Selecting Relevant Features from the Electronic Health Record for Clinical Code Prediction: Supplementary Materials

As appendix, we present the pseudocode for coverage, and an analysis of clinical code prediction results with three different machine learning algorithms – Random Forests, C4.5 decision trees, and Naive Bayes – for three randomly selected specialties. We show predictions for both ICD-9-CM procedure and diagnostic codes. All results were obtained using 10-fold cross-validation, the combination of data sources is the same within a specialty. Scores represent the micro F-measure.

1. Pseudocode for the coverage algorithm

```
for each c in classes do
  classPositiveDataset = getSamplesContainingClass(c)
  for each f in features do
    confidences[f] = calculateConfidence(f, c)
  end for
  confidences.sortDescending

  for each conf in confidences do
    coveredSamples = coveredSamples(conf, classPositiveDataset)
    if coveredSamples.length > 0 then
      for each covSam in coveredSamples do
        covSam.coverage ++
        if covSam.coverage >= 3 then
          classPositiveDataset.remove(covSam)
        end if
      end for
      addFeatureToSelectionList(f)
    end if
```

end for
end for

2. Significance matrix of the results

In table 1, the statistical significance of the results is calculated against the baseline with approximate randomization tests. Since each code is predicted independently (with an independent model), the shown significance is averaged over all codes for which models were generated. The selected baseline for each experiment has the same input sources as the tested experiment. We show the significance of all experiments for a single source, and for total integration. Each column represents a dataset, and each row represents a feature selection method.

The significance score indicates the similarity between the results of two different systems. Lower is better, and we can assume that the models are independent when the significance score is lower than 0.05 with 95% certainty.

Table 1: the similarity of the experiments according to approximate random testing against the baseline. For each experiment with a single source and total integration, the significance of the similarity is shown. Each column represents a dataset, and each row represents a feature selection method.

	UZA1						UZA2		MIMIC-III
	Cardio	**Gastro**	**Onco**	**Ophtal**	**Pneumo**	**Uro**	**Cardio**	**Onco**	**ICU**
Diagnostic codes									
GainRatio single source	0.035	0.046	0.029	0.041	0.020	0.014	0.047	0.044	0.050
GainRatio total integration	0.041	0.052	0.051	0.033	0.046	0.047	0.021	0.040	0.050
ConfCov single source	0.030	0.030	0.038	0.038	0.008	0.038	0.038	0.018	0.033
ConfCov total integration	0.028	0.012	0.005	0.044	0.014	0.040	0.011	0.004	0.035
ConfCov-Neg single source	0.030	0.030	0.038	0.040	0.008	0.025	0.044	0.029	0.033
ConfCov-Neg total integration	0.028	0.012	0.009	0.045	0.017	0.036	0.014	0.004	0.035
InfGainCov single source	0.025	0.015	0.035	0.036	0.025	0.029	0.034	0.033	0.029
InfGainCov total integration	0.027	0.024	0.006	0.047	0.020	0.043	0.017	0.004	0.033
Procedure codes									
GainRatio single source	0.031	0.037	0.036	0.013	0.029	0.059	0.034	0.034	0.038
GainRatio total integration	0.009	0.013	0.028	0.003	0.031	0.043	0.017	0.064	0.038
ConfCov single source	0.030	0.032	0.045	0.034	0.024	0.061	0.027	0.059	0.021
ConfCov total integration	0.009	0.015	0.013	0.029	0.031	0.044	0.017	0.027	0.025
ConfCov-Neg single source	0.020	0.020	0.022	0.035	0.024	0.054	0.024	0.027	0.021
ConfCov-Neg total integration	0.003	0.012	0.018	0.018	0.031	0.036	0.016	0.030	0.025
InfGainCov single source	0.035	0.026	0.037	0.020	0.038	0.056	0.034	0.054	0.029
InfGainCov total integration	0.011	0.037	0.003	0.021	0.033	0.064	0.034	0.048	0.027

3. Results of integration with multiple classifiers and multiple techniques

For Random Forests, table 2 represents integration with confidence coverage, table 3 represents integration with confidence coverage, including both positive and negative feedback, table 4 represents integration with confidence as a scoring mechanism where the top 50 features for each class were selected, table 5 represents integration with gain ratio, and table 6 represents integration with information gain-coverage.

