

EP09.13 Deep learning for fetal congenital heart disease detection in a community screening setting

Laura Gambini, Connor Lau, Adrienne Lightbourne, Wayne Shi, Yili Zhao, Omar Arshad, Ben Li, **Anita J Moon-Grady**, Rima Arnaout. University of California San Francisco, Departments of Pediatrics and Medicine

Introduction

Deep learning (DL) has the potential to improve fetal ultrasound screening for complex congenital heart disease (CHD). Improved detection could have the most impact in the community setting, where clinical expertise varies; however, DL models often fail to generalize to diverse populations. We tested DL’s ability to detect CHD in a community-based practice.

Methods

We applied our own DL model retrospectively to all second trimester obstetric screening ultrasounds (n=1002) from a single community perinatology practice over one year. We calculated sensitivity, specificity, and negative predictive value (NPV) overall and by race, ethnicity, BMI, and gestational week.

Results

86.1% of pregnancies presented in the recommended 18-24 week window; 13.9% presented late (25-32 weeks). Demographics are shown in Figure 1. Five of 752 fetuses (0.66%) had CHD confirmed at birth (atrioventricular septal defect 2, coarctation 1, double outlet right ventricle 1, d-transposition 1). 250 postnatal outcomes are unknown thus excluded from sensitivity/specificity analyses.

Results (cont.)

Clinically 2.4% (18) of the 752 patients were referred for fetal echocardiogram due to abnormal obstetric screening exam; 77.8%(14) of these were normal (false positive referrals) and 4 were abnormal; one infant with CHD was missed (false negative). Sensitivity and specificity of the original clinical ultrasound screening were thus 80% and 99%; NPV was 99%. **Our DL model sensitivity and specificity were 100% and 98% (AUC 0.99, NPV 100%).** Only 0.8% of studies lacked sufficient information for DL prediction. Model performance did not differ by race, ethnicity, or obesity (MWU p-values 0.09-0.96) but had lower specificity for those presenting late for screening (specificity 0.93, p= 0.01).

Conclusion

We show that our DL model can generalize to imaging from a community-based practice serving a diverse population. The model performed well across race, ethnicity, BMI, and timing of presentation, supporting its potential for improving CHD detection during routine screening.

Supported by grants from the NIH to Rima Arnaout and by a generous gift from Georges Harik and Christine Hahn. The authors report no disclosures.

