Comparative Advantage of Humans vs AI in the Long Tail
Online Appendix

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

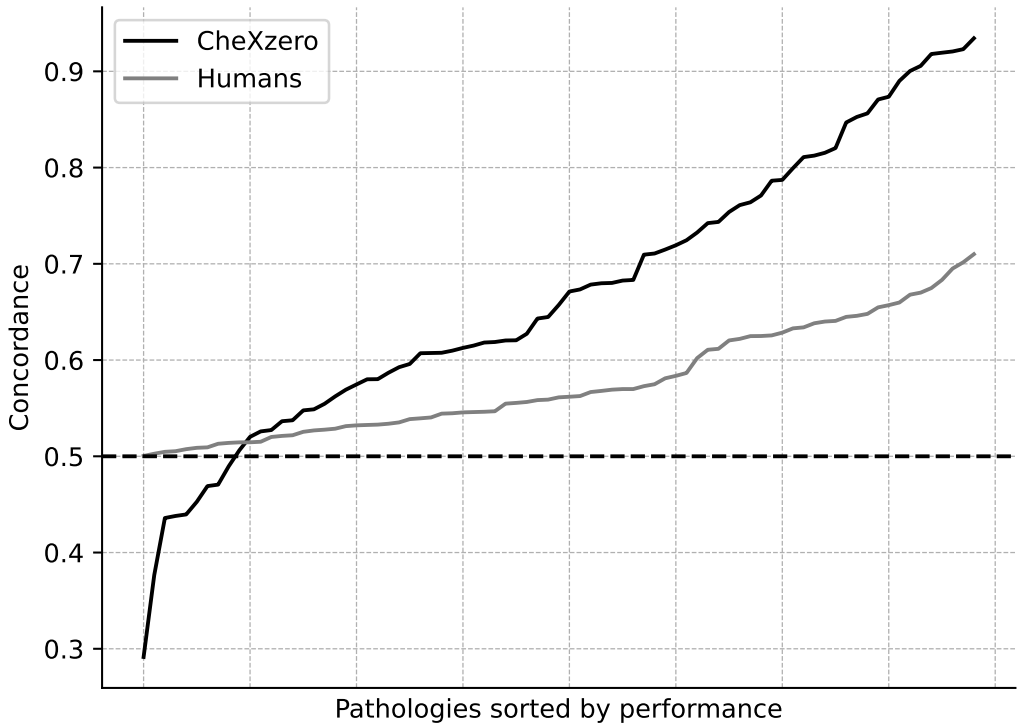

Figure 1: Correlation between human and CheXzero performance

Note: This figure compares human to AI concordance. In the CheXzero line, the pathologies are sorted by CheXzero concordance, while in the human line, the pathologies are sorted by human performance.

