Published in final edited form as:
J Biomed Inform. 2010 June ; 43(3): 442–450. doi:10.1016/j.jbi.2009.10.003.

MESHING MOLECULAR SEQUENCES AND CLINICAL TRIALS: A FEASIBILITY STUDY

Elizabeth S. Chen[1,2,4] and **Indra Neil Sarkar**[1,3,4,*]

[1] Center for Clinical and Translational Science, University of Vermont, Burlington, VT USA

[2] Division of General Internal Medicine, Department of Medicine, University of Vermont, Burlington, VT USA

[3] Department of Microbiology and Molecular Genetics, University of Vermont, Burlington, VT USA

[4] Department of Computer Science, University of Vermont, Burlington, VT USA

Abstract

The centralized and public availability of molecular sequence and clinical trial data presents an opportunity to identify potentially valuable linkages across the bench-to-bedside "T1" translational barrier. In this study, we sought to leverage keyword metadata (Medical Subject Heading [MeSH] descriptors) to infer relationships between molecular sequences and clinical trials, as indexed by GenBank and ClinicalTrials.gov. The results of this feasibility study found that approximately 30% of sequences in GenBank could be linked to trials and over 90% of trials in ClinicalTrials.gov could be linked to sequences through MeSH descriptors. In a cursory evaluation, we were able to consistently identify meaningful linkages between molecular sequences and clinical trials. Based on our findings, there may be promise in subsequent studies aiming to identify linkages across the T1 translational barrier using existing large repositories.

Keywords

GenBank; ClinicalTrials.gov; PubMed/MEDLINE; MeSH; bench-to-bedside; translational bioinformatics; metadata analysis

1. INTRODUCTION

Core to the success of translational bioinformatics endeavors will be ability to link relevant information across the bench-to-bedside translational barrier ("T1"). Insights across this barrier inherently involve the linking of bench-based research data with relevant clinical information. To date, many studies that aim to bridge across the T1 translational barrier have focused on the study of *de novo* data. To this end, there have been limited explorations into the development of a semantic infrastructure for linking clinical trial data to relevant molecular data.

Clinical hypotheses can often involve the direct manipulation of genetic material (such as in animal models) or the development of targeted interventions that interact with a particular

[*] To whom correspondence should be addressed: Indra Neil Sarkar, PhD, MLIS, Center for Clinical and Translational Science, 89 Beaumont Avenue, Given Courtyard N309, Burlington, VT 05405 USA, neil.sarkar@uvm.edu.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

molecule. As clinical hypotheses are tested, putative successful interventions are incorporated into clinical trials. As publications describe the background and results of a clinical trial, they are often catalogued in centralized resources. For example, this may be especially important when considering the identification of molecules or molecular properties that may be relevant in the context of a particular combination of disease phenotypes. As interventions are assessed for their clinical efficacy and effectiveness, it may be imperative to identify linkages to underlying molecular mechanisms. Drawing correlations between bench and bedside research collectively comprise the “holy grail” for translational bioinformatics, yet this can be often hindered by limited availability of data sets that can be cross-linked. Previous work has explored the ability and challenges to navigating information resources (e.g., MedlinePlus, ClinicalTrials.gov, OMIM, and PubMed/MEDLINE) from phenotype to genotype to answer specific questions about genes and gene products related to specific diseases (e.g., “What genes cause the disease?” and “Are there gene therapies or clinical trials for this disease”)[1]. Major challenges to linking or integration of these resources included data complexity, dynamic data, diverse foci and number of resources, and lack of standardized data and knowledge representation. Numerous efforts have emerged to address these data and knowledge integration issues in different contexts (e.g., translational research[2] and biology[3]).

Both the molecular and clinical trial communities have embraced the use of centralized repositories[4]. Within the molecular biology community, almost all reference molecular sequence data are publicly available via GenBank (presently containing over 100 million records)[5]; ClinicalTrials.gov is a publicly accessible resource that catalogues clinical trials (presently containing over 75 thousand records)[6,7]. Both resources are maintained by the National Library of Medicine (NLM) (the former by the National Center for Biotechnology Information [NCBI] and the latter by the Lister Hill National Center for Biomedical Communications in collaboration with the Food and Drug Administration) and readily downloadable.

Many of the NLM/NCBI resources are cross-linked through the Entrez interface and are collectively searchable from a single interface or through the Entrez programming utilities (E-Utilities)[8,9]. For example, it is possible to identify both articles in PubMed/MEDLINE and molecular sequences in GenBank that are associated with a given topic. The strong relationship between GenBank and publications that are largely indexed in MEDLINE[10] can lead to inferred annotations of molecular sequence data. Recent work has shown how these Entrez-based relationships can lead to approaches to link molecular information to relevant literature (e.g., identify keyword descriptors that are associated with publications that are affiliated with a group of related molecular sequences)[11]. Within ClinicalTrials.gov, keyword descriptors are explicitly applied to entries. Additional keyword descriptors can also be inferred, as in the case with GenBank entries, through associated publications.

Keyword descriptors that are part of controlled vocabulary or ontological structures can be leveraged for linking putatively related data elements. Preliminary studies demonstrate the possibility of leveraging existing semantic infrastructures to link genotypic information (e.g., as captured in molecular databases) to phenotypic (e.g., as captured in stores of clinical data). For example, previous work has demonstrated how one might make use of the Unified Medical Language System (UMLS) to link clinical concepts primarily represented in the Systematized Nomenclature of Medicine (SNOMED) to genomic concepts represented in the Gene Ontology (GO)[12,13]. Additional work has explored how the controlled vocabulary primarily associated with MEDLINE, the Medical Subject Headings (MeSH)[14], can be used to link disparate resources (e.g., OMIM, UMLS Metathesaurus, and GenBank[15]) or identify possibly linked concepts[16]. In contrast to specific domain terminologies, such as SNOMED or GO, MeSH is designed primarily as an indexing terminology to capture the breadth of biomedical science that is required to meet information needs associated with the PubMed/

MEDLINE interface. Thus, while lacking potential deeper levels of domain knowledge (as leveraged by other semantic mediated linkage systems[17,18]), MeSH provides a broad list of concepts that may provide some traction for linking relevant objects to each other across the translational spectrum. For example, through a series of MeSH based queries to PubMed/MEDLINE, it is possible to identify relevant citations that are related to each other across the entire spectrum of translational bioinformatics.

