Appendix B

In this section, we explored the distribution of association rankings as a function of an ICD-9 code's frequency in the original clinical dataset. For this experiment we selected three ICD-9 codes from each of five different percentile rankings in the Clinical dataset: 1st, 25th, 50th, 75th, and 100th percentile. Note that the percentile rankings here are based on the frequency of the ICD-9 code among the patients and not an association, which was the ranking discussed in the main manuscript. Further, we selected codes that existed in both the Clinical and Medline datasets (both the Medline 100 and Medline 600 datasets) so that we could compare them across all sets. These codes are listed in Table B1.

For each code we identified all associations from each dataset to which it belonged and plotted the chi square statistic for the associations, so that the distributions of the chi square statistic could be compared across datasets. In each scatter plot Figure below (Figure B1-B15), the distributions of chi square statistics for associations containing the ICD-9 code of interest are shown for each of the six datasets from our analysis (see Table B2 for a description of the rows representing each dataset in the figures). The upper limit for the chi square statistic shown on the x-axis varies from 80,000 (Figures B1-B3) to 10,000 (Figures B10-B15). These were chosen to provide a reasonable range in which the spread of the Chi square statistics for the associations could be seen and compared. In some cases the values surpassed the upper range in the plot and were not displayed. Therefore, these figures are not comprehensive of all associations from each dataset but rather serve as a way to reasonably compare one dataset to another. In the clinical dataset a Chi square statistic of 1,475 roughly corresponded to a p value of 1.0 x 10^{-324} and a Chi square statistic of 454 corresponded to a p value of approximately 1.0 x 10^{-100}.

From these figures it can be seen that the clinical dataset has more scatter of the associations compared to the Medline datasets (based on the Chi Square statistic) and ICD-9 codes that occur more frequently in the Clinical dataset have more scatter compared to when those codes appear in the Medline datasets.

Table B1. ICD-9 codes used for the scatter plots. The percentile rank listed here is based on the number of patients with the ICD-9 code from the Clinical (original) dataset. Different codes that have the same number of patients will be tied for the same rank.

ICD-9 code	Code Description	Rank	Percentile Rank	Figure	Patients with code
427.9	unspecific cardiac dysrhythmia	3	1	B1	148,237
789.00	abdominal pain, unspecified site	6	1	B2	120,539
729.5	pain in limb	10	1	B3	110,363
783.6	polyphagia	2,440	25	B4	452
601.0	acute prostatitis	2,468	25	B5	431
523.8	other specified periodontal diseases	3,601	25	B6	422
114.9	Coccidioidomycosis, unspecified	2,834	50	B7	54
375.01	acute dacryoadenitis	2,835	50	B8	53
615.1	chronic inflammatory diseases of uterus (except cervix)	2,837	50	B9	51
388.41	diplacusis	2,879	75	B10	9
676.40	failure of lactation	2,879	75	B11	9
982.2	toxic effect of carbon disulfide	2,879	75	B12	9
E849.2	mine and quarry accidents	2,887	100	B13	1
902.27	injury to inferior mesenteric artery	2,887	100	B14	1
837	dislocation of ankle	2,887	100	B15	1

Table B2. Dataset used for each row in the scatter plots. Note that for the clinical dataset, codes with less than 30 patients were not considered in the association analysis and thus are not displayed in the scatter plots.

Scatter plot row	Dataset
f	Medline 600 (Simplified)
e	Medline 600 (Original)
d	Medline 1000 (Simplified)
c	Medline 1000 (Original)
b	Clinical (Simplified)
a	Clinical (Original)