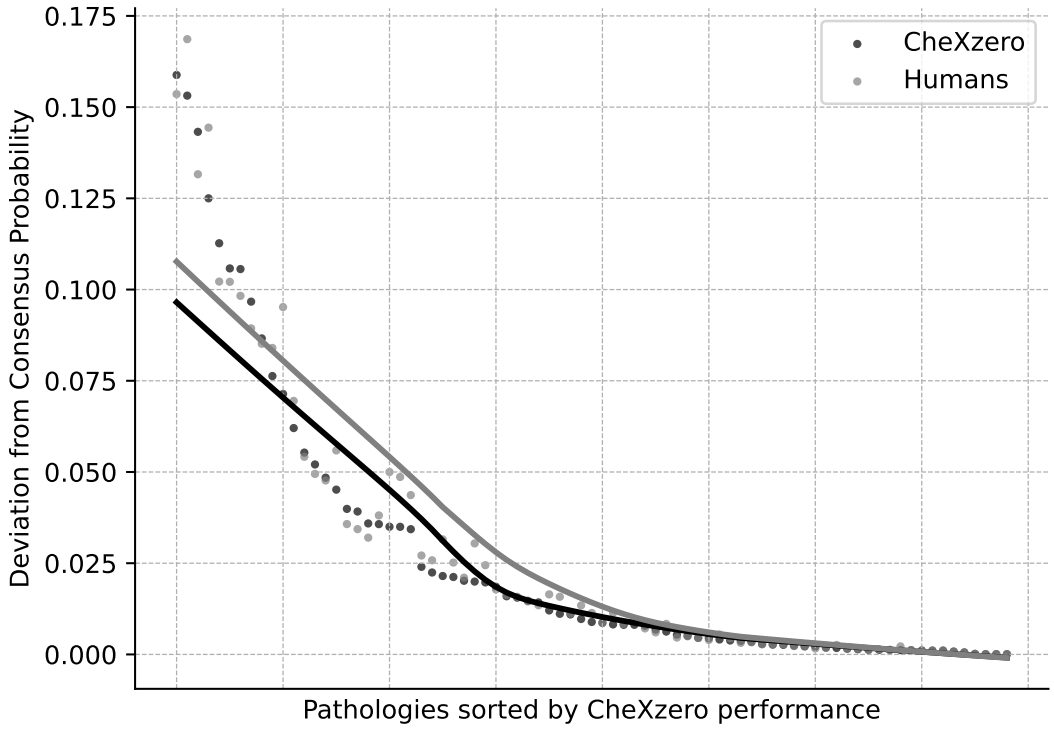

Figure 2: Correlation between human and CheXzero performance

Note: This figure compares human to AI deviation from consensus probability with pathologies sorted by CheXzero concordance. Each scatter point represents a pathology, and the locally weighted scatter plot smoothing (LOWESS) estimates are displayed.

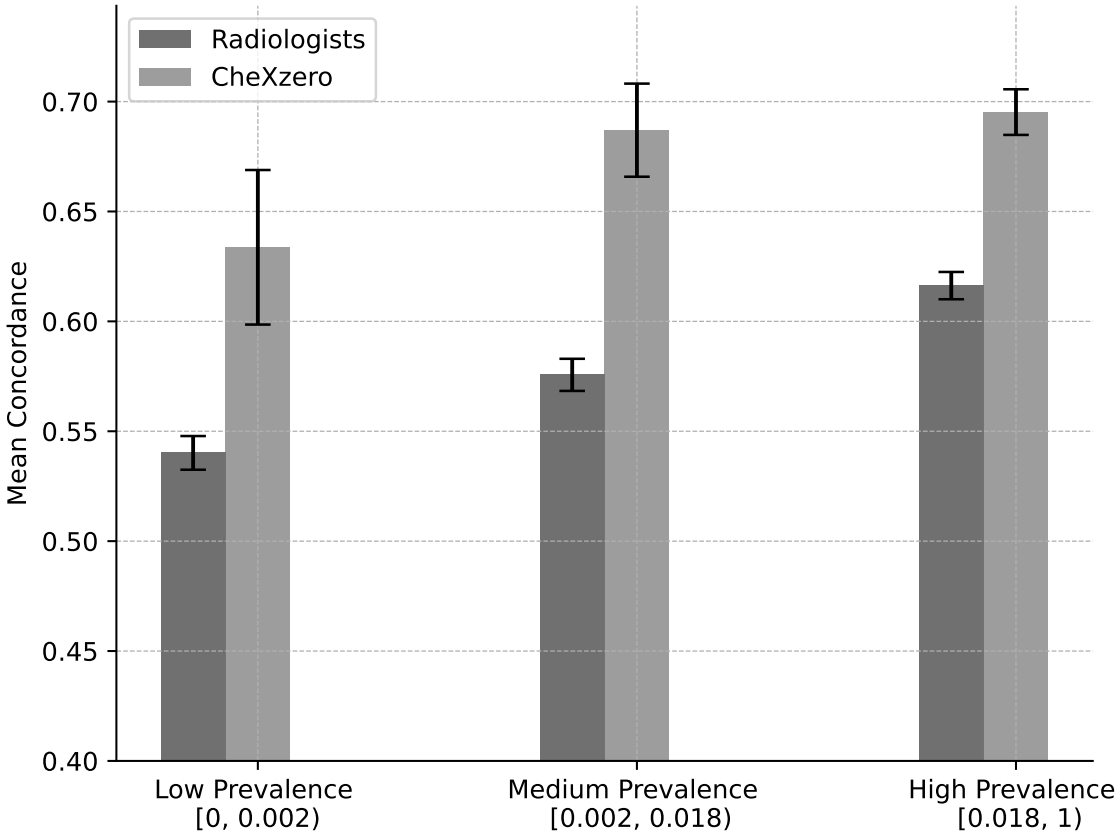

Figure 3: Human vs AI Concordance

Note: This bar chart is contains 79 pathologies for which there exist radiologist and CheXzero predictions, resulting in 26 pathologies per bin. 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. Bootstrapped standard errors computed separately for each bin are used to calculate 95% confidence intervals. For 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.

