace{[f(\mathbf{x}_{\text{true}}|y_{\text{observed}}) - f(\mathbf{x}_{\text{observed}}|y_{\text{observed}})]}_{\substack{\text{Measurement gain from } \mathbf{x}\\ \text{(Epistemic)}}} \\
&+ \underbrace{y_{\text{predicted}}}_{\substack{\text{Current}\\ \text{prediction}}} + \underbrace{\varepsilon}_{\substack{\text{Irreducible}\\ \text{(Aleatoric)}}}.
\end{aligned}
$$

Algorithm

Y — Construct Validity

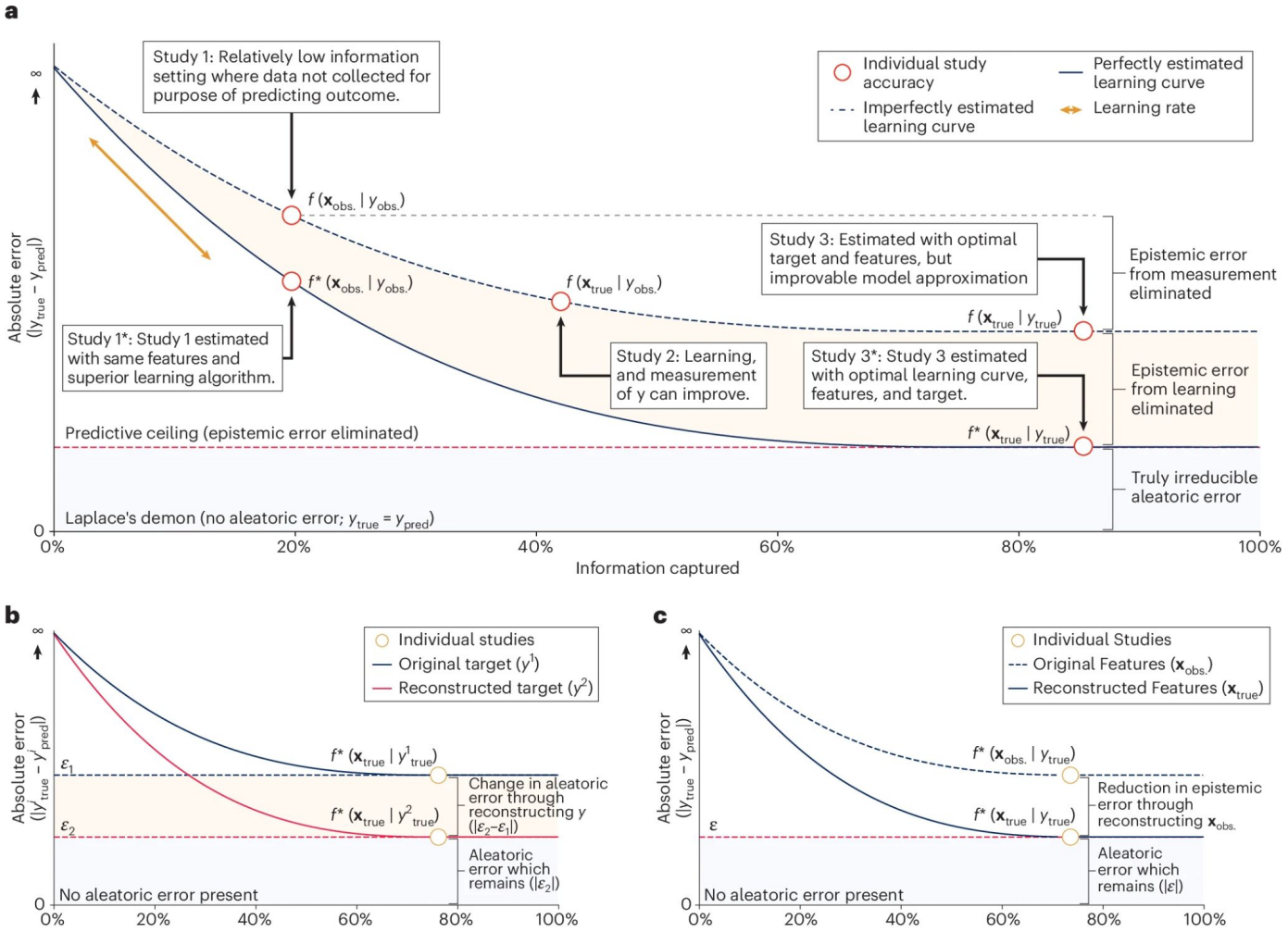
Representation

X — Relevant information inclusion

We claim that people should be responsibly using terms while thinking about the limits of human endeavour:

We can only say how predictable things are *conditional* on features and predictive machines.

