Comparative Advantage of Humans vs 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 amounts of labeled data to perform specific 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. (Liu et al. (2019); Lai et al. (2021); Mullainathan and Obermeyer (2019); Kleinberg et al. (2017)). Many anticipate significant job displacements due to these developments, especially in diagnostic radiology.[1] A counterargument holds that the short-term risk of job displacement is limited because the workflow in most jobs requires a number of different tasks to be performed, not all of which are squarely about prediction (see Agrawal, Gans and Goldfarb (2019); Langlotz (2019) for example).

One hypothesis is that humans may remain relevant even within prediction domains, at least in the medium-run, because humans can learn from relatively few examples (see Malaviya et al. (2022); Kühl et al. (2022); Lake et al. (2011); Coutanche and Thompson-Schill (2014); and Casler and Kelemen (2005)).[2] Specifically in radiology, Langlotz (2019) argued 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.[3] Similar arguments can be made for other important applications where AI has made inroads. Autonomous cars, for instance, suffer from a "curse of rarity" (Liu and Feng (2022)), because specific constellations are rarely encountered due to the high dimensionality of the prediction problem. Humans can overcome this curse by drawing on their knowledge outside the specific domain of driving.

This paper examines whether zero-shot learning algorithms – which learn broadly because they do not require structured labels – have diminished the advantage of human radiologists in diagnosing rare diseases. Specifically, we compare the performance of CheXzero (Tiu et al. (2022)), a zero-shot algorithm for diagnosing chest pathologies using X-rays, to human radiologists across 79 diseases. As a point of comparison, 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.[4] To examine the hypothesis that humans will remain relevant in the long-tail of diseases, we study how the com-

* Agarwal: Department of Economics, MIT and NBER, 50 Memorial Dr, Cambridge, MA 02142 (email: agarwaln@mit.edu). Huang: MIT Blueprint Labs, 30 Wadsworth St. Cambridge, MA 02142 (email: rayhuang@mit.edu). Moehring: MIT Sloan School of Management, 100 Main St, Cambridge, MA 02142 (email: moehring@mit.edu). Rajpurkar: Department of Biomedical Informatics, Harvard Medical School, 25 Shattuck St, Boston, MA 02115 (email: pranav_rajpurkar@hms.harvard.edu). Salz: Department of Economics, MIT and NBER, 50 Memorial Dr, Cambridge, MA 02142 (email: tsalz@mit.edu). Yu: Department of Biomedical Informatics, Harvard Medical School and Department of Computer Science, Stanford University, 25 Shattuck St, Boston, MA 02115 (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 (2022-17182), JPAL Healthcare Delivery Initiative, and MIT SHASS. This paper reports on secondary analysis from data collected and analyzed in Agarwal et al. (2023)

[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" – Geoffrey Hinton (in 2016).

[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).

[4] We exclude support devices and an overall assessment of whether there are "no findings" from the analysis.

parative performance of these algorithms and human radiologists varies with disease prevalence.

I. Background and Data

A. CheXzero vs CheXpert

CheXzero is a self-supervised learning algorithm based on zero-shot learning methods (Tiu et al. (2022)). It is trained on 377,110 chest X-ray images paired with a radiological report taken from the MIMIC-CXR dataset (Johnson et al. (2019)). It uses contrastive learning methods to predict whether a positive prompt for a pathology is a better pairing for an image as compared to a negative prompt. This allows the algorithm to score X-rays for multiple labels in a self-supervised manner without the need for explicit labels during training.

The CheXpert algorithm is a supervised deep learning algorithm trained on 224,316 radiographs taken from Stanford hospital (Irvin et al. (2019)). It can predict the presence of the twelve pathologies for which the training data contain explicit labels. In prior studies, it has been shown to match or surpass the performance of professional radiologists on each of these diseases (Irvin et al. (2019)).

B. Data Collection

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

Participants use a remote interface we developed, as shown in figure 1. This interface mimics typical clinical practice, but instead of a free-text report, it elicits structured data on radiologists for 79 pathologies. Radiologists' probability reports are elicited through a continuous slider.[5] These assessments are obtained using a pathology hierarchy.