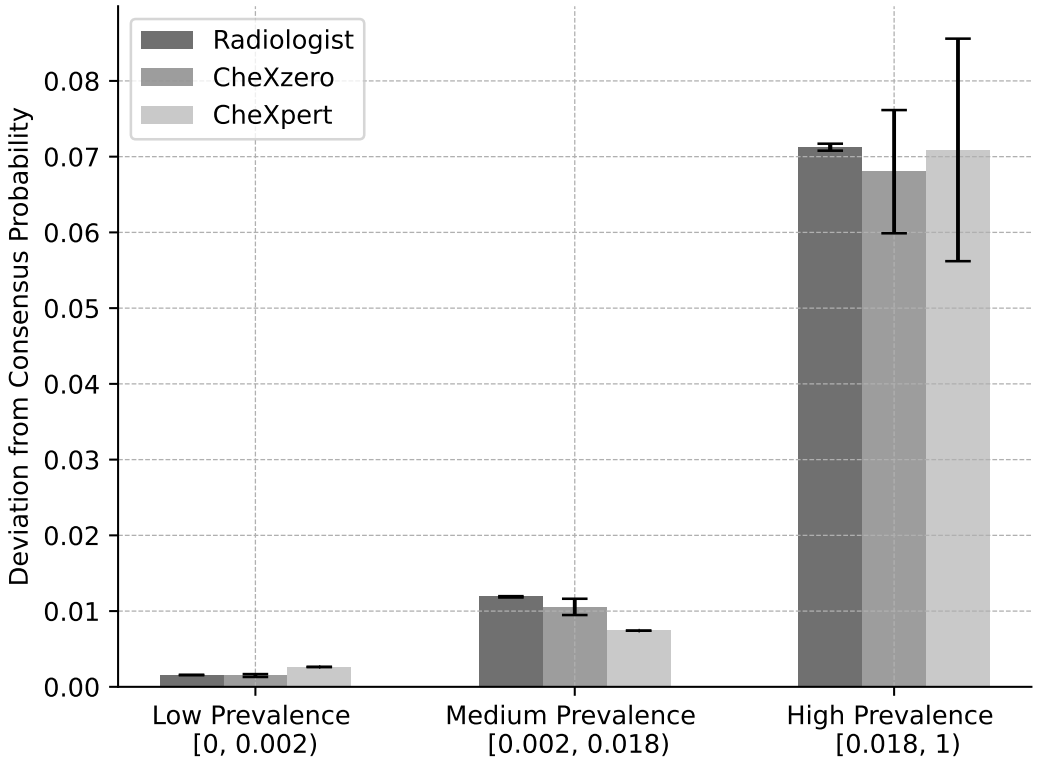

Figure 4: Human vs AI Deviation from Consensus Probability

Note: This bar chart compares the deviation from consensus probability 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. Clustered standard errors are used to calculate 95% confidence intervals.