We emphasize the importance – and develop a new style – of ‘learning curve’.


a
Study 1: Relatively low information setting where data not collected for purpose of predicting outcome.
Individual study accuracy
Perfectly estimated learning curve
Imperfectly estimated learning curve
Learning rate
Absolute error (|y_true − y_pred|)
∞
0
f (x_obs. | y_obs.)
f* (x_obs. | y_obs.)
f (x_true | y_obs.)
Study 3: Estimated with optimal target and features, but improvable model approximation
f (x_true | y_true)
Epistemic error from measurement eliminated
Study 1*: Study 1 estimated with same features and superior learning algorithm.
Study 2: Learning, and measurement of y can improve.
Study 3*: Study 3 estimated with optimal learning curve, features, and target.
Epistemic error from learning eliminated
Predictive ceiling (epistemic error eliminated)
f* (x_true | y_true)
Truly irreducible aleatoric error
Laplace's demon (no aleatoric error; y_true = y_pred)
0%
20%
40%
60%
80%
100%
Information captured
b
Individual studies
Original target (y¹)
Reconstructed target (y²)
Absolute error (|y¹_true − y¹_pred|)
f* (x_true | y¹_true)
ε1
Change in aleatoric error through reconstructing y (|ε2−ε1|)
f* (x_true | y²_true)
ε2
Aleatoric error which remains (|ε2|)
No aleatoric error present
c
Individual Studies
Original Features (x_obs.)
Reconstructed Features (x_true)
Absolute error (|y_true − y_pred|)
f* (x_obs. | y_true)
f* (x_true | y_true)
Reduction in epistemic error through reconstructing x_obs.
ε
Aleatoric error which remains (|ε|)
No aleatoric error present

What can we do? Comprehensive algorithms and datasets

Article | Open access | Published: 17 September 2025

Learning the natural history of human disease with generative transformers

Artem Shmatko, Alexander Wolfgang Jung, Kumar Gaurav, Søren Brunak, Laust Hvas Mortensen, Ewan Birney, Tom Fitzgerald & Moritz Gerstung

Nature (2025) | Cite this article

146k Accesses | 3 Citations | 1446 Altmetric | Metrics

Abstract

Decision-making in healthcare relies on understanding patients' past and current health states to predict and, ultimately, change their future course[1,2,3]. Artificial intelligence (AI) methods promise to aid this task by learning patterns of disease progression from large corpora of health records[4,5]. However, their potential has not been fully investigated at scale. Here we modify the GPT[6] (generative pretrained transformer) architecture to model the progression and competing nature of human diseases. We train this model, Delphi-2M, on data from 0.4 million UK Biobank participants and validate it using external data from 1.9 million Danish individuals with no change in parameters. Delphi-2M predicts the rates of more than 1,000 diseases, conditional on each individual's past disease history, with accuracy comparable to that of existing single-disease models. Delphi-2M's generative nature also enables sampling of synthetic future health trajectories, providing meaningful estimates of potential disease burden for up to 20 years, and enabling the training of AI models that have never seen actual data. Explainable AI methods[7] provide insights into Delphi-2M's predictions, revealing clusters of co-morbidities within and across disease chapters and their time-dependent consequences on future health, but also highlight biases learnt from training data. In summary, transformer-based models appear to be well suited for predictive and generative health-related tasks, are applicable to population-scale datasets and provide insights into temporal dependencies between disease events, potentially improving the understanding of personalized health risks and informing precision medicine approaches.

Download PDF

Associated content

This AI tool predicts your risk of 1,000 diseases – by looking at your medical records