We use data from 227 radiologists, each reading between 30 and 120 cases (approximately 46 cases on average) from a sample of 324 cases from Stanford hospital. None of these cases were used to train the AI models we study. We refer the reader to AMRS for further details on the samples and the data collection process.

II. Comparing Human and AI Performance

We compare the performance of AI and radiologist predictions using the concordance statistic C, which is a generalization of area under the receiving operating characteristic curve (AUROC) to a continuous setting.[6] C_{rt} is defined as the proportion of concordant pairs: $C_{rt} = P(p_{irt} > p_{i'rt} | \bar{p}_{it} > \bar{p}_{i't})$ where i represents a case and $\bar{p}_{it}$ represents an aggregated probability assessment from a panel of radiologists specializing in chest radiology, which we call the *consensus probability* given the available diagnostic information for a case. This approach mirrors methods used in computer science for evaluating AI algorithms (see Sheng, Provost and Ipeirotis (2008), and references therein). [7] C_{rt} is computed separately for each radiologist r and pathology t, which we then average across r to obtain C_t. AI concordance is represented as C_t^A. Because concordance is correlated across pathologies and within radiologists, we will employ a bootstrap for inference.[8]

The choice of concordance as a performance metric is based on the property that it is invariant to prevalence and does

[5] AMRS also collect radiologist assessments for the presence of support devices and hardware and an overall assessment if the case is normal. We exclude these assessments and focus on the 79 remaining pathologies.

[6] Like AUROC, a random classifier has a concordance of 0.5, while a perfect classifier has a concordance of 1.

[7] The method circumvents the challenge, discussed in AMRS, that oftentimes a diagnostic test that is more definitive than an X-ray does not exist or is not administered.

[8] Specifically, we estimate the standard error of $\hat{C}_{rt}$ using a block bootstrap that samples radiologists and then the cases that we read within within each bootstrap iteration. To estimate the standard error of $\hat{C}_t^A$, we sample cases within each pathology.

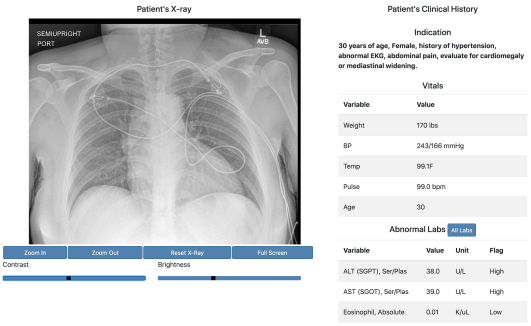

FIGURE 1. DATA COLLECTION INTERFACE

Note: This figure presents an example of the interface seen by radiologists with clinical history information.

not depend on preferences. In fact, it is calculable even in the absence of cases that are positive with very high probability. This feature is vital in our context, where 45 pathologies show no cases for which the consensus probability, $\bar{p}_{it}$, exceeds 0.5. Concordance will provide an informative signal about the performance of the classifier as long as there is some variation across cases in $\bar{p}_{it}$.

The main disadvantage to using concordance is that it is an ordinal measure of performance. As an ordinal classifier that compares two cases, a translation to treatment recommendation is not immediate.

An alternative measure of performance that we consider in the appendix is the deviation from consensus probability, $|p_{irt} - \bar{p}_{it}|$. Like concordance, it is calculable for all pathologies. However, as a performance measure it is misleading for low prevalence pathologies – those with a distribution of $\bar{p}_{it}$ concentrated at low values – because any classifier that predicts low probabilities will perform well and yield very small deviations. It therefore under-weights performance in the long-tail relative to concordance. Some of our conclusions are therefore sensitive to the use of this alternative.

III. Overall Performance

We now turn to comparing the classification performance of human radiologists to the CheXzero and CheXpert algorithms.