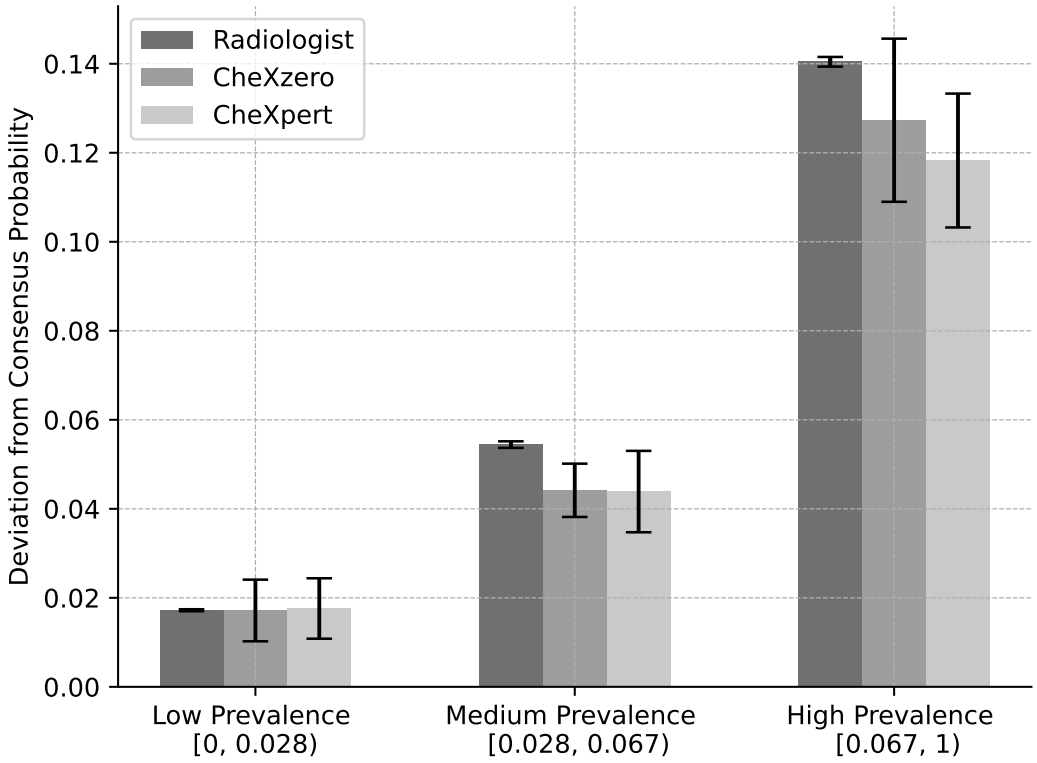

Figure 5: Human vs AI Deviation from Consensus Probability

Note: This bar chart is limited to pathologies for which there exist CheXpert predictions, resulting in four pathologies per bin. This bar chart compares the deviation from consensus probability 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. Clustered standard errors are used to calculate 95% confidence intervals.

