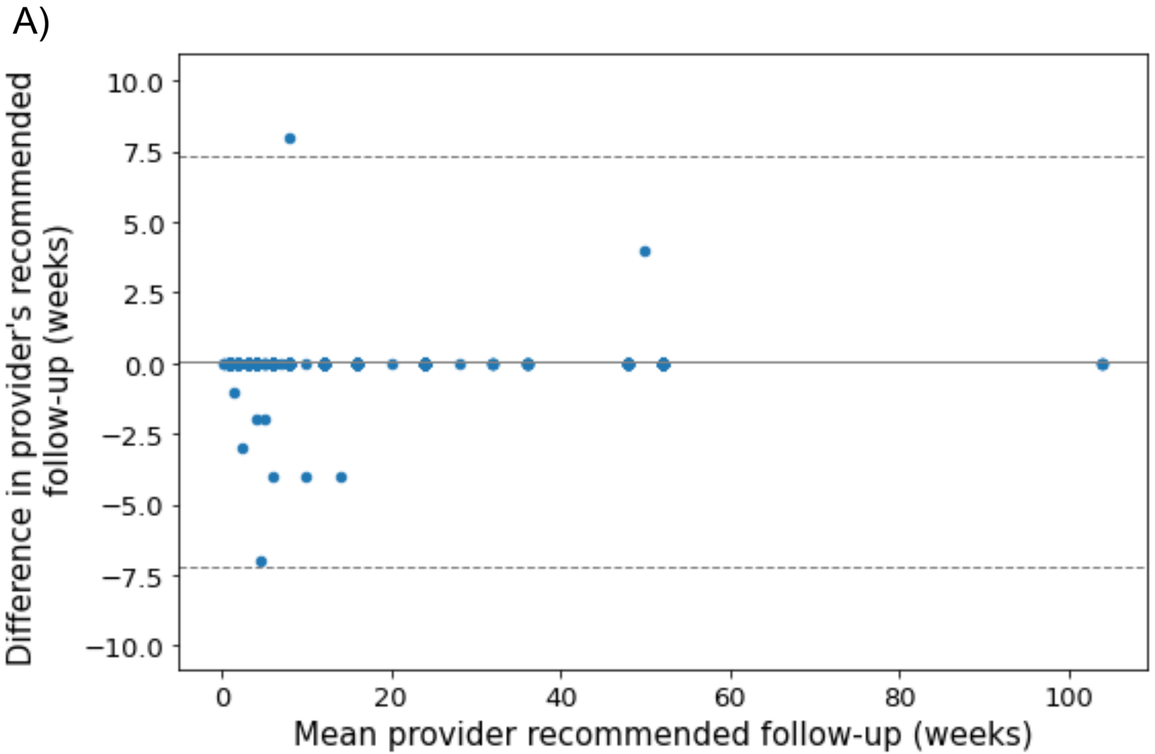

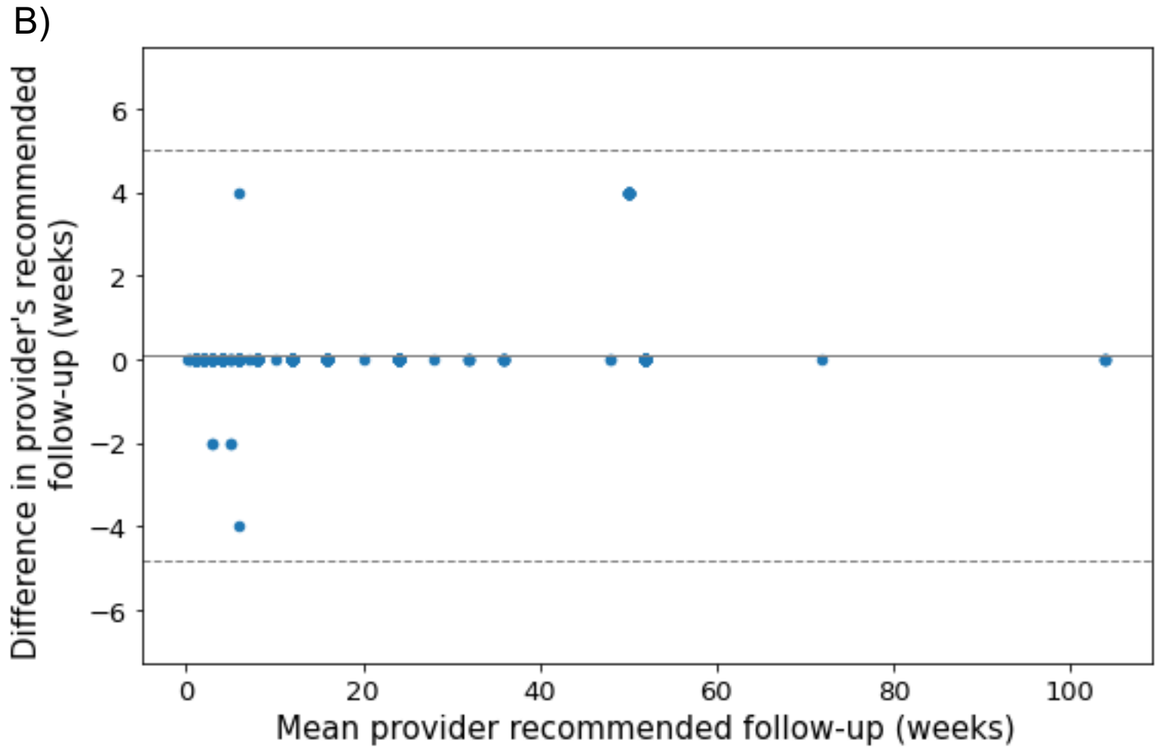

Supplemental Figure 3: Bland-Altman plot showing the difference in extraction of providers' recommended follow-up timeframe by the NLP algorithm and gold standard grader (A), and between the two graders (B). The mean difference in timeframe extraction is shown in the black solid line and the 95% limits of agreement are shown in the black dashed line. (A) The mean difference between the NLP algorithm and gold standard grader is 0.05 weeks, and the 95% limits of agreement are between -7.20 and +7.30 weeks. (B) The mean difference between another grader and the gold standard grader is 0.09 weeks, and the 95% limits of agreement are between -4.82 and +5.00 weeks.