Table 2: Integration with confidence-coverage, using a Random Forest Classifier. Rows represent the number of used sources, columns represent the code system and the specialty, respectively.

# sources	Diag, Opht	Diag, Pneu	Diag, Uro	Proc, Opth	Proc, Pneu	Proc, Uro
1	0.538	0.447	0.324	0.9	0.748	0.653
2	0.541	0.462	0.336	0.899	0.751	0.67
3	0.545	0.461	0.352	0.906	0.765	0.673
4	0.553	0.521	0.379	0.909	0.768	0.681
5	0.552	0.487	0.382	0.909	0.768	0.683
6	0.552	0.488	0.382	0.909	0.77	0.682
7	0.552	0.508	0.416	0.909	0.769	0.684
8	0.554	0.523	0.44	0.908	0.77	0.681

For naive Bayes, table 7 represents integration with confidence coverage, table 8 represents integration with confidence coverage, including both positive and negative feedback, table 9 represents integration with gain ratio, and table 10 represents integration with information gain-coverage.

For C4.5, table 11 represents integration with confidence coverage, table 12 represents integration with confidence coverage, including both positive and negative feedback, table 13 represents integration with confidence as a scoring mechanism where the top 50 features for each class were selected, table 14 represents integration with gain ratio, and table 15 represents integration with information gain-coverage.

Table 3: Integration with confidence-coverage with positive and negative feedback, using a Random Forest Classifier. Rows represent the number of used sources, columns represent the code system and the specialty, respectively.

# sources	Diag, Opht	Diag, Pneu	Diag, Uro	Proc, Opth	Proc, Pneu	Proc, Uro
1	0.537	0.447	0.316	0.9	0.748	0.659
2	0.538	0.461	0.332	0.9	0.751	0.661
3	0.548	0.459	0.354	0.906	0.765	0.676
4	0.553	0.509	0.382	0.909	0.768	0.681
5	0.554	0.485	0.383	0.909	0.769	0.683
6	0.554	0.485	0.385	0.909	0.77	0.683
7	0.554	0.496	0.412	0.908	0.769	0.681
8	0.554	0.515	0.44	0.908	0.77	0.685

Table 4: Integration with confidence, where the top 50 best features are selected per class, using a Random Forest Classifier. Rows represent the number of used sources, columns represent the code system and the specialty, respectively.

# sources	Diag, Opht	Diag, Pneu	Diag, Uro	Proc, Opth	Proc, Pneu	Proc, Uro
1	0.527	0.442	0.323	0.898	0.745	0.651
2	0.529	0.463	0.331	0.898	0.746	0.651
3	0.533	0.463	0.359	0.904	0.759	0.666
4	0.534	0.515	0.387	0.906	0.764	0.679
5	0.534	0.497	0.389	0.905	0.762	0.68
6	0.534	0.497	0.391	0.905	0.762	0.681
7	0.535	0.523	0.42	0.905	0.762	0.681
8	0.535	0.534	0.441	0.905	0.761	0.682

Table 5: Integration with gain ratio, where the top 50 best features are selected per class, using a Random Forest Classifier. Rows represent the number of used sources, columns represent the code system and the specialty, respectively.

# sources	Diag, Opht	Diag, Pneu	Diag, Uro	Proc, Opth	Proc, Pneu	Proc, Uro
1	0.524	0.404	0.286	0.902	0.748	0.643
2	0.524	0.419	0.301	0.903	0.740	0.646
3	0.529	0.421	0.34	0.905	0.744	0.668
4	0.53	0.481	0.345	0.909	0.74	0.67
5	0.529	0.433	0.347	0.909	0.743	0.669
6	0.529	0.432	0.346	0.909	0.746	0.674
7	0.528	0.459	0.387	0.909	0.748	0.67
8	0.53	0.449	0.39	0.91	0.743	0.671

Table 6: Integration with information gain-coverage with positive and negative feedback, using a Random Forest Classifier. Rows represent the number of used sources, columns represent the code system and the specialty, respectively.