Building on the premise that MeSH-based annotations can be used to link related concepts, this study explores the feasibility of linking molecular sequence and clinical trial data leveraging MeSH descriptors either directly associated with clinical trials or inferred from published literature (i.e., PubMed/MEDLINE).

2. MATERIALS AND METHODS

For this feasibility study, the overall approach involved three major phases: (1) collecting PubMed Identifiers (PMIDs) and MeSH descriptors from GenBank, ClinicalTrials.gov, and PubMED/MEDLINE; (2) charactering and filtering by MeSH descriptors; and, (3) identifying linkages between molecular sequences and clinical trials based on PMIDs and MeSH.

2.1 Data Collection

Literature references and MeSH descriptors associated with sequences in GenBank and studies in ClinicalTrials.gov were collected. Figure 1 depicts the approach for extracting this information from GenBank, ClinicalTrials.gov, and PubMed/MEDLINE. Each step in the overall process is further described in the following sections.

2.1.1 Molecular Sequences from GenBank—Molecular sequence data represent a fundamental component in many bench research endeavors. GenBank is a centralized repository for cataloguing molecular sequence data[5]. Through partnerships via the International Nucleotide Sequence Database Collaboration (INSDC), GenBank data are globally synchronized with European (EMBL) and Asian (DDBJ) repositories, thus representing a complete collection of nearly all publicly available molecular sequence data. Associated with each sequence entry in GenBank is a detailed set of metadata elements, organized into general bibliographic metadata and “Feature Table” as defined by the INSDC. Included among the over 70 metadata elements are fields like “ACCESSION” representing the unique identifier for the sequence and “REFERENCE” for literature relevant to the sequence that includes “AUTHOR,” “TITLE,” “JOURNAL,” and “PUBMED” conveying the citation and PMID. In a recent study, a significant portion (~30%) of GenBank records was found to be associated with PubMed/MEDLINE records through the PMIDs[11]. Preliminarily analyses have demonstrated the ability to leverage PMIDs to identify MeSH descriptors that are associated with a given GenBank record.

The entirety of GenBank is freely downloadable from NCBI[19]. A series of scripts was developed to extract and load GenBank metadata into a MySQL database thus enabling rapid query of GenBank records (e.g., “Which sequences have reference information?”). For sequences with references, E-Utilities was used to obtain the associated MeSH descriptors using the PMIDs (Figure 1.A). This set of references and MeSH descriptors derived from information in GenBank was supplemented by information from PubMed/MEDLINE. Specifically, E-Utilities was used to query PubMed/MEDLINE directly to identify records containing information about molecular sequence data through the “Secondary Source ID” (SI) field (e.g., SI–GENBANK/AF306859)[20]; MeSH descriptors were then obtained for the corresponding PMIDs (Figure 1.B). The combined sets of PMIDs and MeSH descriptors will henceforth be referred to as *GB/PMID* and *GB/P-MeSH* respectively.

2.1.2 Clinical Trials from ClinicalTrials.gov—Clinical trials represent a primary means for evaluating the safety and efficacy of new therapies and other interventions that have the potential for improving clinical practice. ClincialTrials.gov was developed in response to legislation mandating a comprehensive, publically accessible registry of federally and privately funded clinical trials[6,7]. This Web-based registry maintains a common set of data elements for each trial including 40 major required and optional elements for descriptive information, recruitment information, location and contact information, administrative data, and optional supplementary information. Information about studies registered in ClinicalTrials.gov is viewable from the public Web site and can be downloaded as plain text or XML files from this site[21]. To facilitate registration of trial information from organizations, a Web-based data entry and management system called the Protocol Registration System was created[22,23].

Required elements associated with each registered study include study identifier, brief title, recruitment status, sponsor or funding source, eligibility criteria, study design, condition(s), and intervention(s). Optional elements may include investigators, references for background citations, references for completed studies, and keywords. The use of MeSH for conditions, interventions, and keywords is requested if possible; however, part of the data preparation phase of the registration process involves mapping to MeSH descriptors as needed and high-level MeSH categories[6,7].

Based on a preliminary review of content within the XML files and Web pages for ClinicalTrials.gov, we observed that a combination of sources and methods would be needed to extract a more complete set of literature references and MeSH descriptors associated with each study. Table 1 itemizes the relevant elements, how they are represented in the ClinicalTrials.gov XML files, what headers they appear under in the ClinicalTrials.gov Web pages, and sources used to collect this information. The specific sources are: (1) XML files from ClinicalTrials.gov for the full studies (Figure 1.1), (2) HTML pages from the ClinicalTrials.gov public Web site (Figure 1.2), and (3) PubMed/MEDLINE (Figure 1.3).

