AEA Papers and Proceedings 2024, 114: 618–622
https://doi.org/10.1257/pandp.20241071

ALGORITHMIC DECISION-MAKING[‡]

Comparative Advantage of Humans versus AI in the Long Tail[†]

By Nikhil Agarwal, Ray Huang, Alex Moehring,
Pranav Rajpurkar, Tobias Salz, and Feiyang Yu*

Supervised machine learning algorithms use large labeled datasets to perform predictive tasks (see LeCun, Bengio, and Hinton 2015 for an early review). These algorithms have demonstrated superior performance compared to human experts in several key areas (Lai et al. 2021; Mullainathan and Obermeyer 2019). Many anticipate significant job displacements, especially in diagnostic radiology.[1] A counterargument holds that the short-term risk of job displacement is limited because most jobs require several different tasks to be performed, not all of which are about prediction (see Agrawal, Gans, and Goldfarb 2019 and Langlotz 2019 for example).

The long-tail hypothesis holds that humans may remain relevant even within prediction domains, at least in the medium run, as humans can learn from relatively few examples (see Malaviya et al. 2022; Casler and Kelemen 2005).[2] In radiology, Langlotz argues that humans will remain relevant because "radiologists know the 'long tail'" of diseases, each of which are uncommon but are together relevant for a large proportion of patients (2019, 2).[3] Similar arguments apply to other applications where artificial intelligence (AI) has made inroads. Autonomous cars, for instance, suffer from a "curse of rarity" because specific constellations are seldom encountered (Liu and Feng 2022, 1). Humans can overcome this curse by using knowledge outside the driving domain. Thus, the long-tail hypothesis holds that job displacement may be limited if a job requires performing several complementary or essential tasks that are hard to automate because the tasks are individually rarely encountered.

A challenge in studying the long-tail hypothesis is to find a class of similar problems that can be ordered by how commonly they are encountered. We argue that the interpretation of chest X-rays for various pathologies offers such a class. Here, disease prevalence can be used to parametrize the long tail, as rarer diseases will have fewer training examples.

This paper examines whether self-supervised learning algorithms—which learn broadly because they do not require structured labels—have diminished the advantage of human radiologists in diagnosing the long tail of diseases. Specifically, we compare the performance of

[‡] *Discussants:* Ziad Obermeyer, University of California, Berkeley; Paul Goldsmith-Pinkham, Yale University; Jann Spiess, Stanford University; Michael Luca, Harvard University.

* Agarwal: Deptartment of Economics, MIT, and NBER (email: agarwaln@mit.edu); Huang: MIT Blueprint Labs (email: rayhuang@mit.edu); Moehring: MIT Sloan School of Management (email: moehring@mit.edu); Rajpurkar: Deptartment of Biomedical Informatics, Harvard Medical School (email: pranav_rajpurkar@hms.harvard.edu); Salz: Deptartment of Economics, MIT, and NBER (email: tsalz@mit.edu); Yu: Deptartment of Biomedical Informatics, Harvard Medical School (email: fyu9@cs.stanford.edu). We are grateful to Stanford University Hospital for facilitating data access. The authors acknowledge support from the Alfred P. Sloan Foundation (grant 2022-17182), Abdul Latif Jameel Poverty Action Lab's Healthcare Delivery Initiative, and MIT School of Humanities, Arts, and Social Sciences. We thank David Yanigazawa-Drott for helpful comments and suggestions. This paper reports on secondary analysis from data collected and analyzed in Agarwal et al. (2023).

[†] Go to https://doi.org/10.1257/pandp.20241071 to visit the article page for additional materials and author disclosure statement(s).

[1] "We should stop training radiologists now. It's just completely obvious that within five years, deep learning is going to do better than radiologists" (Geoff Hinton, "On Radiology," November 24, 2016, YouTube video, https://www.youtube.com/watch?v=2HMPRXstSvQ).