Table 1 summarizes the data and overall performance. The average prevalence, defined as the average value of $\bar{p}_{it}$ across pathologies and cases, is low at approximately 2.42%. The distribution of its mean across pathologies is heavily skewed, with the prevalence $\bar{p}_t$ of some pathologies exceeding 15%.

Radiologists perform worse than CheXzero and CheXpert with an average concordance of 0.58. CheXpert performs slightly better than CheXzero. However, the comparison between AIs should be interpreted with caution, as CheXpert only has predictions for 12 pathologies, while CheXzero has predictions for 79.

One hypothesis is that human and AI performance is highly correlated – pathologies in which humans perform well would be the same as those in which CheXzero does well. Figure 2 examines this relationship. Pathologies in the figure are sorted by CheXzero performance, each point represents a pathology overlayed with a local regression. Interestingly, we find that human performance is only weakly correlated with CheXzero performance. This suggests that CheXzero and human radiologists focus on different features of an X-ray. Humans still outperform zero-shot algorithms on select pathologies, but there are few of them.

Another important takeaway from the figure is that CheXzero's performance is more varied than that of humans – its concordance across pathologies spans the range from 0.45 and 0.94, while human concordance lies in a narrower range between 0.52

Table 1—Summary statistics

	Mean	Std. Dev.
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	
Number of radiologists	227	
Number of pathologies	79	

Note: Means and standard deviations are calculated across pathologies, except for the number of reads per radiologist, which is calculated across radiologists. Prevalence is a calculated from the consensus probability and multiplied by 100.

and 0.72.

IV. The Long Tail

Figure 3 assess the importance of the long tail. It shows the empirical CDF of prevalence ($\bar{p}_t$) across pathologies. The pathologies are arranged so that those with CheXpert assessments appear before others. Within the two groups, the pathologies are ordered by prevalence. The twelve pathologies with CheXpert assessments together constitute less than 60% of the overall prevalence. Thus, a significant proportion of relevant pathologies are not predicted by the supervised learning algorithm that we study. These, together with the four most prevalent pathologies covered by CheXzero, constitute 80% over the overall prevalence.

We next examine whether or not humans out-perform the AI algorithms for low prevalence pathologies. Figure 4 compares human performance to the performance of the two algorithms across bins of pathologies classified by low, medium, and high prevalence in our sample of cases. The bar charts represent the mean concordance, and the error bars show 95% confidence intervals.

Both human and AI performance increases with pathology prevalence. CheXpert displays the largest improvements when moving from medium to high prevalence. Although not reported, a test for the difference in performance between the high prevalence and either the medium or low prevalence bin is statistically significant at the 1% level for CheXpert. This increase may be expected because CheXpert relies heavily on both the quantity and quality of training data. Humans display modest and consistent improvements as prevalence increases, with all pairwise differences in performance between bins being statistically significant. CheXzero's performance is notably less sensitive to prevalence, perhaps due to the zero shot learning method it relies on. In fact, the differences in performance are either insignificant or marginally significant at the 1% level between any of the two pathology groups.

Within any pathology group, CheXzero outperform humans, with statistically significant differences in each group. Humans perform significantly worse than CheXzero in the low prevalence bin and only marginally better than a purely random classifier, providing initial evidence against the argument that humans will remain relevant in the long tail of diseases.

We next turn to analyzing the effect of the long tail on the overall concordance of humans versus the two algorithms we consider. Figure 5 shows how the overall concordance of these modes of diagnosis varies as the number of pathologies that we rely on them for increases. In each case, the concordance is set to 0.5, which corresponds to a random guess, for any pathology that is not assessed. Thus, the graph begins at 0.5 if no pathologies are assessed and increases to the average concordance. Since CheXpert only produces assessments for twelve out of the 79 pathologies, its graph ends

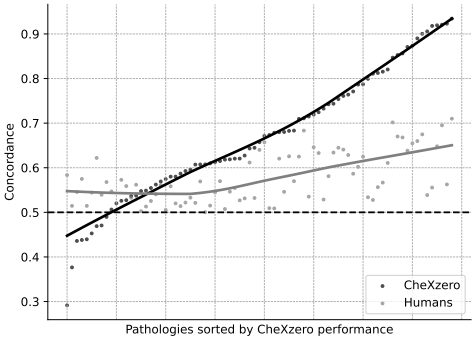

Figure 2. Correlation between human and CheXzero performance

Note: This figure compares human to AI concordance with pathologies sorted by CheXzero concordance. Each scatter point represents a pathology, and the locally weighted regression curve estimated using lowess are displayed.