# sources	Diag, Opht	Diag, Pneu	Diag, Uro	Proc, Opth	Proc, Pneu	Proc, Uro
1	0.52	0.447	0.3	0.899	0.751	0.649
2	0.517	0.437	0.311	0.898	0.748	0.649
3	0.523	0.44	0.339	0.903	0.746	0.652
4	0.533	0.525	0.346	0.908	0.741	0.653
5	0.533	0.449	0.346	0.909	0.741	0.656
6	0.533	0.448	0.346	0.909	0.741	0.657
7	0.534	0.496	0.393	0.909	0.742	0.654
8	0.535	0.495	0.4	0.908	0.74	0.654

Table 7: Integration with confidence-coverage, using a naive Bayes Classifier. Rows represent the number of used sources, columns represent the code system and the specialty, respectively.

# sources	Diag, Opht	Diag, Pneu	Diag, Uro	Proc, Opth	Proc, Pneu	Proc, Uro
1	0.345	0.277	0.242	0.851	0.664	0.577
2	0.345	0.274	0.248	0.845	0.643	0.576
3	0.307	0.209	0.185	0.831	0.660	0.560
4	0.305	0.235	0.231	0.851	0.683	0.562
5	0.304	0.239	0.232	0.851	0.682	0.562
6	0.304	0.239	0.232	0.851	0.683	0.562
7	0.304	0.251	0.232	0.85	0.684	0.562
8	0.304	0.255	0.238	0.85	0.684	0.562

Table 8: Integration with confidence-coverage with positive and negative feedback, using a naive Bayes Classifier. Rows represent the number of used sources, columns represent the code system and the specialty, respectively.

# Sources	Diag, Opht	Diag, Pneu	Diag, Uro	Proc, Opht	Proc, Pneu	Proc, Uro
1	0.343	0.308	0.246	0.852	0.665	0.577
2	0.346	0.301	0.25	0.845	0.643	0.577
3	0.294	0.292	0.184	0.849	0.666	0.555
4	0.291	0.277	0.197	0.85	0.683	0.559
5	0.29	0.237	0.199	0.85	0.682	0.558
6	0.29	0.236	0.199	0.85	0.683	0.557
7	0.29	0.216	0.186	0.85	0.684	0.558
8	0.291	0.24	0.206	0.85	0.684	0.558

Table 9: Integration with gain ratio, where the top 50 best features are selected per class, using a naive Bayes Classifier. Rows represent the number of used sources, columns represent the code system and the specialty, respectively.

# sources	Diag, Opht	Diag, Pneu	Diag, Uro	Proc, Opth	Proc, Pneu	Proc, Uro
1	0.332	0.292	0.22	0.798	0.394	0.359
2	0.341	0.288	0.225	0.788	0.431	0.377
3	0.32	0.27	0.189	0.759	0.477	0.376
4	0.308	0.264	0.213	0.741	0.5	0.431
5	0.308	0.217	0.213	0.737	0.503	0.425
6	0.308	0.217	0.214	0.736	0.504	0.424
7	0.308	0.201	0.19	0.736	0.504	0.424
8	0.308	0.217	0.216	0.736	0.503	0.427

Table 10: Integration with information gain-coverage, using a naive Bayes Classifier. Rows represent the number of used sources, columns represent the code system and the specialty, respectively.

# sources	Diag, Opht	Diag, Pneu	Diag, Uro	Proc, Opth	Proc, Pneu	Proc, Uro
1	0.392	0.307	0.24	0.838	0.623	0.53
2	0.394	0.317	0.238	0.835	0.397	0.537
3	0.331	0.314	0.211	0.817	0.539	0.486
4	0.386	0.309	0.244	0.836	0.631	0.497
5	0.386	0.283	0.258	0.836	0.637	0.508
6	0.386	0.282	0.258	0.836	0.637	0.509
7	0.386	0.259	0.238	0.835	0.638	0.509
8	0.391	0.296	0.268	0.834	0.643	0.514

Table 11: Integration with confidence-coverage, using a C4.5 Classifier. Rows represent the number of used sources, columns represent the code system and the specialty, respectively.