[2] A large literature in cognitive philosophy questions how humans establish knowledge with limited observation (see Russell 2009 for example), with some hypothesizing that aspects of human knowledge must be innate (see Chomsky 1986 for example).

[3] A similar idea within economics posits that the long tail of products together can account for a large fraction of total surplus (Waldfogel 2017).

CheXzero (Tiu et al. 2022), a zero-shot algorithm for diagnosing chest pathologies, to human radiologists across 79 diseases. We compare the two to predictions from the CheXpert algorithm (Irvin et al. 2019), a traditional supervised deep learning algorithm capable of diagnosing 12 chest pathologies. To examine the hypothesis that humans will remain relevant in the long tail of diseases, we study how the performance of these classifiers vary with disease prevalence.

I. Background and Data

A. *CheXzero versus CheXpert*

CheXzero is a self-supervised learning method that detects diseases on chest X-rays directly from natural-language descriptions in clinical reports (Tiu et al. 2022). The model was trained on more than 377,000 chest X-rays and more than 227,000 corresponding radiology reports from the MIMIC-CXR dataset (Johnson et al. 2019). CheXzero scores whether a positive or negative prompt for a pathology is a better pairing for the image to form a prediction. Tiu et al. (2022) show that this algorithm performs comparably to radiologists on a set of five diseases.

CheXpert is a supervised deep learning algorithm trained on 224,316 radiographs from the Stanford hospital (Irvin et al. 2019). It predicts the presence of 12 pathologies that are explicitly labeled in the training data. It has previously been shown to match or surpass the performance of radiologists on the majority of diseases (Irvin et al. 2019; Agarwal et al. 2023).

B. *Data Collection*

We use data first reported in Agarwal et al. (2023)—henceforth, AMRS. Our analysis focuses exclusively on the treatment arms where no AI assistance was provided as we seek to document the comparative performance of humans and AI across pathology prevalence rather than the use of AI by humans, which is the focus of AMRS.

Participants use a remote interface (see the online Appendix). The interface mimics clinical practice, but elicits structured data on 79 pathologies instead of free-text reports.[4]

We use data from 227 radiologists, each reading between 30 and 120 cases (average of 46 cases) from a sample of 324 cases from the Stanford hospital. None of these cases are in the AI training sets. We refer readers to AMRS for details of the data collection.

II. Human versus AI Performance

We compare AI and radiologist performance using the concordance statistic C, which generalizes the area under the receiving operating characteristic curve (AUROC) to a continuous setting. C_{rt} is defined as the share of concordant pairs: $C_{rt} = \Pr\left(p_{irt} > p_{i'rt} \middle| \bar{p}_{it} > \bar{p}_{i't}\right)$, where i, r and t index cases, radiologists, and pathologies, respectively. $\bar{p}_{it}$ is the average probability assessment from a panel of radiologists specializing in chest radiology, which we call the consensus probability.[5] We average C_{rt} across r to obtain C_t. C_t^A denotes AI concordance. We use a block bootstrap for inference to account for correlation within cases and radiologists.

Concordance measures the probability a classifier's ranking of disease likelihood aligns with that of an experienced expert panel. The use of an expert panel mirrors approaches in computer science (see Sheng, Provost, and Ipeirotis 2008 and references therein). The method circumvents the challenge, discussed in AMRS, that a definitive diagnostic test often does not exist. Although the consensus may err, higher concordance implies greater fidelity to the aggregate opinion of experienced experts.

We choose concordance as a performance metric because it is invariant to prevalence and preferences. It is calculable even with no cases that are positive with high probability. This feature is vital in our context as 45 pathologies contain no cases where the consensus probability exceeds 0.5. Concordance provides an informative signal about the performance of the classifier as long as there is some variation across cases in $\bar{p}_{it}$.

A drawback of concordance is that it is an ordinal measure of performance that does not immediately result in treatment recommendations. Turning ordinal measures of disease risk into treatment recommendations requires estimates of the relative costs of various decisions (see Chan,

[4] AMRS also collect assessments for the presence of support devices/hardware and an assessment if the case is normal. We exclude these assessments.

