Appendix A All Significant Results

Table 6 in Section 4 only included the statistically relevant results relevant to the analysis and discussion. Table A1 lists *all 24* statistically significant results obtained from the 150 comparisons that were made.

Task	**Weaker model**		**Stronger model**		**p-value**
	P	**F**	**P**	**F**	
ICD-10	✗	✓	✗	✗	0.0378
Factuality NER	✗	✓	✗	✗	0.0014
ICD-10	✗	✓	✓	✓	0.0014
Clinical NER	✗	✓	✓	✓	0.0070
ICD-10	✗	✓	✓	+	0.0002
Clinical NER	✗	✓	✓	+	0.0320
Factuality NER	✗	✓	✓	+	0.0002
ICD-10	✗	✓	✓	✗	0.0007
Clinical NER	✗	✓	✓	✗	0.0029
Factuality NER	✗	✓	✓	✗	0.0005
Clinical NER	✗	+	✗	✗	0.0269
ICD-10	✗	+	✓	✓	0.0057
ADE Classification	✗	+	✓	✓	0.0320
Clinical NER	✗	+	✓	✓	0.0029
ICD-10	✗	+	✓	+	0.0006
ADE Classification	✗	+	✓	+	0.0481
Clinical NER	✗	+	✓	+	0.0129
Factuality NER	✗	+	✓	+	0.0188
ICD-10	✗	+	✓	✗	0.0011
Clinical NER	✗	+	✓	✗	0.0023
Factuality NER	✗	+	✓	✗	0.0156
ADE Classification	✗	✗	✓	✓	0.0078
ICD-10	✗	✗	✓	✗	0.0086
Clinical NER	✗	✗	✓	✗	0.0226

Table A1 All 24 statistically significant results are listed alongside the task, the model configurations and the *p-values*. Similarly to Table **??** and Table **??**, **P** denotes whether pre-training was done using pseudonymized data, and **F** if fine-tuning was done using pseudonymized data. As in the previously mentioned tables, a ✗ denotes that no pseudonymization was done, a ✓ that it was done using the *pseudo* model and a + means that pseudonymization was performed using the *pseudo+* model.