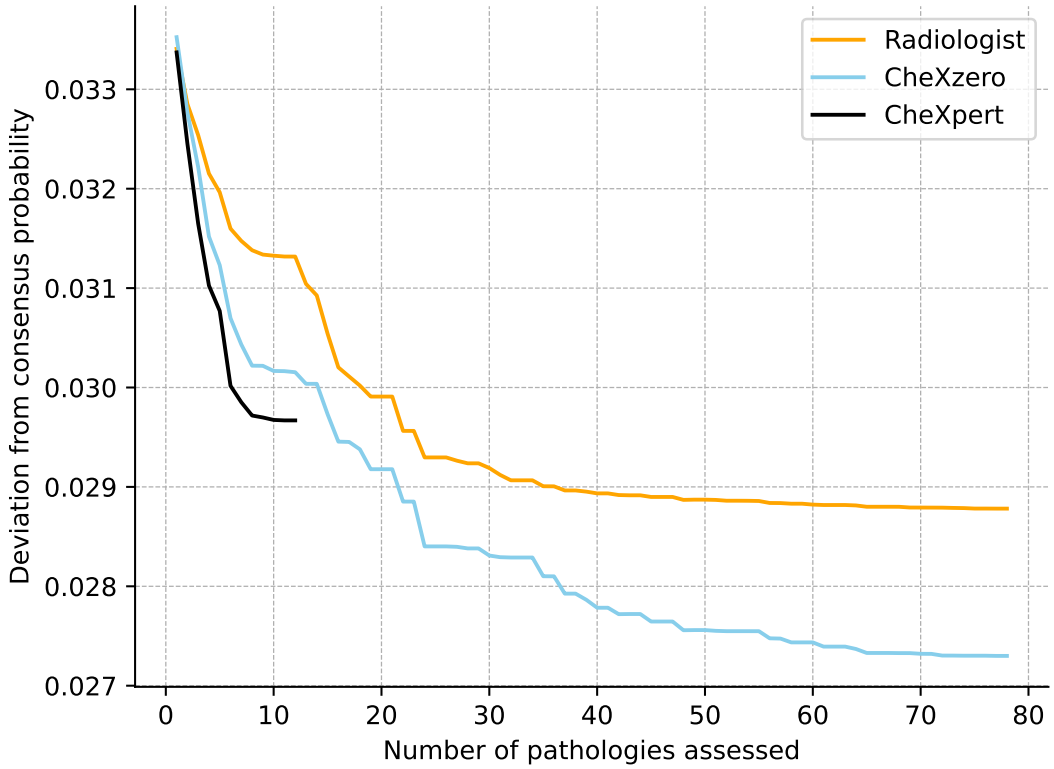

Figure 6: Deviation vs Number of Pathologies Assessed

Note: This figure compares the performance of humans versus AI depending on the number of pathologies assessed. The y-axis concordance is computed as the average between the deviation from consensus probability of the pathologies assessed and the prior (i.e. absolute difference between the pathology's prevalence and diagnostic standard for that case) for the pathologies without assessments. The number of pathologies assessed is ordered from most prevalent to least, with the condition that pathologies with CheXpert reads are displayed first.

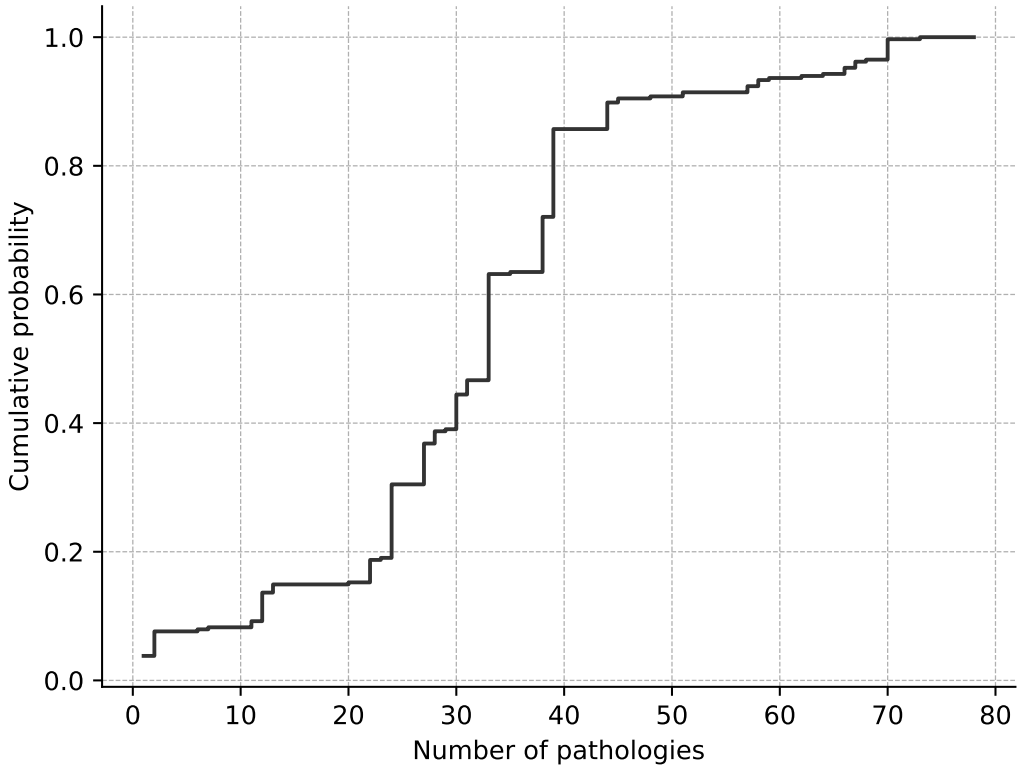

Figure 7: Empirical CDF: Total Share of Positive Cases

Note: This figure plots the empirical CDF for the total share of positive cases. Pathologies are ordered by CheXzero concordance from greatest to smallest, with the condition that pathologies with CheXpert reads are displayed first.