[5] We score ties $(p_{irt} = p_{i'rt})$ as 0.5 in computing C_{rt}.

Table 1—Summary Statistics

	Mean	SD
Pathology prevalence	2.42	3.87
Radiologist concordance	0.58	0.06
CheXzero concordance	0.67	0.15
CheXpert concordance	0.72	0.12
Reads per radiologist	46.2	

Notes: Summary statistics at the pathology level. CheXpert concordance uses 12 pathologies, while radiologist and CheXzero concordances use 79 pathologies.

Gentzkow, and Yu 2022 and AMRS for examples within radiology).

An alternative performance measure we consider in the online Appendix is the deviation from consensus probability, $|p_{irt} - \bar{p}_{it}|$. It is calculable for all pathologies but is a misleading performance measure for rare diseases as there is less variance in $\bar{p}_{it}$ and overestimates performance in the long tail relative to concordance. Some conclusions are sensitive to the use of this alternative.

Table 1 summarizes the data and compares performance of human radiologists and AI algorithms. Average prevalence is low at 2.4 percent. The distribution of prevalence is heavily skewed. Radiologists perform worse than both algorithms. CheXpert performs slightly better than CheXzero on the 12 pathologies for which it has predictions.

III. The Long Tail

Disease prevalence provides a measure to parameterize the long tail. Figure 1 plots the cumulative share of positive cases by pathology to assess the importance of the long tail. We arrange pathologies so that those with CheXpert assessments appear first. Within the two groups, we order pathologies by prevalence. The 12 pathologies with CheXpert assessments constitute less than 60 percent of the overall prevalence. Thus, a significant proportion of key pathologies are not predicted by CheXpert.

We next examine if humans outperform AI for rare pathologies. Figure 2 compares human and AI performance across the 12 pathologies that have CheXpert predictions, grouped by low, medium, and high prevalence in our sample.

Both human and CheXpert performance increases with pathology prevalence. CheXpert displays the largest improvement between the low and medium prevalence bins. The difference in CheXpert performance in the low bin and both the medium and high bins is statistically significant at the 1 percent level. Humans display modest and statistically significant improvements as prevalence increases. CheXzero's performance is less sensitive to prevalence. Notably, CheXzero outperforms humans in all bins, providing initial evidence against the claim that

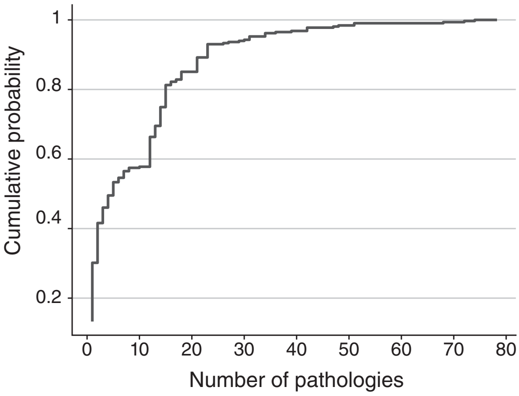

Figure 1. Cumulative Share of Positive Cases

Notes: Cumulative share of positive cases by pathology. Pathologies are ordered from most prevalent to least, with CheXpert pathologies first.

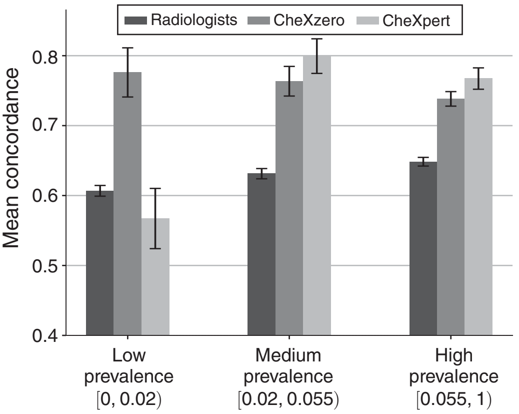

Figure 2. Radiologists versus AI Performance

Notes: Average concordance by classifier for three prevalence bins. We limit our sample to pathologies with CheXpert reads, resulting in four pathologies per bin. The 95 percent confidence intervals are block bootstrapped.