Figure B1

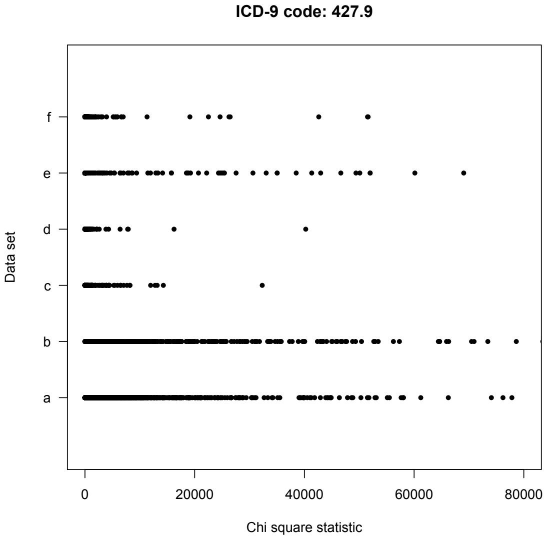


Figure B2

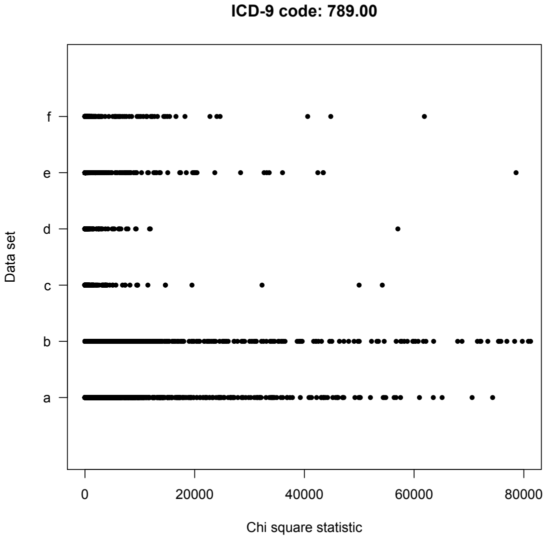


Figure B3

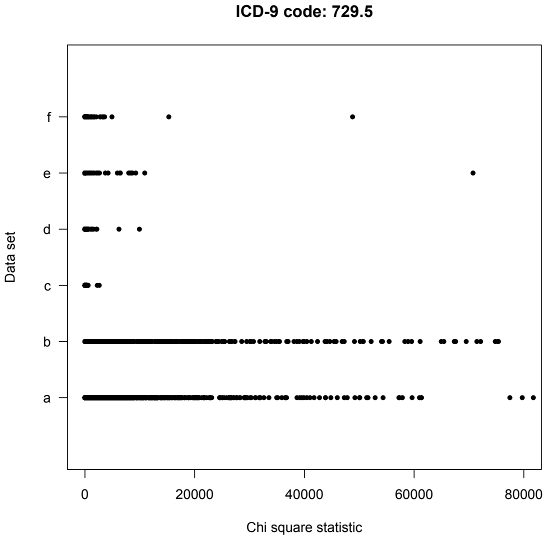


Figure B4

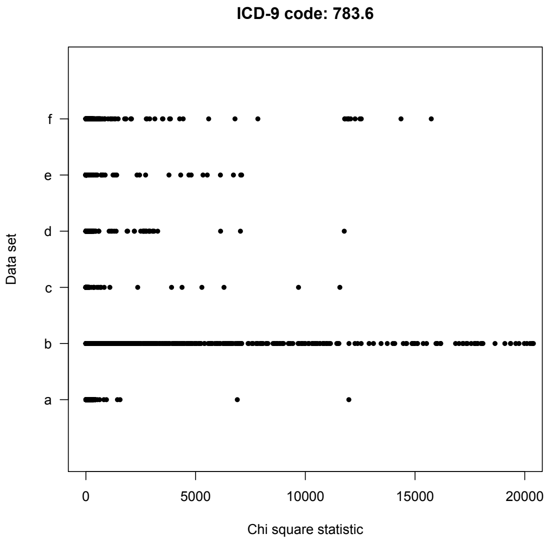


Figure B5

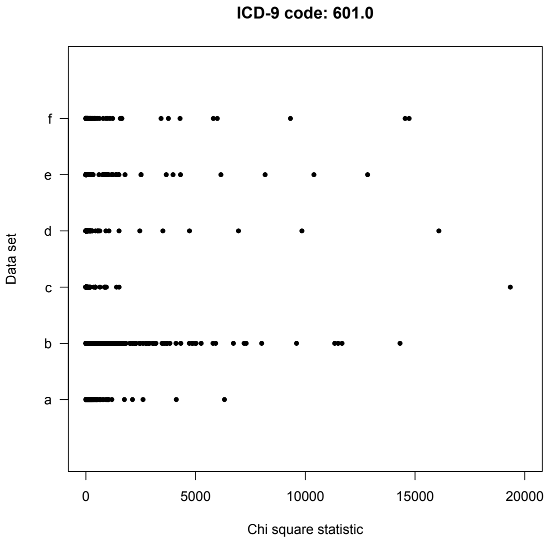


Figure B6