# sources	Diag, Opht	Diag, Pneu	Diag, Uro	Proc, Opth	Proc, Pneu	Proc, Uro
1	0.526	0.357	0.276	0.895	0.737	0.64
2	0.526	0.412	0.275	0.895	0.737	0.64
3	0.528	0.413	0.299	0.905	0.741	0.638
4	0.524	0.477	0.328	0.904	0.745	0.637
5	0.525	0.451	0.33	0.904	0.747	0.641
6	0.525	0.451	0.33	0.904	0.743	0.641
7	0.525	0.5	0.36	0.904	0.744	0.641
8	0.527	0.516	0.386	0.904	0.744	0.641

Table 12: Integration with confidence-coverage with positive and negative feedback, using a C4.5 Classifier. Rows represent the number of used sources, columns represent the code system and the specialty, respectively.

# sources	Diag, Opht	Diag, Pneu	Diag, Uro	Proc, Opth	Proc, Pneu	Proc, Uro
1	0.526	0.358	0.276	0.896	0.737	0.645
2	0.529	0.414	0.278	0.895	0.737	0.642
3	0.529	0.417	0.3	0.905	0.741	0.642
4	0.525	0.477	0.329	0.904	0.745	0.641
5	0.525	0.45	0.333	0.904	0.747	0.642
6	0.525	0.45	0.332	0.904	0.743	0.642
7	0.525	0.498	0.365	0.904	0.744	0.642
8	0.524	0.514	0.389	0.904	0.744	0.642

Table 13: Integration with confidence, where the top 50 best features are selected per class, using a C4.5 Classifier. Rows represent the number of used sources, columns represent the code system and the specialty, respectively.

# sources	Diag, Opht	Diag, Pneu	Diag, Uro	Proc, Opth	Proc, Pneu	Proc, Uro
1	0.513	0.354	0.279	0.893	0.731	0.635
2	0.515	0.409	0.282	0.894	0.732	0.632
3	0.517	0.409	0.302	0.901	0.734	0.636
4	0.516	0.478	0.33	0.905	0.734	0.638
5	0.516	0.449	0.335	0.905	0.734	0.638
6	0.516	0.449	0.335	0.905	0.732	0.639
7	0.516	0.493	0.37	0.905	0.732	0.639
8	0.516	0.505	0.382	0.905	0.732	0.639

Table 14: Integration with gain ratio, where the top 50 best features are selected per class, using a C4.5 Classifier. Rows represent the number of used sources, columns represent the code system and the specialty, respectively.

# sources	Diag, Opht	Diag, Pneu	Diag, Uro	Proc, Opth	Proc, Pneu	Proc, Uro
1	0.519	0.32	0.262	0.892	0.731	0.633
2	0.513	0.374	0.271	0.892	0.728	0.631
3	0.516	0.376	0.3	0.898	0.727	0.644
4	0.516	0.443	0.316	0.898	0.732	0.644
5	0.517	0.442	0.319	0.897	0.731	0.639
6	0.517	0.442	0.318	0.897	0.733	0.639
7	0.516	0.478	0.36	0.897	0.732	0.639
8	0.516	0.49	0.372	0.897	0.733	0.64

Table 15: Integration with information gain-coverage with positive and negative feedback, using a C4.5 Classifier. Rows represent the number of used sources, columns represent the code system and the specialty, respectively.

# sources	Diag, Opht	Diag, Pneu	Diag, Uro	Proc, Opth	Proc, Pneu	Proc, Uro
1	0.511	0.356	0.257	0.891	0.735	0.635
2	0.509	0.374	0.258	0.891	0.736	0.635
3	0.513	0.38	0.286	0.9	0.729	0.627
4	0.519	0.473	0.3	0.902	0.722	0.635
5	0.518	0.417	0.3	0.901	0.723	0.632
6	0.518	0.417	0.3	0.901	0.725	0.631
7	0.52	0.483	0.346	0.901	0.726	0.631
8	0.514	0.489	0.354	0.901	0.725	0.629