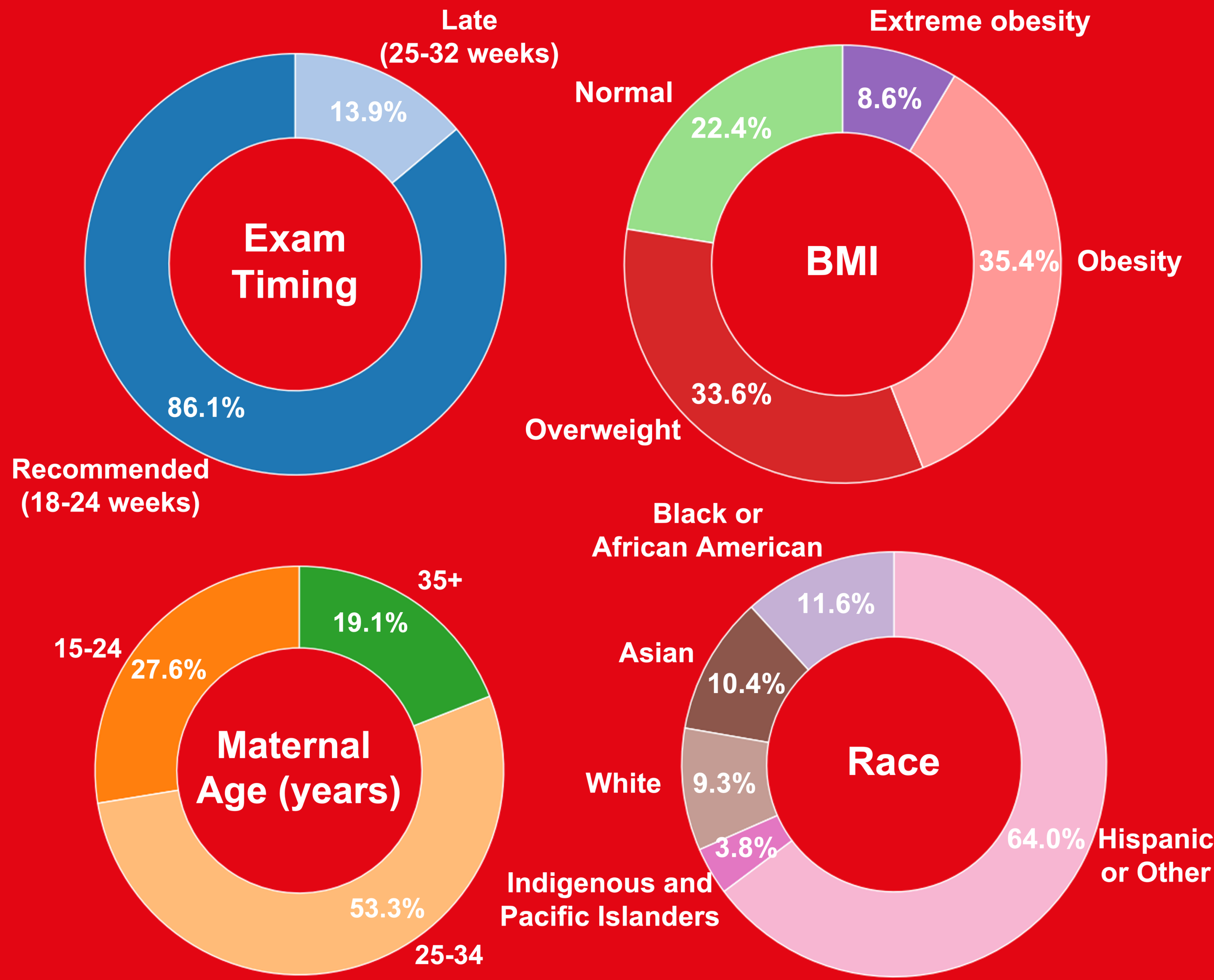


Figure 1. Population Demographics

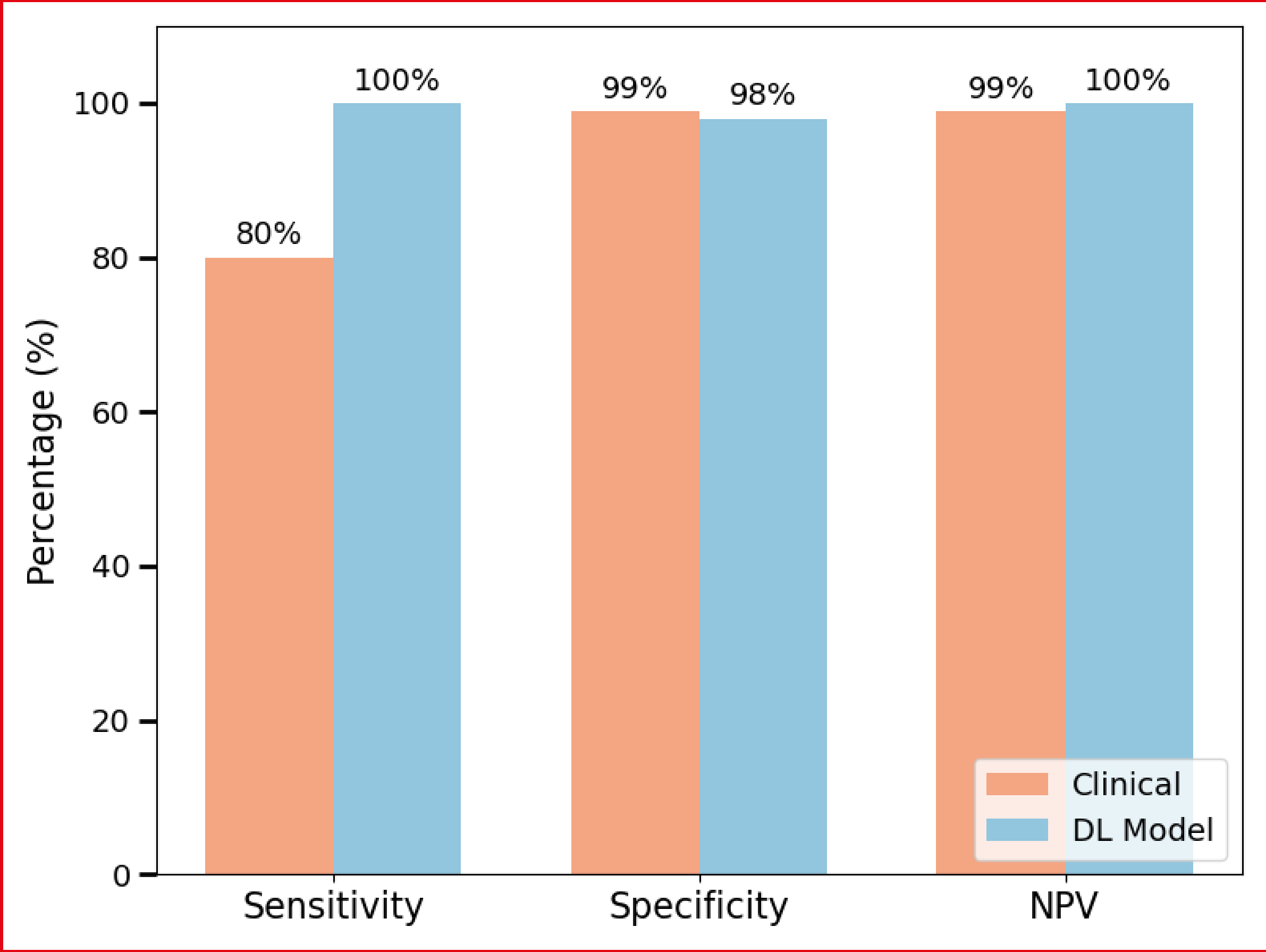


Figure 2. Performance comparison: clinical vs deep-learning model