Using the "Download Options" feature in ClinicalTrials.gov, the XML files for all full studies were downloaded. Each XML file was parsed according to the Document Type Definition (DTD)[24] to extract basic metadata (e.g., National Clinical Trials Identifiers [NCT ID] and title) and metadata associated with references. These references either represent literature that provide background for the study ("Background References") or report on results from the study ("Results References"). For either type of reference, the PMID, full citation, or both may be provided. In cases where a PMID was available, E-Utilities was used to retrieve associated MeSH descriptors. Similar to our GenBank analysis, E-Utilities was used to query PubMed/MEDLINE to identify any additional references for clinical trials; these include references displayed on the ClinicalTrials.gov Web pages that are not in the corresponding XML file and are difficult to extract from the Web pages or any others that may be indicated by the "SI" field in PubMed/MEDLINE (e.g., SI–ClinicalTrials.gov/NCT00000419). MeSH descriptors associated with these records were similarly obtained for the corresponding PMIDs. The combined sets of PMIDs and MeSH descriptors will henceforth be referred to as *CT/PMID* and *CT/P-MeSH* respectively.

Finally, in examining the content displayed on the ClinicalTrials.gov public Web site for studies, we found that some information was not included in the XML files such as keywords, topic categories, and MeSH descriptors. To add these MeSH descriptors into the data set, we processed each Web page associated with a study to extract this information. This set of MeSH descriptors is referred to as *CT/C-MeSH* to represent MeSH descriptors obtained directly from ClinicalTrials.gov rather than through PubMed/MEDLINE; the set resulting from the combination of this set with CT/P-MeSH will be called *CT/CP-MeSH*.

2.2 Characterizing and Filtering by MeSH

An analysis of the four sets of MeSH descriptors collected for sequences in GenBank and studies in ClinicalTrials.gov (GB/P-MeSH, CT/P-MeSH, CT/C-MeSH, and CT/CP-MeSH) was performed to determine the distribution of descriptors across the 16 top-level MeSH 2009 categories. This analysis revealed that the hierarchies with the highest frequency based on the combined set of MeSH descriptors for clinical trials (CT/CP-MeSH) were "Diseases" (Category C) and "Chemicals and Drugs" (Category D). These two categories were used to filter each of the MeSH descriptor sets in order to use more focused and condensed sets to test the feasibility of linking molecular sequences and clinical trials through MeSH. In addition, due to the large size of GenBank, GB/P-MeSH was further filtered to include only the MeSH descriptors also occurring in the sets for clinical trials.

2.3 Linking Molecular Sequences and Clinical Trials

Based on the PMID and MeSH descriptor sets, two major types of linkages can be obtained: (1) links between sequences and trials through PMID and (2) links between sequences and trials through MeSH. The former can enable the ability the answer questions like: "For a given sequence, what clinical trials may be of interest, what literature is associated with both the sequence and trials, and what other literature associated with the trials may be of interest?" or "For a given trial, what sequences may be of interest, what literature is associated with both the sequences and trial, and what other literature associated with the sequences may be of interest?". The latter may enhance the answers provided by the PMIDs alone by identifying additional linkages through common MeSH descriptors.

Using the GB/PMID and CT/PMID sets, links between sequences and trials were identified based on common PMIDs. Three sets of linkages were identified between the MeSH descriptor set for sequences (GB/P-MeSH) and MeSH descriptor sets for trials (CT/C-MeSH, CT/P-MeSH, and CT/CP-MeSH). Due to the potentially large number of and possibly irrelevant links, a score was calculated to measure the relevancy of each link based on the number of common MeSH descriptors between a given sequence and trial. In the case of links from trials to sequences, the general algorithm was as follows: (1) get all associated MeSH descriptors for the trial, (2) identify sequences with at least two MeSH descriptors in common with the trial, and (3) determine the strength of the link by calculating the proportion of MeSH descriptors in common. For example, if Trial A = {MeSH1, MeSH2, MeSH3, MeSH4, MeSH5, MeSH6}, Sequence B = {MeSH1, MeSH4}, and Sequence C = {MeSH1, MeSH2, MeSH3, MeSH5}, then Score(Trial A, Sequence B) = 0.33 and Score(Trial A, Sequence C) = 0.67.

3. RESULTS

3.1 Molecular Sequences and Clinical Trials

GenBank (Release 172) was downloaded, processed, and loaded into a MySQL database for further querying in July 2009. Out of the over one hundred million sequences, 30.13% are associated with references from PubMed/MEDLINE, specifically 43,540,315 total PMIDs and 298,646 unique PMIDs. An E-Utilities query to PubMed/MEDLINE performed on July 9, 2009 identified 99,230 additional unique articles with 1,493,439 associated Accession IDs representing sequences in GenBank. The combination of these sets included 35,128,847 (30.36%) sequences with 313,612 unique PMIDs (*GB/PMID* set)

Full studies for all clinical trials were downloaded from ClinicalTrials.gov as separate XML files on July 6, 2009. This set included 74,853 trials from 167 countries. Each XML file was processed to extract basic information and references associated with the study. Out of the total number of trials, 14,320 (19.13%) included 70,307 background and/or results references where 91.98% of these references are associated with 64,669 total PMIDs and 56,167 unique PMIDs.

An E-Utilities query to PubMed/MEDLINE performed on July 8, 2009 identified 3,790 articles with 3,397 associated NCT IDs representing studies in ClinicalTrials.gov. Combining these two sets resulted in 15,674 unique trials (20.94%) with 59,234 unique PMIDs (*CT/PMID* set). Table 2 depicts the breakdown of references by type (background or results) and source (ClinicalTrials.gov or PubMed/MEDLINE).

Using E-Utilities to query PubMed/MEDLINE, MeSH descriptors were obtained for each PMID in GB/PMID and CT/PMID to produce *GB/P-MeSH* and *CT/P-MeSH*, respectively. In addition, MeSH descriptors for each trial were extracted from the corresponding ClinicalTrials.gov Web pages to generate *CT/C-MeSH*. The CT/P-MeSH and CT/C-MeSH sets were then combined to create *CT/CP-MeSH*. Table 3 summarizes the number of unique sequences or trials, total number of MeSH descriptors, and total number of unique MeSH descriptors associated with each of these sets.