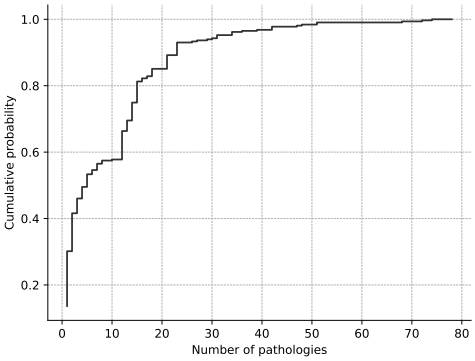

Figure 3. Empirical CDF of prevalence

Note: This figure plots the empirical CDF for the total share of positive cases. Pathologies are ordered from most prevalent to least, with the condition that pathologies with CheXpert reads are displayed first.

early. The pathologies in this figure are arranged the same as in figure 3.

Figure 3 shows that performance in the long tail is critical for assessing the overall quality of an algorithm or human radiologists. While CheXpert beats humans and matches CheXzero for pathologies for which it makes predictions, its ability to only predict a smaller subset of pathologies hinders its overall performance. Considering all pathologies, its concordance is less than 0.54 which indicates that the table 1 estimate of 0.72, which was calculated only on the twelve pathologies, is severely overestimated. Indeed, when considering all pathologies, its performance is lower than that of human radiologists.

Perhaps the most important takeaway from the figure that deserves discussion is that CheXzero performance is significantly higher than human performance, suggesting that the AI may have humans beat even in the long tail (c.f. 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 several additional steps, such as calibrating the algorithm and de-

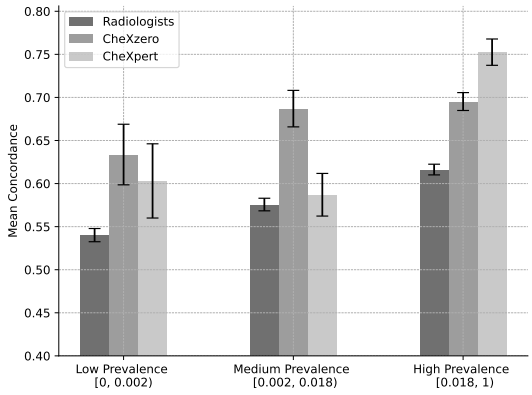

FIGURE 4. HUMAN VS AI PERFORMANCE

Note: This bar chart compares the concordance between humans and AI separately for low, medium, and high prevalence pathologies, and the lower and upper bounds of each bin are displayed on the x-axis. Humans and CheXzero have reads for all the pathologies, so there are 26 pathologies in each bin. CheXpert has reads for 12 pathologies, and there are 4 pathologies in each CheXpert bin. Bootstrapped standard errors computed separately for each bin are used to calculate 95% confidence intervals. For CheXpert and CheXzero, we use a block bootstrap, in which cases are drawn to account for correlations in performance across pathologies within a case. For human radiologists, we use a block bootstrap in which radiologists are drawn followed by cases.

termining an appropriate threshold for decisions. These tasks are particularly challenging for uncommon pathologies.

V. 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 the large number of uncommon cases – the long tail. Zero-shot learning algorithms are one attempt to make progress in the long tail by avoiding the need for large data sets with specifically annotated labels.

We compared the quality of 227 radiologists' assessments on 79 pathologies to two leading algorithms for diagnosing chest pathologies using X-rays – CheXpert, a supervised learning algorithm capable of assessing twelve pathologies, and CheXzero, which can produce assessments for any pathology. Our results suggest that self-supervised algorithms are quickly catching up or surpassing humans in the long tail of diseases in terms of predictive ability.

