Appendix C

In this section, we explored the associations contained in the UMLS relationships files to see how they related to those found with our other approaches. These relationships can be found in the file MRREL.RRF, which we obtained from the 202AA version of UMLS. This file contains 62.2 million rows and is 5.53 GB in size.

Each row contains two CUIs that are connected via a relationship (e.g., SIB: “has sibling relationship in a Metathesaurus source vocabulary”, and RO: “has relationship other than synonymous, narrower, or broader”). The relationships can also be defined by attributes, of which there are about 675, including “disease may have associated disease” and “clinically associated with”.

Further details about the relationships can be found here:

https://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/release/abbreviations.html

From this file we created two datasets. The first dataset included all rows in the MRREL.RRF file regardless of the relationship or relationship attribute. This we called the “UMLS complete” dataset. The second dataset was created by including only those rows that had one of 25 relationship attributes that we selected as being potentially relevant for describing an association between two clinical concepts as we might find in our clinical dataset. This was called the “UMLS subset” dataset and the included relationship attributes are shown in Table C1. This dataset included 383,094 rows.

For both datasets we extracted the CUI pairs and removed duplicates, leaving the “UMLS complete” dataset with 26,837,582 rows and the “UMLS subset” dataset with 241,195 rows. The CUIs in each dataset were then mapped to ICD-9 codes using the same approach described in the main body of the manuscript, and only the relationships where both elements in a pair contained an ICD-9 code were retained. After this mapping the Complete dataset contained 22,189 distinct ICD-9 codes and the Subset dataset contained 4,869 distinct ICD9 codes.

Association analyses were then run on both of these datasets using the same approach described in the manuscript. The Venn diagrams below show the overlap of the ICD-9 codes (Figures C1-C12) and associations (Figures C13-C24) found in this analysis.

From these figures and what was presented in the main body of the manuscript, it is evident that there is substantial variability in the overlap of the codes from each dataset, which further depends on the initial parameters for creating the datasets (e.g., UMLS-subset vs. UMLS-complete, and Medline 1000 vs. Medline 600). The ICD-9 codes obtained from the MRREL.RRF file generally have more coverage than the other datasets, although this is substantially restricted when we limited the relationship attributes to the 25 listed in Table C1. Further, despite the UMLS relationship dataset having more ICD-9 codes than the clinical dataset, the clinical dataset still contained many more associations than were found in the UMLS dataset.

Table C1. Relationship attributes and their respective frequencies, obtained from the UMLS file MRREL.RRF. These were used to construct the “UMLS subset” dataset.

relationship attribute	row count
alternative_of	181
associated_disease	2,309
associated_etiologic_finding_of	1,404
associated_finding_of	22,554
associated_with	160,462
cause_of	7,169
clinically_associated_with	29,632
clinically_similar	9,270
co-occurs_with	13,768
disease_has_associated_disease	814
disease_has_finding	19,141
disease_may_have_associated_disease	1,479
has_associated_etiologic_finding	1,404
has_associated_finding	22,554
has_associated_procedure	9,236
has_sign_or_symptom	377
is_associated_disease_of	814
is_finding_of_disease	19,141
may_be_associated_disease_of_disease	1,479
may_be_finding_of_disease	12,960
occurs_after	10,008
occurs_before	10,008
related_to	11,818
temporally_followed_by	7,556
temporally_follows	7,556
TOTAL	383,094

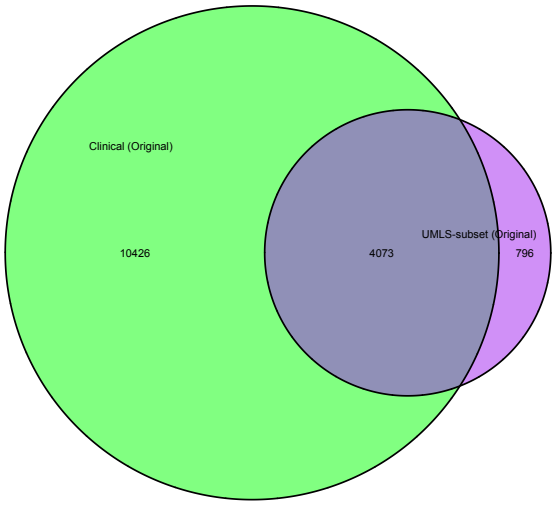

Figure C1. ICD-9 code overlap between Clinical (Original) and UMLS subset (Original) datasets