3.2 MeSH Hierarchy Characterization

The top-level hierarchy for each MeSH descriptor in GB/P-MeSH, CT/P-MeSH, CT/C-MeSH, and CT/CP-MeSH was identified. Table 4 and Figure 2 depict the distribution of descriptors across the sixteen hierarchies. The top three hierarchies were “Phenomena and Processes” (37.23%), “Organisms” (15.43%), and “Chemicals and Drugs” (12.63%) for GB/P-MeSH; “Diseases” (53.80%), “Chemicals and Drugs” (41.15%), and “Psychiatry and Psychology” (3.40%) for CT/C-MeSH; “Analytical, Diagnostic and Therapeutic Techniques and Equipment” (20.02%), “Technology, Industry, Agriculture” (20.2%), and “Chemical and Drugs” (13.77%) for CT/P-MeSH; and, “Diseases” (31.48%), “Chemicals and Drugs” (26.74%), “Analytical, Diagnostic and Therapeutic Techniques and Equipment” (10.75%), and “Technology, Industry, Agriculture” (10.75%) for CT/CP-MeSH. The sets were subsequently filtered by “Diseases” and “Chemicals and Drugs” (the top two hierarchies based on CT/C-MeSH and CT/CP-MeSH) and GB/P-MeSH was further filtered to include descriptors only occurring in the clinical trials sets.

3.3 Molecular Sequence and Clinical Trial Linkages

3.3.1 Linking Molecular Sequences and Clinical Trials through PMID—The PMIDs from GB/PMID and CT/PMID were used to identify links between molecular sequences and clinical trials. A total number of 78,259 links were identified involving 64,181 sequences and 426 trials. For a given sequence, the minimum number of links to trials was 1 and the maximum was 6; for a given trial, the number of links to sequences ranged from a minimum of 1 to a maximum of 41,152. Figure 3 depicts the sequences and literature references linked to the trial “Family Studies of Inherited Heart Disease” (NCT ID = NCT00001225) and Figure 4 lists the trials and literature references linked to the sequence “*Homo sapiens* Huntington’s Disease (HD) mRNA, complete cds.” (Accession ID = HUMHDA).

3.3.2 Linking Molecular Sequences and Clinical Trials through MeSH—Through MeSH descriptors in the “Diseases” and “Chemicals and Drugs” hierarchies, three sets of linkages between sequences and trials were identified (Table 5). Over 39 billion links were generated from GB/P-MeSH and CT/C-MeSH covering 90.41% of the trials, over 800 million were identified for CT/P-MeSH involving 18.72% of the total trials, and the combined set of CT/CP-MeSH resulted in almost 40 million links for 91.6% of the trials. The estimated upper bound for the number of sequences involved in each of these sets was about 35 million (30.00%).

For each link, the relevancy score was calculated based on the number of common MeSH descriptors to enable the ranking of sequences linked to a given trial. Figure 5 plots the distribution of sequences relative to trials based on the relevancy scores. CT/C-MeSH by itself was able to identify relevant sequences across all clinical trials. By contrast, CT/P-MeSH by itself does not offer as many significant sequences that are related to clinical trials. The combined set of MeSH descriptors (CT/CP-MeSH) was not dramatically affected by the CT/P-MeSH; however, there were a number of instances where even if no CT/C-MeSH descriptors were available that CT/P-MeSH was able to facilitate the identification of relevant molecular sequences. As an example, Figure 6 presents MeSH descriptors and the MeSH-based links to sequences for the same trial in Figure 3.

4. DISCUSSION

The development of approaches for identifying potential linkages across disparate data sources is essential towards the generation of putative testable hypotheses. Within the scope of translational bioinformatics, such hypotheses include those that aim to identify potentially related molecular sequences and clinical trials, thus transcending the “T1” bench-to-bedside translational barrier. In the present study, we were able to demonstrate the ability to link molecular sequences from GenBank to relevant clinical trials in ClinicalTrials.gov and vice versa. Using MeSH descriptors associated with GenBank records (through PubMed/MEDLINE) and MeSH descriptors associated with ClinicalTrials.gov records (either directly, through PubMed/MEDLINE, or in combination), linkable molecular sequences and clinical trials can be identified.

The linkage between human-centric (e.g., ClinicalTrials.gov) and organism-agnostic (e.g., GenBank) resources presents an opportunity to identify putative model organisms for a given disease. We are currently working to develop an approach that leverages the results of the present study to discover potentially interesting organisms that may be associated with a clinical condition. In this feasibility study, the entirety of GenBank was considered to determine the feasibility of developing potentially meaningful linkages from ClinicalTrials.gov; however, the nature of sequences in GenBank is such that there are often numerous similar or even identical sequences. As a possible way to address this, we are planning to leverage curated molecular sequence resources such as RefSeq[25] or UniGene [26]. Additionally, we anticipate that the use of RefSeq or UniGene will help lead to a more statistically relevant understanding of the linkages between clinical trials and available molecular sequence data (by filtering out redundant sequences).

In this study, literature references associated with either GenBank or ClinicalTrials.gov were used as a surrogate for annotations. There is a keyword field in GenBank; however, it is not associated with a specific controlled vocabulary like MeSH. References in ClinicalTrials.gov are categorized as either “background” or “result” citations. In the present study, we treated these two references equivocally; however, it may be interesting to study the effect of using either of these category types exclusively compared with their combination. During our processing of ClinicalTrials.gov data, we observed that there were references that did not have a PMID (thus preventing an easy linkage to PubMed/MEDLINE). Future work could therefore involve reviewing these to determine why there is no PMID (e.g., abstract/poster or not indexed in PubMed/MEDLINE).