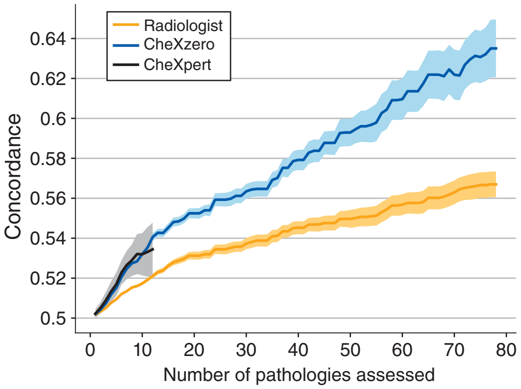

Figure 3. Performance versus Pathologies Assessed

Notes: Performance of human versus AI. The *y*-axis concordance is computed as the average between the concordance of the pathologies assessed and a random guess $C_t = 0.5$ for the pathologies without assessments. Pathology ordering follows Figure 1. The 95 percent pointwise confidence bands are block bootstrapped.

humans will remain relevant in the long tail of diseases.

Figure 3 shows how the average concordance of these classifiers varies as the number of pathologies that are predicted increases. In each case, the concordance is set to 0.5 (random guess) for any pathology that is not assessed. Thus, the line ranges from 0.5 to the average concordance.

The performance in the long tail is critical for assessing the overall quality of a classifier (Figure 3). While CheXpert beats humans and matches CheXzero for its 12 pathologies, only being able to predict a smaller subset of pathologies hinders its overall performance. Across all pathologies, CheXpert's concordance is less than 0.54, suggesting its performance is overestimated when restricted to CheXpert pathologies.

Perhaps the most important takeaway from the figure is that CheXzero's performance is significantly higher than human performance, suggesting that the AI may have humans beat even in the long tail (cf. Langlotz 2019). A note of caution on this conclusion is that although concordance is a reasonable metric for comparing classifiers, it is an ordinal metric for comparing algorithms. Converting ordinal algorithmic output to diagnostic decisions requires additional steps, including determining an appropriate decision threshold. This is particularly challenging for rare pathologies.

IV. Conclusion

While supervised machine learning algorithms have surpassed human performance in specific prediction tasks, humans may continue to add value because of their superior ability to deal with uncommon cases—the long tail. Zero-shot learning algorithms attempt to make progress in the long tail by avoiding the need for large datasets with specifically annotated labels.

We studied the long-tail hypothesis in the iconic example of radiological diagnosis. Our results show that self-supervised algorithms have surpassed human radiologists in the long tail of diseases in terms of predictive ability, providing evidence against the long-tail hypothesis at least in the interpretation of chest X-rays.

Yet, there are hurdles remaining before algorithms—even those based on zero-shot learning methods—are deployed or result in job displacement. The output of the algorithm does not immediately yield probabilities, recommendations, or decisions. These and other technical and regulatory hurdles need to be circumvented before AI displaces human chest X-ray interpretations. Moreover, prediction is just one task in a job (Agrawal, Gans, and Goldfarb 2019). For this and other reasons, it is possible that these tools will assist humans, as opposed to replace them. To the extent that humans remain in the loop, there are a number of important economic issues surrounding the use of AI predictions by humans (cf. AMRS; Angelova, Dobbie, and Yang 2022). It is also possible that an important role of human expertise may be generating high-quality data for training algorithms or auditing their outputs. In our opinion, understanding factors that determine the optimal way of combining human expertise and the use or training of AI tools is a fruitful avenue for research in economics—these issues are central to how AI tools should be incorporated into workflows and how they reshape jobs.