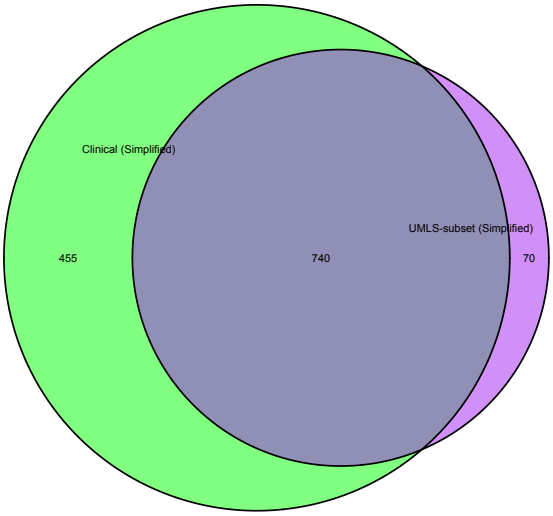

Figure C2. ICD-9 code overlap between Clinical (Simplified) and UMLS subset (Simplified) datasets

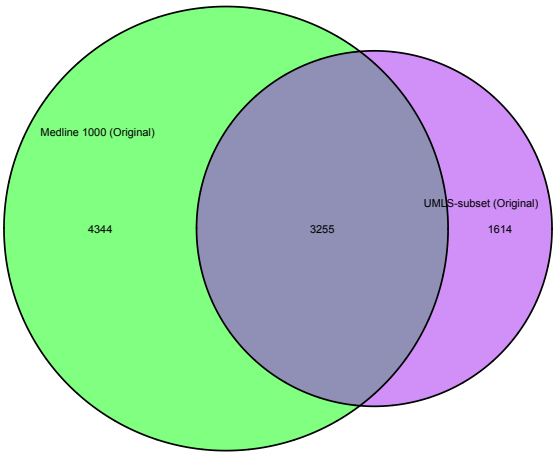

Figure C3. ICD-9 code overlap between the Medline 1000 (Original) and UMLS subset (Original) datasets

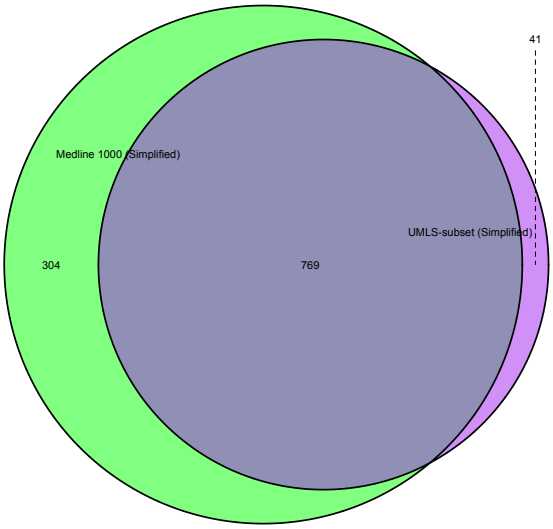

Figure C4. ICD-9 code overlap between the Medline 1000 (Simplified) and UMLS subset (Simplified) datasets

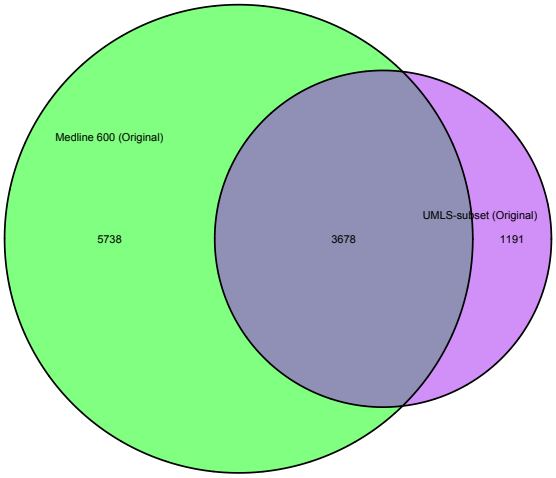

Figure C5. ICD-9 code overlap between the Medline 600 (Original) and UMLS subset (Original) datasets

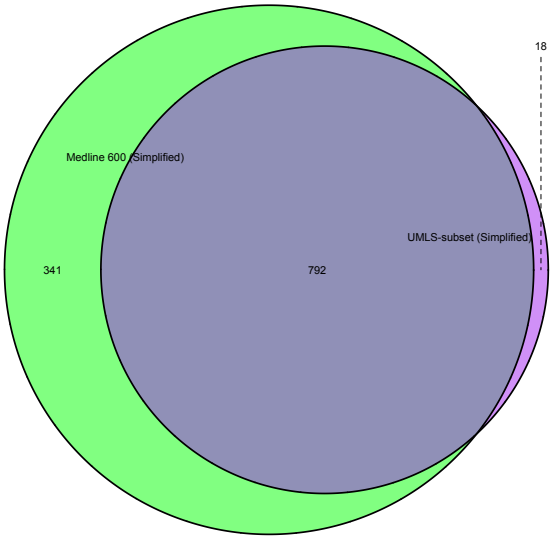

Figure C6. ICD-9 code overlap between the Medline 600 (Simplified) and UMLS subset (Simplified) datasets

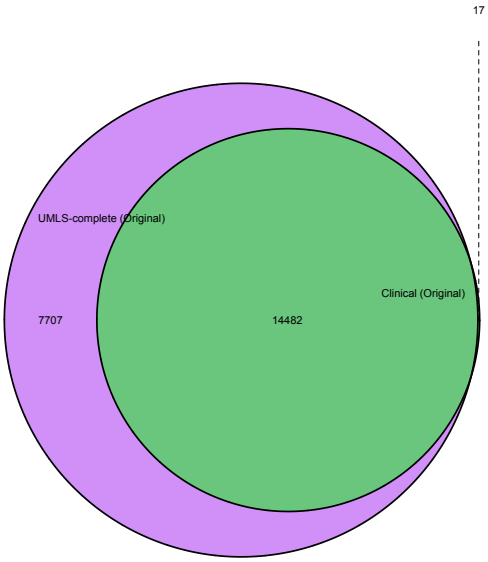

Figure C7. ICD-9 code overlap between the Clinical (Original) and UMLS complete (Original) datasets

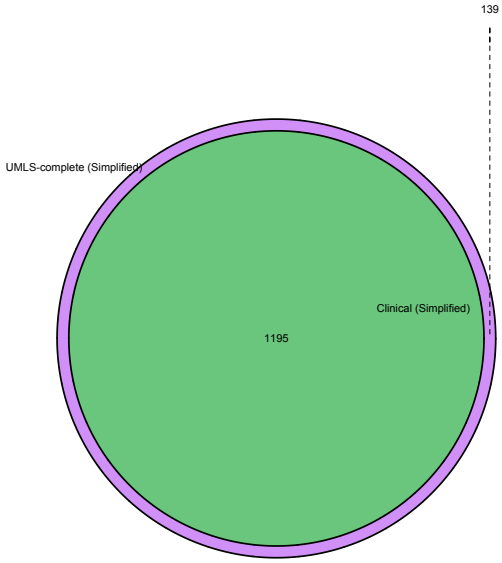

Figure C8. ICD-9 code overlap between the Clinical (Simplified) and UMLS complete (Simplified) datasets

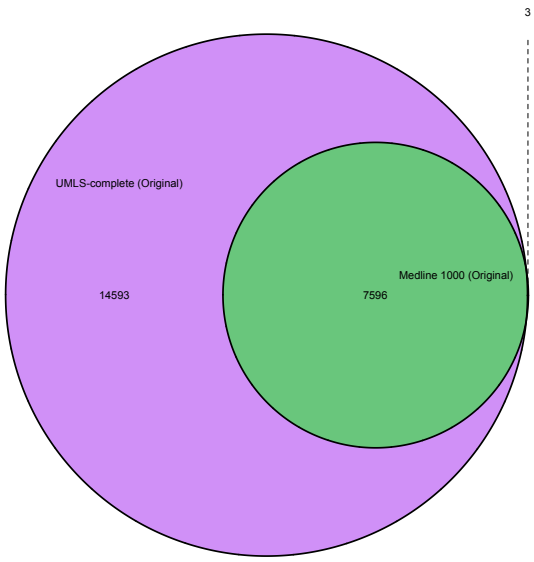

Figure C9. ICD-9 code overlap between the Medline 1000 (Original) and UMLS complete (Original) datasets

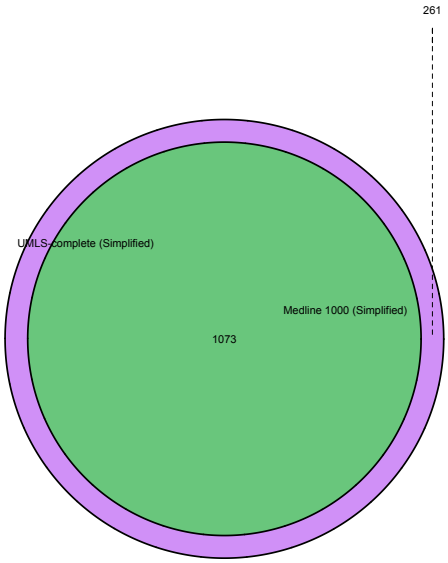

Figure C10. ICD-9 code overlap between the Medline 1000 (Simplified) and UMLS complete (Simplified) datasets

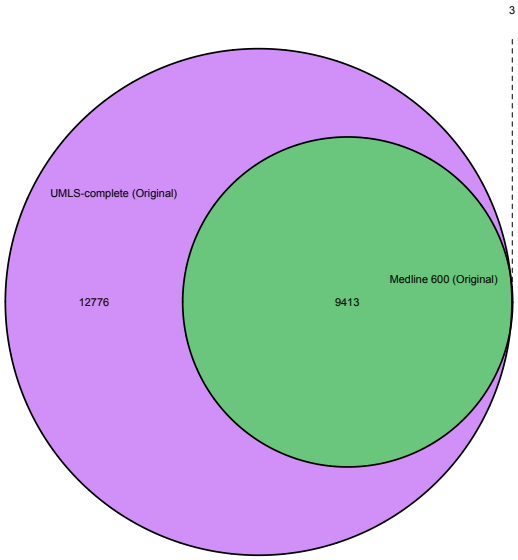

Figure C11. ICD-9 code overlap between the Medline 600 (Original) and UMLS complete (Original) datasets

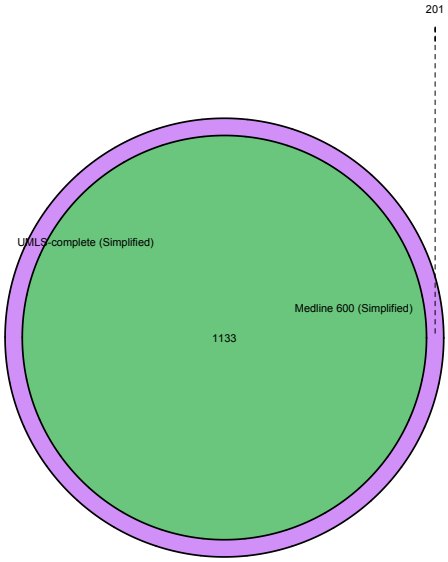

Figure C12. ICD-9 code overlap between the Medline 600 (Simplified) and UMLS complete (Simplified) datasets

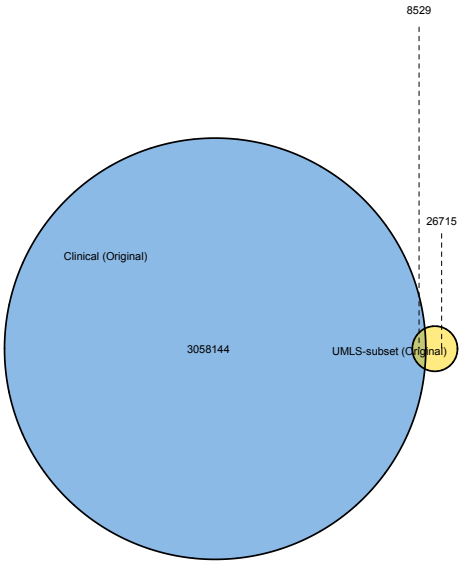

Figure C13. Association overlap between the Clinical (Original) and UMLS subset (Original) datasets

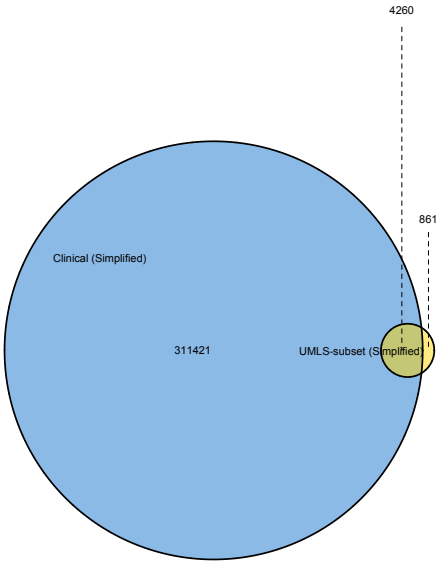

Figure C14. Association overlap between the Clinical (Simplified) and UMLS subset (Simplified) datasets

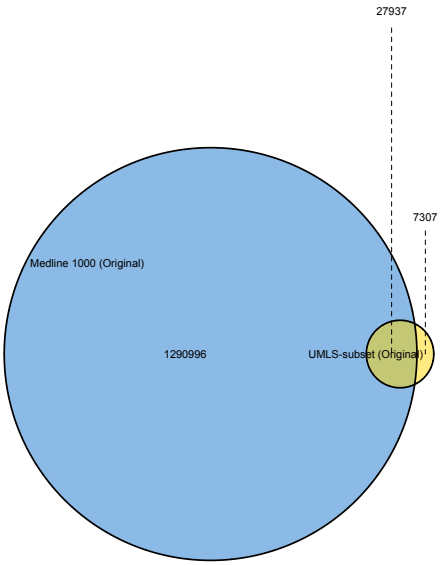

Figure C15. Association overlap between the Medline 1000 (Original) and UMLS subset (Original) datasets

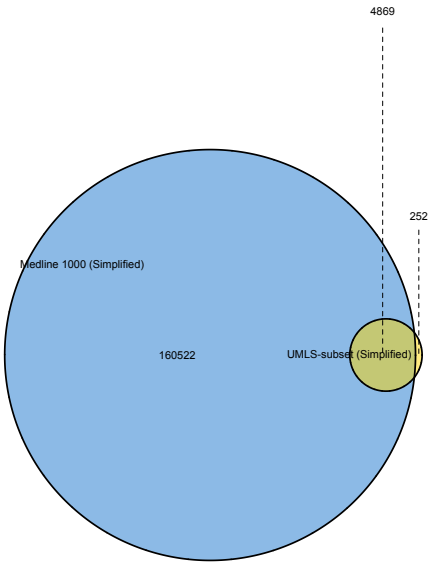

Figure C16. Association overlap between the Medline 1000 (Simplified) and UMLS subset (Simplified) datasets

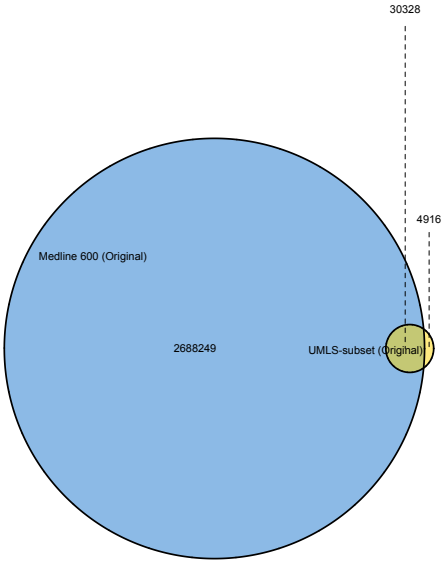

Figure C17. Association overlap between the Medline 600 (Original) and UMLS subset (Original) datasets

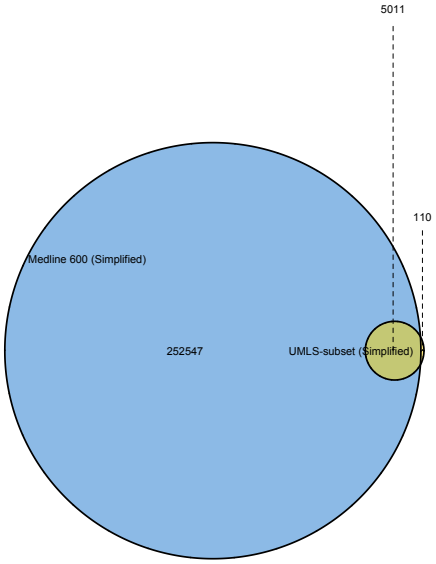

Figure C18. Association overlap between the Medline 600 (Simplified) and UMLS subset (Simplified) datasets

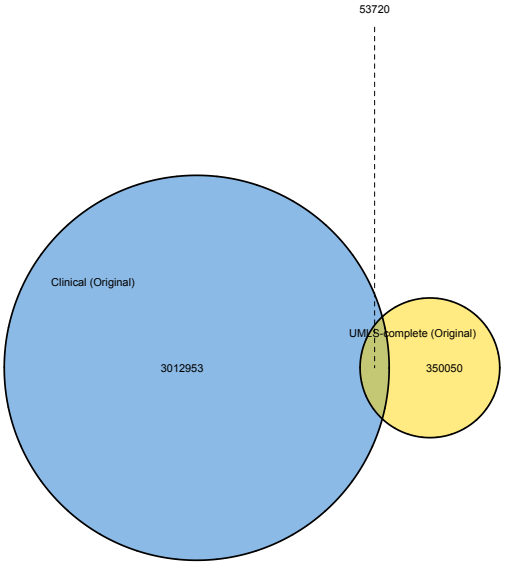

Figure C19. Association overlap between the Clinical (Original) and UMLS complete (Original) datasets

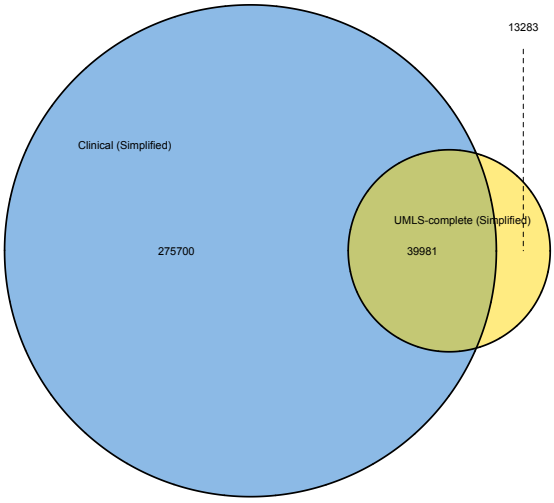

Figure C20. Association overlap between the Clinical (Simplified) and UMLS complete (Simplified) datasets

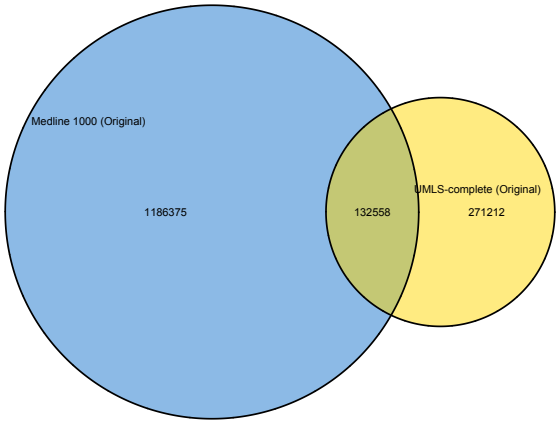

Figure C21. Association overlap between the Medline 1000 (Original) and UMLS complete (Original) datasets

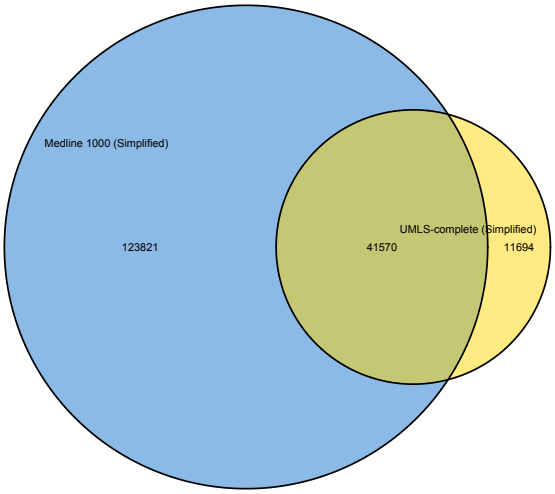

Figure C22. Association overlap between the Medline 1000 (Simplified) and UMLS complete (Simplified) datasets

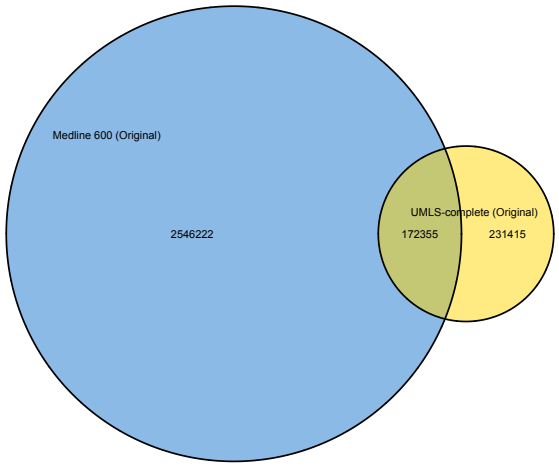

Figure C23. Association overlap between the Medline 600 (Original) and UMLS complete (Original) datasets

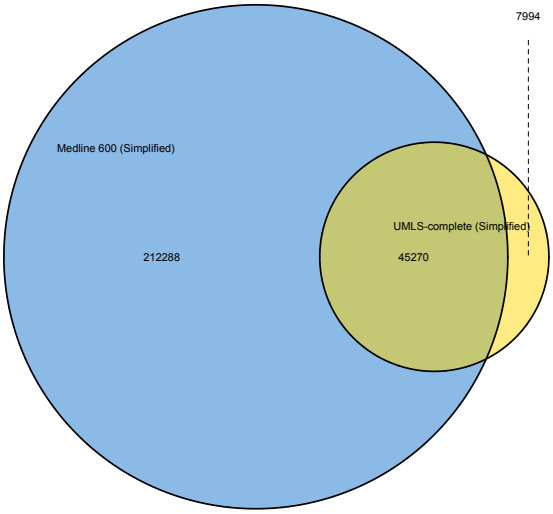

Figure C24. Association overlap between the Medline 600 (Simplified) and UMLS complete (Simplified) datasets