Yet, there are a number of hurdles remaining before algorithms, even those based on zero-shot learning methods, are deployed or result in job displacement. The output of the algorithm doesn't immediately yield either probabilities, recommendations or decisions. Perhaps more importantly, prediction is just one task in a job (Agrawal, Gans and Goldfarb (2019)). For this and other reasons, it is possible that these tools are more likely to provide assistance to humans, as opposed to replace them. In our opinion, factors that determine the optimal use of predictive AI tools are understudied and a fruitful avenue for research in economics.

REFERENCES

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

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

Casler, Krista, and Deborah Kelemen. 2005. "Young children's rapid

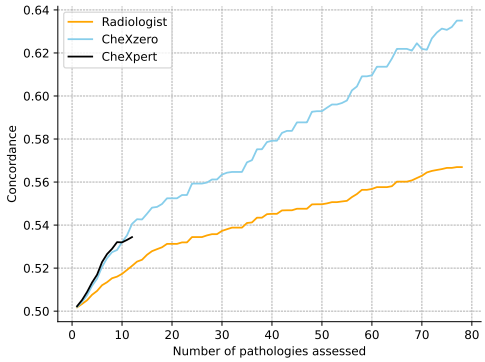

FIGURE 5. CONCORDANCE VS NUMBER OF PATHOLOGIES ASSESSED

Note: This figure compares the performance of human versus AI depending on the number of pathologies assessed. 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. The number of pathologies assessed are ordered from most prevalent to least, with the condition that pathologies with CheXpert reads are displayed first.

learning about artifacts." *Dev. Sci.*, 8(6): 472–480.

Chomsky, Noam. 1986. "Knowledge of language: Its nature, origin, and use." 97(4): 567.

Coutanche, Marc N, and Sharon L Thompson-Schill. 2014. "Fast mapping rapidly integrates information into existing memory networks." *J. Exp. Psychol. Gen.*, 143(6): 2296–2303.

Irvin, Jeremy, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, Jayne Seekins, David A Mong, Safwan S Halabi, Jesse K Sandberg, Ricky Jones, David B Larson, Curtis P Langlotz, Bhavik N Patel, Matthew P Lungren, and Andrew Y Ng. 2019. "CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison." Vol. 33, 590–597.

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." *Sci Data*, 6(1): 317.

Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2017. "Human Decisions and Machine Predictions." *Q. J. Econ.*, 133(1): 237–293.

Kühl, Niklas, Marc Goutier, Lucas Baier, Clemens Wolff, and Dominik Martin. 2022. "Human vs. supervised machine learning: Who learns patterns faster?" *Cogn. Syst. Res.*, 76: 78–92.

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."

Lake, Brenden, Ruslan Salakhutdinov, Jason Gross, and Joshua Tenenbaum. 2011. "One shot learning of simple visual concepts." Vol. 33.

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): 436–444.

Liu, Henry X, and Shuo Feng. 2022. "" Curse of rarity" for autonomous vehicles." *arXiv preprint arXiv:2207.02749*.

Liu, Xiaoxuan, Livia Faes, Aditya U Kale, Siegfried K Wagner, Dun Jack Fu, Alice Bruynseels, Thushika Mahendiran, Gabriella Moraes, Mohith Shamdas, Christoph Kern, Joseph R Ledsam, Martin K Schmid, Konstantinos Balaskas, Eric J Topol, Lucas M Bachmann, Pearse A Keane, and Alastair K Denniston. 2019. "A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis." *The Lancet Digital Health*, 1(6): e271–e297.

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

Mullainathan, S, and Z Obermeyer. 2019. "A machine learning approach to low-value health care: wasted tests, missed heart attacks and mispredictions."

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

Sheng, Victor S, Foster Provost, and Panagiotis G Ipeirotis. 2008. "Get another label? improving data quality and data mining using multiple, noisy labelers." *KDD '08*, 614–622. New York, NY, USA: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." *Nat Biomed Eng*, 6(12): 1399–1406.

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