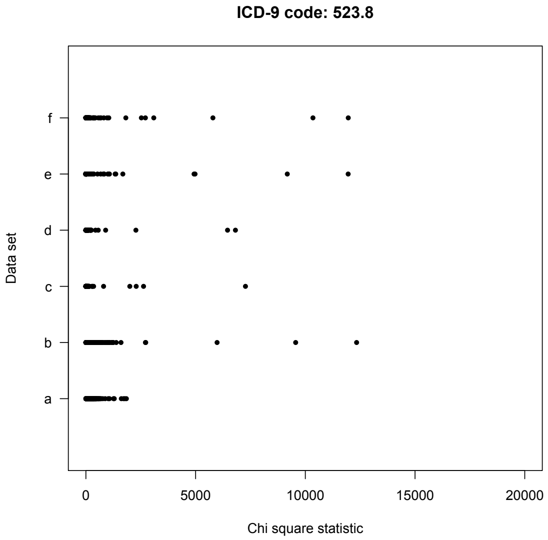


Figure B7

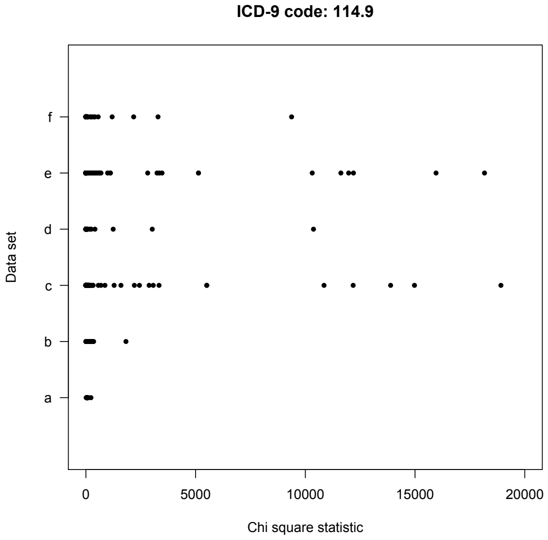


Figure B8

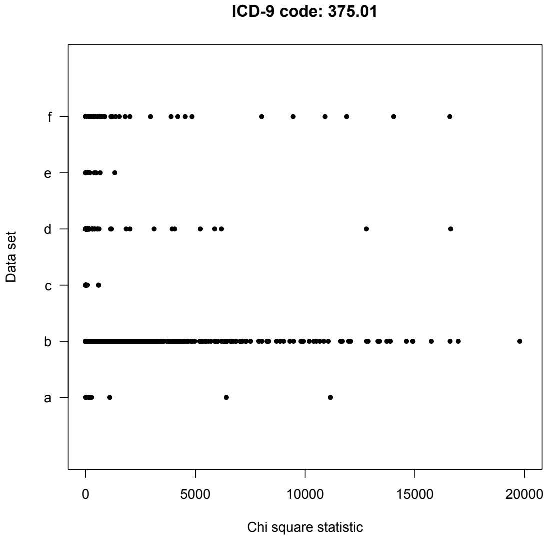


Figure B9

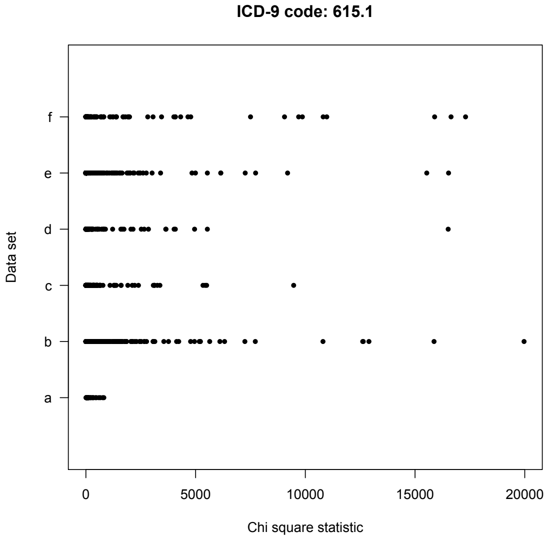


Figure B10

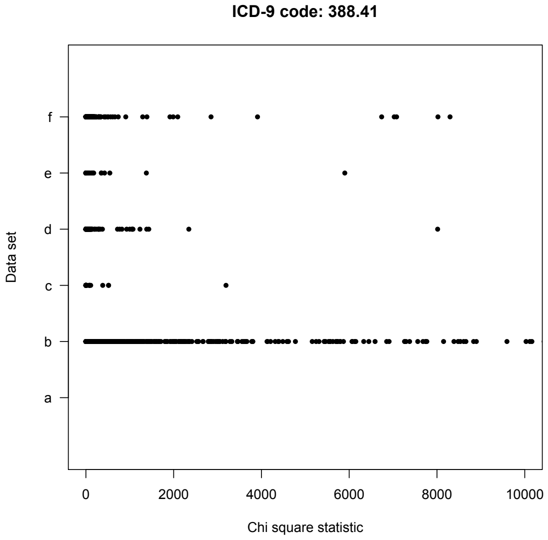


Figure B11

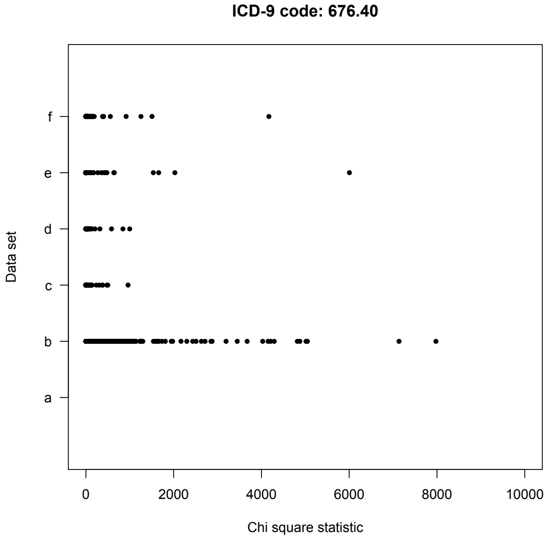