REFERENCES

Agarwal, Nikhil, Alex Moehring, Pranav Rajpurkar, and Tobias Salz. 2023. "Combining Human Expertise with Artificial Intelligence: Experimental Evidence from Radiology." NBER Working Paper 31422.

Agrawal, Ajay, Joshua S. Gans, and Avi Goldfarb. 2019. "Artificial Intelligence: The Ambiguous Labor Market Impact of Automating Prediction." *Journal of Economic Perspectives* 33 (2): 31–50.

Angelova, Victoria, Will S. Dobbie, and Crystal Yang. 2022. "Algorithmic Recommendations and Human Discretion." Unpublished.

Casler, Krista, and Deborah Kelemen. 2005. "Young Children's Rapid Learning about Artifacts." *Developmental Science* 8 (6): 472–80.

Chan, David C., Matthew Gentzkow, and Chuan Yu. 2022. "Selection with Variation in Diagnostic Skill: Evidence from Radiologists." *Quarterly Journal of Economics* 137 (2): 729–83.

Chomsky, Noam. 1986. *Knowledge of Language: Its Nature, Origin, and Use*. New York: Praeger.

Irvin, Jeremy, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, et al. 2019. "CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison." In *AAAI'19/IAAI'19/EAAI'19: Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, 590–97. Honolulu: AAAI Press.

Johnson, Alistair E. W., Tom J. Pollard, Seth J. Berkowitz, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih-Ying Deng, Roger G. Mark, and Steven Horng. 2019. "MIMIC-CXR, a De-identified Publicly Available Database of Chest Radiographs with Free-Text Reports." *Scientific Data* 6 (1): 317.

Lai, Vivian, Chacha Chen, Q. Vera Liao, Alison Smith-Renner, and Chenhao Tan. 2021. "Towards a Science of Human-AI Decision Making: A Survey of Empirical Studies." arXiv: 2112.11471.

Langlotz, Curtis P. 2019. "Will Artificial Intelligence Replace Radiologists?" *Radiology: Artificial Intelligence* 1 (3): e190058.

LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. 2015. "Deep Learning." *Nature* 521 (7553): 436–44.

Liu, Henry X., and Shuo Feng. 2022. "'Curse of Rarity' for Autonomous Vehicles." arXiv: 2207.02749.

Malaviya, Maya, Ilia Sucholutsky, Kerem Oktar, and Thomas L. Griffiths. 2022. "Can Humans Do Less-than-One-Shot Learning?" arXiv: 2202.04670.

Mullainathan, Sendhil, and Ziad Obermeyer. 2019. "A Machine Learning Approach to Low-Value Health Care: Wasted Tests, Missed Heart Attacks and Mis-Predictions." NBER Working Paper 26168.

Russell, Bertrand. 2009. *Human Knowledge: Its Scope and Limits*. London: Routledge.

Sheng, Victor S., Foster Provost, and Panagiotis G. Ipeirotis. 2008. "Get Another Label? Improving Data Quality and Data Mining Using Multiple, Noisy Labelers." In *KDD '08, Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 614–622. New York: Association for Computing Machinery.

Tiu, Ekin, Ellie Talius, Pujan Patel, Curtis P. Langlotz, Andrew Y. Ng, and Pranav Rajpurkar. 2022. "Expert-Level Detection of Pathologies from Unannotated Chest X-ray Images via Self-Supervised Learning." *Nature Biomedical Engineering* 6 (12): 1399–406.

Waldfogel, Joel. 2017. "The Random Long Tail and the Golden Age of Television." *Innovation Policy and the Economy* 17: 1–25.

This article has been cited by:

1. Joshua Hatherley, Anne Kinderlerer, Jens Christian Bjerring, Lauritz Aastrup Munch, Lynsey Threlfall. 2024. The FHJ debate: Will artificial intelligence replace clinical decision making within our lifetimes?. *Future Healthcare Journal* **11**:3, 100178. [Crossref]