Because our approach to supplement the MeSH descriptors for both molecular sequences and clinical trials depended on PMIDs, a complete list of associated PMIDs was required for each sequence or trial. However, an interesting finding was that it was not possible to reliably get a complete list directly from either GenBank or ClinicalTrials.gov. For GenBank, previous work has characterized the types of discordance between which PMIDs are reported from GenBank compared with those that are in PubMed/MEDLINE[11]. Performing a similar type of analysis on ClinicalTrials.gov data reveals a similar trend (Figure 7). This further underscores the challenges in curating, maintaining, and synchronizing large repositories like GenBank, ClinicalTrials.gov, and PubMed/MEDLINE. Nonetheless, the infrastructure is in place for explicitly linking between key resources (e.g., PubMed/MEDLINE to GenBank or PubMed/MEDLINE to ClinicalTrials.gov). Interestingly, a PubMed/MEDLINE search for records that are associated with both GenBank and ClinicalTrials.gov records (as recorded in the "Secondary Identifier" [SI] field) only retrieves two articles (PMID 16525138 and 16525137; query performed on August 31, 2009).

The present study focused on keyword metadata in the form of MeSH descriptors. Both GenBank and ClinicalTrials.gov have associated metadata fields that contain additional potentially valuable keyword terms that might be used to develop more reliable and complete linkages. For example, examination of the GenBank keyword field reveals that there are over 2 million terms (compared to the over 25,000 MeSH descriptors available). A possible future area of work may thus be to map these keywords to a hierarchy such as MeSH to get a more direct annotation of molecular sequence data as a comparison to those inferred by PubMed/MEDLINE. With respect to ClinicalTrials.gov, other potential keywords (many of which might be directly derived from MeSH) include keywords provided by the sponsors or organizations (e.g., "Keywords provided by National Institutes of Health Clinical Center (CC)"), topic categories, conditions, and interventions. We have begun an analysis of these terms and have started identifying mechanisms to map them to MeSH descriptors (e.g., using natural language processing techniques). Finally, in considering PubMed/MEDLINE, we only focused on the use of MeSH descriptors. Next steps include making better use of MeSH qualifiers and substances to help identify additional linkages as well as help quantify the importance of linkages that are identified using solely MeSH.

One part of this study involved characterizing MeSH descriptors by identifying their corresponding hierarchies or categories in MeSH 2009. This characterization provided insights on the distribution of descriptors across MeSH hierarchies for molecular sequences compared with clinical trials and assisted in identifying an initial filtering strategy (i.e., by "Diseases" and "Chemicals and Drugs"). It is worth noting that in some cases a descriptor could not be found (e.g., may be from an earlier version of MeSH) and some descriptors fall into multiple hierarchies. Future work includes characterizing descriptors by semantic type or MeSH sub-hierarchies and exploring additional filtering strategies (e.g., including other hierarchies such as "Psychiatry and Psychology").

The initial approach of using PMIDs and MeSH descriptors to identifying potential relationships between sequences and trials revealed a vast number of linkages. Further evaluation of these linkages will be valuable for assessing their relevance and guiding the development of techniques for enhancing the ranking of results.

5. CONCLUSION

The ability to integrate disparate publicly available biomedical resources could be valuable for supporting and promoting research across the bench-to-bedside ("T1") translational barrier. In this study, we explored the feasibility of linking molecular sequences in GenBank and clinical trials in ClinicalTrials.gov leveraging literature from PubMed/MEDLINE and keyword metadata in the form of MeSH descriptors. The results obtained in this feasibility study indicate that this is a promising approach for identifying relevant linkages between sequences and trials.

Acknowledgments

The authors thank colleagues from the University of Vermont Center for Clinical and Translational Science for insightful discussions regarding this study. This work is supported in part by National Library of Medicine Grant R01LM009725.