Shamini Bundell & Nick Petrić Howe
Nature | Nature Podcast | 17 Sept 2025

Which diseases will you have in 20 years? This AI accurately predicts your risks

Gemma Conroy
Nature | News | 17 Sept 2025

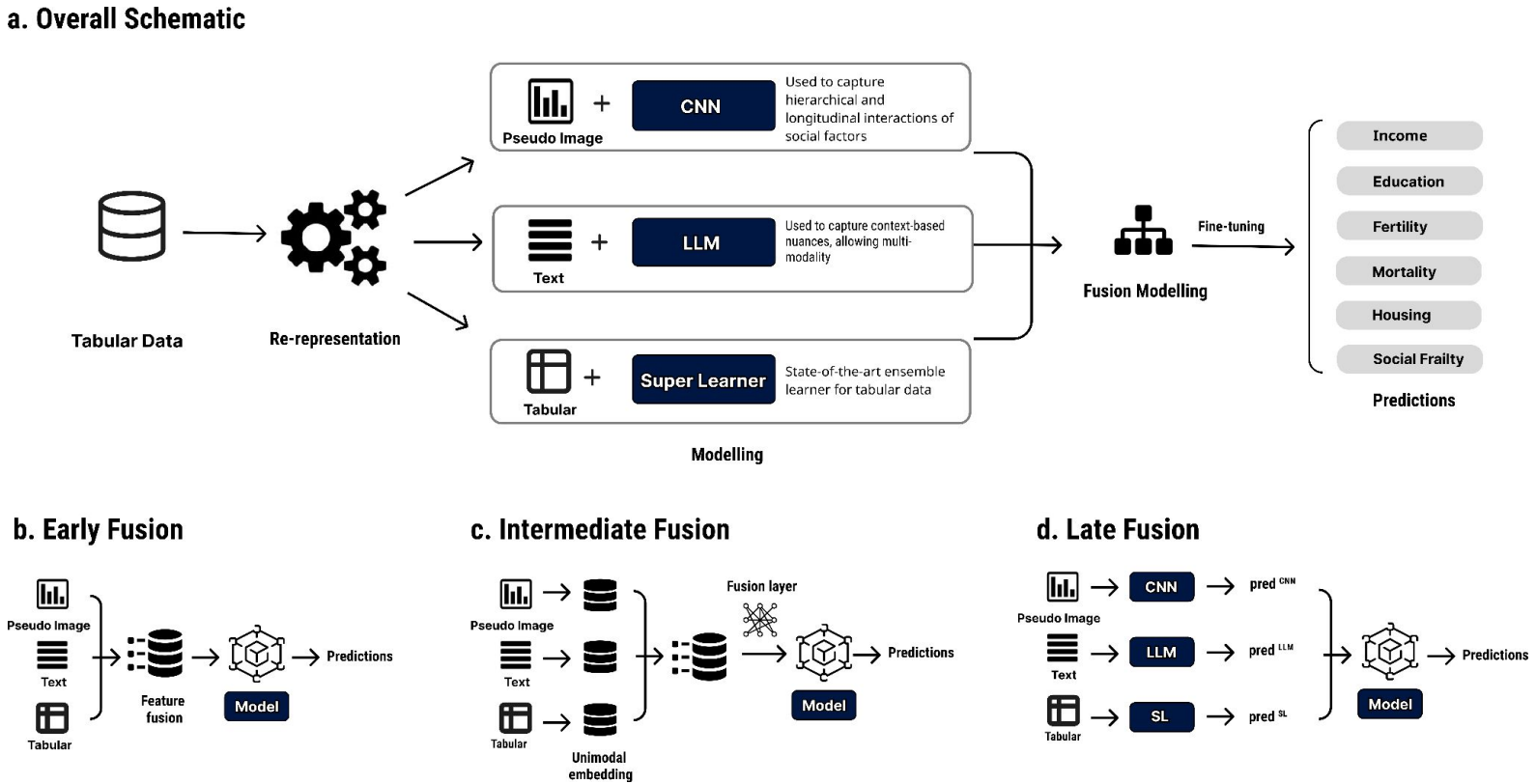
AI uses medical records to accurately predict onset of disease 20 years into the future

Yonghui Wu
Nature | News & Views | 30 Sept 2025

Sections | Figures | References

Abstract
Main
A transformer model for health records
Modelling multi-disease incidences
Sampling future disease trajectories
Explaining Delphi-2M predictions
External validation and bias assessment
Discussion
Methods

What can we do? Comprehensive algorithms and datasets



Can we ever know where we are in the process?

- We have up to seven new approaches for approximating epistemic error across data, model, and measurement.
- These methods include learning curves (Mohr & van Rijn, 2021), Taylor expansions with the law of propagation of uncertainty, and bootstrap variants.
- We also developed a large-scale integrated simulation bench that uses user-defined data-generating mechanisms.

- Aleatoric error is unmodelable; fixed at the point of confirming your research question.
- Aleatoric error levels can vary depending on the research object; some topics are easier to predict than others.
- Epistemic error can be eliminated through
  - Better capture of information
    - Better representativity
    - Better variable construction
    - Better relevant information inclusion
  - Better algorithms

“ Until we have reason to believe that we have eliminated error through better measurement and construction of features and outputs — in addition to the elimination of learning error — we should only tentatively discuss how predictable things are with current data and technology.”

Thank you for your time and attention!

Paper in Nature Computational Science:

We’re also recruiting a DPhil student: