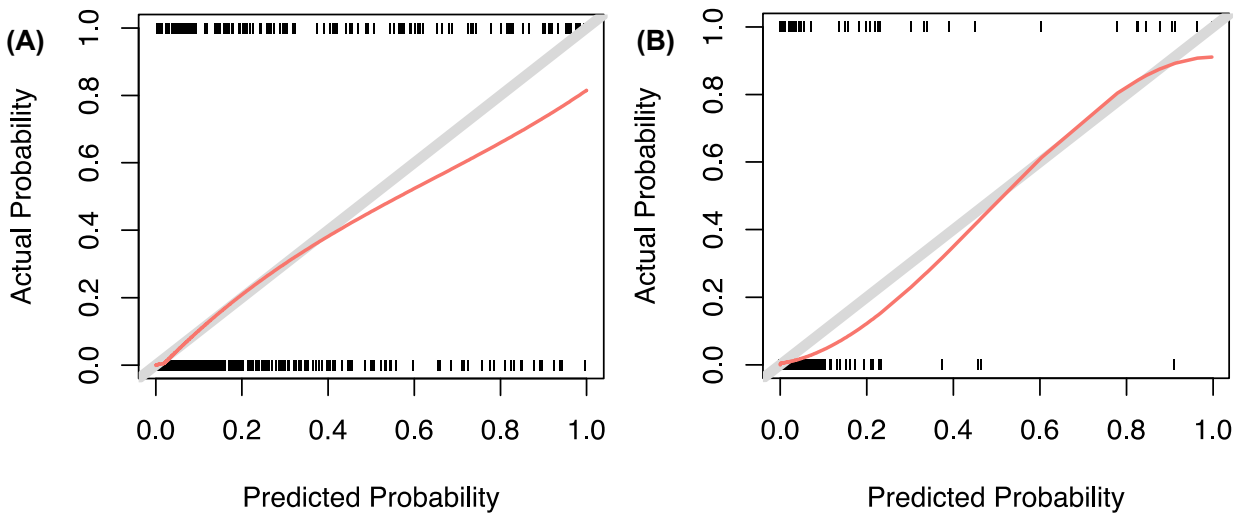

Figure 2. Calibration curves for the deep neural network (DNN). The Cincinnati Children's Hospital Medical Center dataset is on the left (**A**) and the University of Cincinnati dataset is on the right (**B**). Perfect calibration is denoted by the grey identity line, y = x. Tick marks represent the DNN's estimated probability for non-surgical and surgical patients at y = 0 and y = 1, respectively. "Predicted Probability" is the DNN's estimated probability that the patient was a surgical candidate. For a given value of "Predicted Probability", the "Actual Probability" is the observed proportion of patients who underwent surgical treatment. In (**A**), the average difference between the calibration curve and the identity line for all patients was 0.81%, the maximum difference was 18.5%, and the 90% quantile was 1.32%. For (**B**), the average difference between the calibration curve and the identity line was 0.19%, the maximum difference was 8.64%, and the 90% quantile was 0.33%.