References

1. Mitchell JA, McCray AT, Bodenreider O. From phenotype to genotype: issues in navigating the available information resources. Methods Inf Med 2003;42(5):557–63. [PubMed: 14654891]
2. Burgun A, Bodenreider O. Accessing and integrating data and knowledge for biomedical research. Yearb Med Inform 2008:91–101. [PubMed: 18660883]
3. Donelson L, Tarczy-Hornoch P, Mork P, Dolan C, Mitchell JA, Barrier M, et al. The BioMediator system as a data integration tool to answer diverse biologic queries. Stud Health Technol Inform 2004;107(Pt 2):768–72. [PubMed: 15360916]
4. Lindberg D, Humphreys B. Rising expectations: access to biomedical information. Yearb Med Inform 2008:165–72. [PubMed: 18587496]
5. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. GenBank. Nucleic Acids Res 2009 Jan;37(Database issue):D26–31. [PubMed: 18940867]
6. McCray A. Better access to information about clinical trials. Ann Intern Med 2000 Oct;133(8):609–14. [PubMed: 11033590]
7. McCray A, Ide N. Design and implementation of a national clinical trials registry. J Am Med Inform Assoc 7(3):313–23. [PubMed: 10833169]
8. Sayers EW, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, et al. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 2009 Jan;37(Database issue):D5–15. [PubMed: 18940862]
9. Baxevanis AD. Searching NCBI databases using Entrez. Curr Protoc Bioinformatics 2008 Dec;Chapter 1(Unit 1–3)
10. Lindberg DA, Schoolman HM. The National Library of Medicine and medical informatics. West J Med 1986 Dec;145(6):786–90. [PubMed: 3544508]
11. Miller H, Norton C, Sarkar I. GenBank and PubMed: How connected are they? BMC Res Notes 2009;2:101. [PubMed: 19508734]
12. Cantor M, Sarkar I, Bodenreider O, Lussier Y. Genestrace: phenomic knowledge discovery via structured terminology. Pac Symp Biocomput 2005:103–14. [PubMed: 15759618]
13. Sarkar I, Cantor M, Gelman R, Hartel F, Lussier Y. Linking biomedical language information and knowledge resources: GO and UMLS. Pac Symp Biocomput 2003:439–50. [PubMed: 12603048]
14. Sewell W. Medical Subject Headings in Medlars. Bull Med Libr Assoc 1964 Jan;52:164–70. [PubMed: 14119288]
15. Sperzel WD, Abarbanel RM, Nelson SJ, Erlbaum MS, Sherertz DD, Tuttle MS, et al. Biomedical database inter-connectivity: an experiment linking MIM, GENBANK, and META-1 via MEDLINE. Proc Annu Symp Comput Appl Med Care 1991:190–3. [PubMed: 1807585]
16. Srinivasan P, Rindflesch T. Exploring text mining from MEDLINE. Proc AMIA Symp 2002:722–6. [PubMed: 12463919]
17. Muller HM, Kenny EE, Sternberg PW. Textpresso: an ontology-based information retrieval and extraction system for biological literature. PLoS Biol 2004 Nov;2(11):e309. [PubMed: 15383839]
18. Lussier Y, Borlawsky T, Rappaport D, Liu Y, Friedman C. PhenoGO: assigning phenotypic context to gene ontology annotations with natural language processing. Pac Symp Biocomput 2006:64–75. [PubMed: 17094228]
19. ftp://ftp.ncbi.nih.gov/genbank/
20. http://www.nlm.nih.gov/bsd/mms/medlineelements.html
21. Madden TL, Tatusov RL, Zhang J. Applications of network BLAST server. Methods Enzymol 1996;266:131–41. [PubMed: 8743682]
22. Gillen J, Tse T, Ide N, McCray A. Design, implementation and management of a web-based data entry system for ClinicalTrials.gov. Stud Health Technol Inform 2004;107(Pt 2):1466–70. [PubMed: 15361058]
23. http://prsinfo.clinicaltrials.gov/definitions.html
24. http://www.clinicaltrials.gov/ct2/html/images/info/public.dtd
25. http://www.ncbi.nlm.nih.gov/RefSeq/
26. http://www.ncbi.nlm.nih.gov/unigene

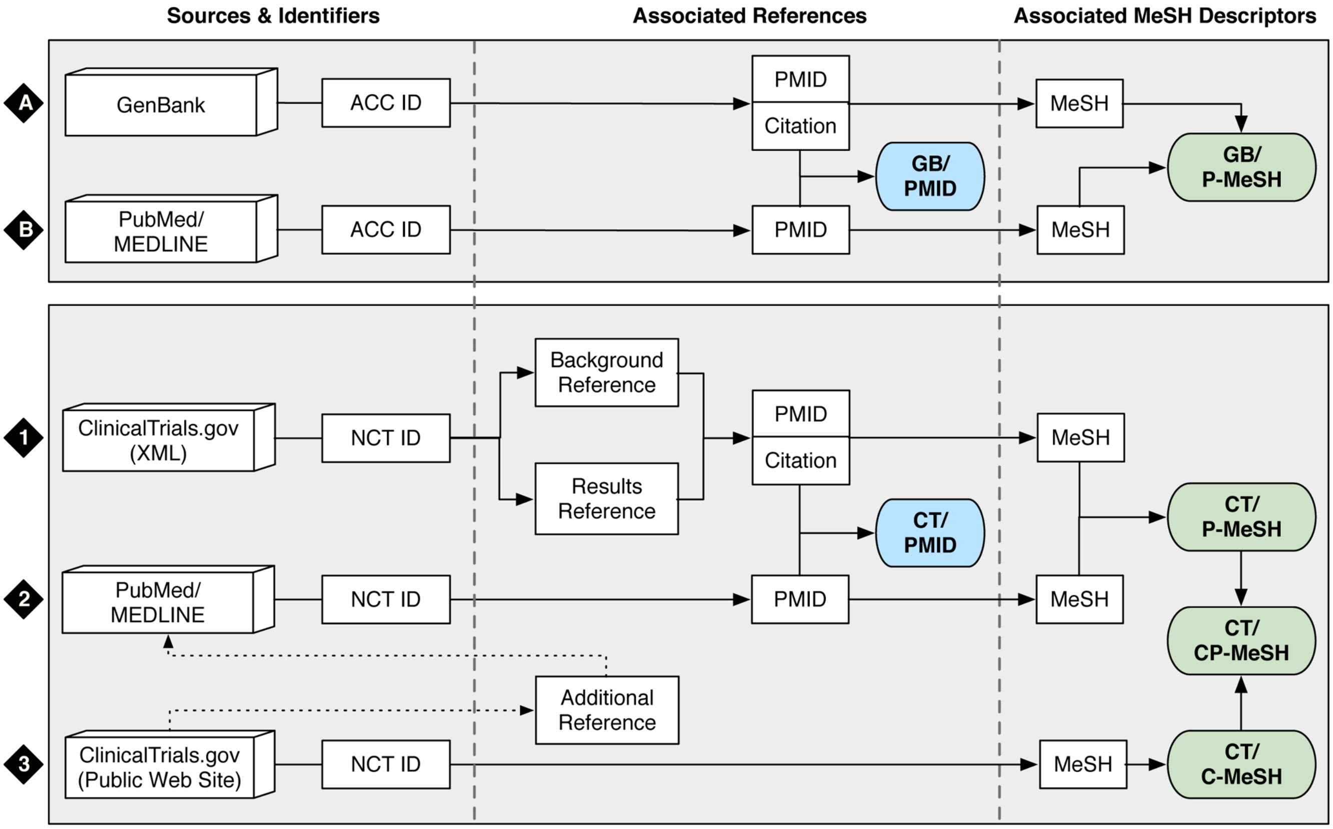


Figure 1. Collecting Literature References and MeSH Descriptors for Molecular Sequences and Clinical Trials

References (citation and/or PubMed Identifier [PMID]) associated with sequences (indicated by a unique Accession Identifier [ACC ID]) in GenBank were obtained (A). Records (indicated by a unique PMID) in PubMed/MEDLINE that include Accession IDs for sequences in GenBank were also identified (B). The combined set of PMIDs from GenBank and PubMed/MEDLINE is “GB/PMID” and the set of MeSH descriptors for these PMIDs is “GB/P-MeSH”. Background and results references (citation and/or PubMed Identifier [PMID]) were obtained from XML files for studies (indicated by a National Clinical Trials Identifier [NCT ID]) in ClinicalTrials.gov (1). Records (indicated by a unique PMID) in PubMed/MEDLINE that include NCT IDs for studies in ClinicalTrials.gov were also identified; these include additional references listed on the CLinicalTrials.gov public Web site (2). The combined set of PMIDs from GenBank and PubMed/MEDLINE is “CT/PMID” and the set of MeSH descriptors for these PMIDs is “CT/P-MeSH”. The set of MeSH descriptors obtained directly for studies from the ClinicalTrials.gov Web pages is “CT/C-MeSH” (3). The combination of MeSH descriptors (explicit and obtained through PMIDs) for ClinicalTrials.gov is “CT/CP-MeSH”.

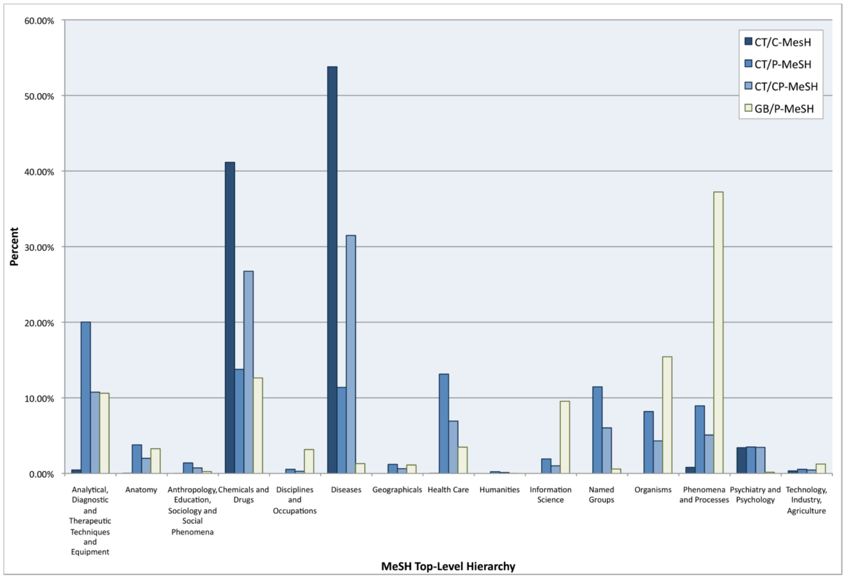
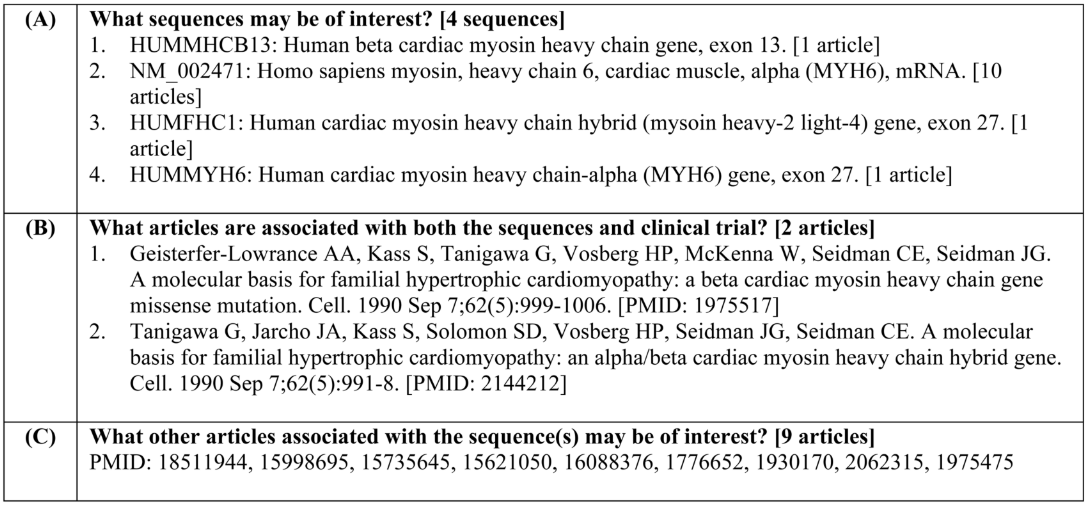