Figure B12

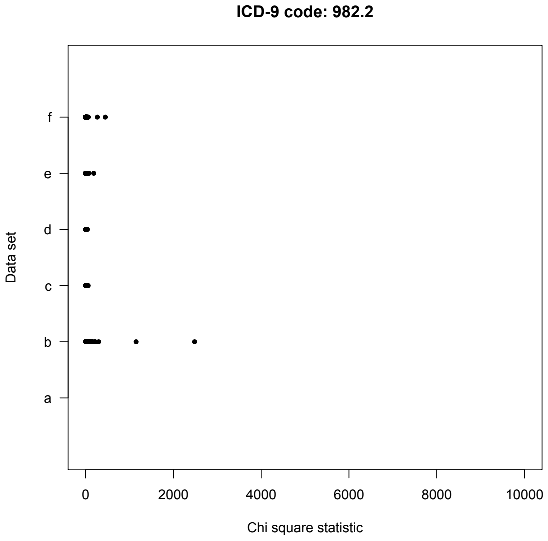


Figure B13

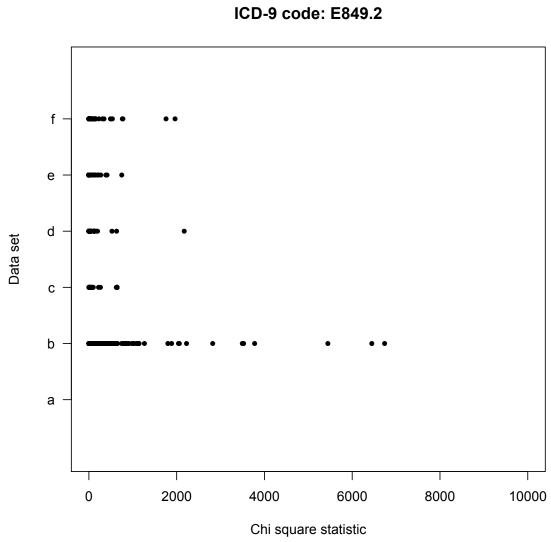


Figure B14

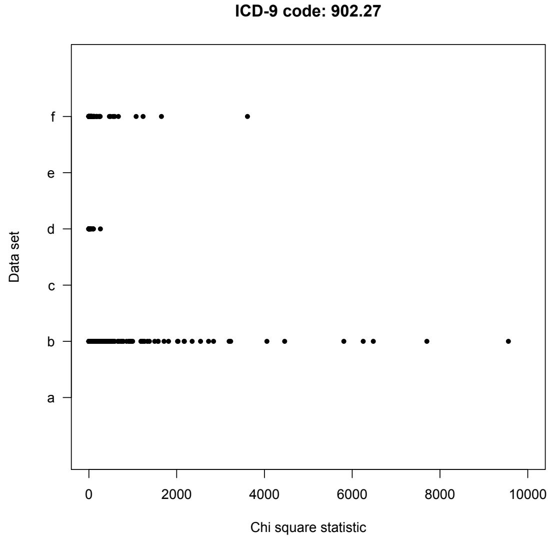
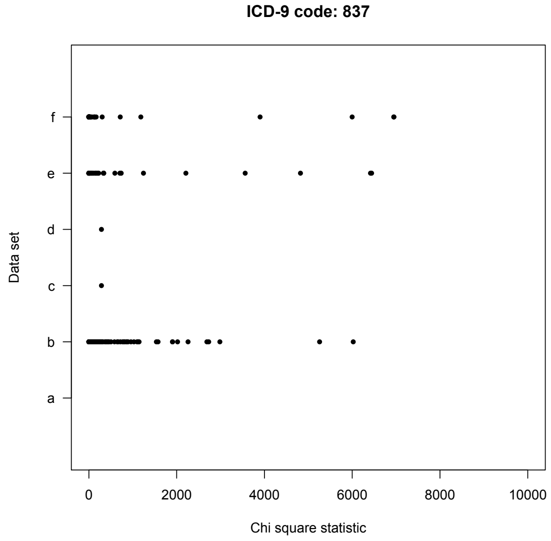


Figure B15


ICD-9 code: 837
Data set
f
e
d
c
b
a
Chi square statistic