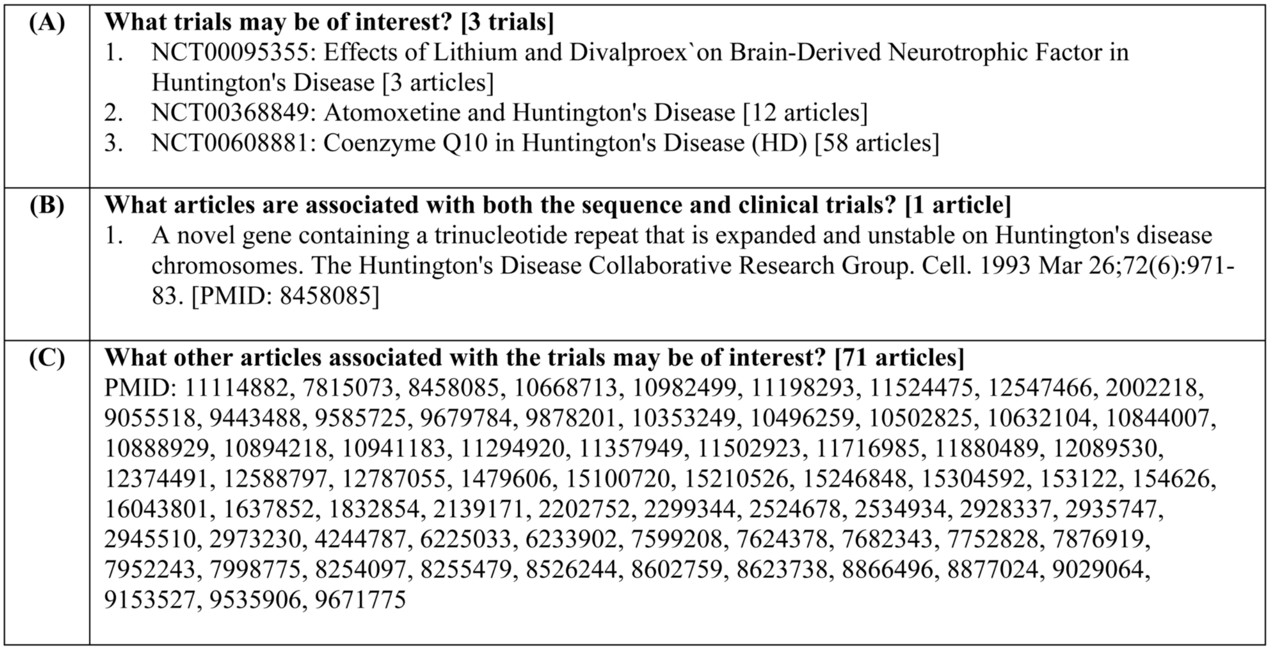
Figure 2.
Distribution of MeSH Descriptors Across the Top-Level Hierarchies for Sequences and Trials.



(A)	**What sequences may be of interest? [4 sequences]** 1. HUMMHCB13: Human beta cardiac myosin heavy chain gene, exon 13. [1 article] 2. NM_002471: Homo sapiens myosin, heavy chain 6, cardiac muscle, alpha (MYH6), mRNA. [10 articles] 3. HUMFHC1: Human cardiac myosin heavy chain hybrid (mysoin heavy-2 light-4) gene, exon 27. [1 article] 4. HUMMYH6: Human cardiac myosin heavy chain-alpha (MYH6) gene, exon 27. [1 article]
(B)	**What articles are associated with both the sequences and clinical trial? [2 articles]** 1. Geisterfer-Lowrance AA, Kass S, Tanigawa G, Vosberg HP, McKenna W, Seidman CE, Seidman JG. A molecular basis for familial hypertrophic cardiomyopathy: a beta cardiac myosin heavy chain gene missense mutation. Cell. 1990 Sep 7;62(5):999-1006. [PMID: 1975517] 2. Tanigawa G, Jarcho JA, Kass S, Solomon SD, Vosberg HP, Seidman JG, Seidman CE. A molecular basis for familial hypertrophic cardiomyopathy: an alpha/beta cardiac myosin heavy chain hybrid gene. Cell. 1990 Sep 7;62(5):991-8. [PMID: 2144212]
(C)	**What other articles associated with the sequence(s) may be of interest? [9 articles]** PMID: 18511944, 15998695, 15735645, 15621050, 16088376, 1776652, 1930170, 2062315, 1975475

Figure 3. PMID-based Links for Trial in ClinicalTrials.gov, NCT00001225 (Family Studies of Inherited Heart Disease)

This study is associated with three articles (PMID 1975517, 2022018, and 2144212). Links to four sequences were identified (A) where two articles were found to be common to the trial and sequences (B) and nine other articles associated with the sequences may be of interest (C).



(A)	**What trials may be of interest? [3 trials]** 1. NCT00095355: Effects of Lithium and Divalproex`on Brain-Derived Neurotrophic Factor in Huntington's Disease [3 articles] 2. NCT00368849: Atomoxetine and Huntington's Disease [12 articles] 3. NCT00608881: Coenzyme Q10 in Huntington's Disease (HD) [58 articles]
(B)	**What articles are associated with both the sequence and clinical trials? [1 article]** 1. A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. The Huntington's Disease Collaborative Research Group. Cell. 1993 Mar 26;72(6):971-83. [PMID: 8458085]
(C)	**What other articles associated with the trials may be of interest? [71 articles]** PMID: 11114882, 7815073, 8458085, 10668713, 10982499, 11198293, 11524475, 12547466, 2002218, 9055518, 9443488, 9585725, 9679784, 9878201, 10353249, 10496259, 10502825, 10632104, 10844007, 10888929, 10894218, 10941183, 11294920, 11357949, 11502923, 11716985, 11880489, 12089530, 12374491, 12588797, 12787055, 1479606, 15100720, 15210526, 15246848, 15304592, 153122, 154626, 16043801, 1637852, 1832854, 2139171, 2202752, 2299344, 2524678, 2534934, 2928337, 2935747, 2945510, 2973230, 4244787, 6225033, 6233902, 7599208, 7624378, 7682343, 7752828, 7876919, 7952243, 7998775, 8254097, 8255479, 8526244, 8602759, 8623738, 8866496, 8877024, 9029064, 9153527, 9535906, 9671775

Figure 4. PMID-based Links for Sequence in GenBank, HUMHDA (*Homo sapiens* Huntington's Disease (HD) mRNA, complete cds. ())

This sequence is associated with one article (PMID: 8458085). Links to three trials were identified (A) where one article was found to be common to the sequence and trials (B) and seventy-one other articles associated with the trials may be of interest (C).

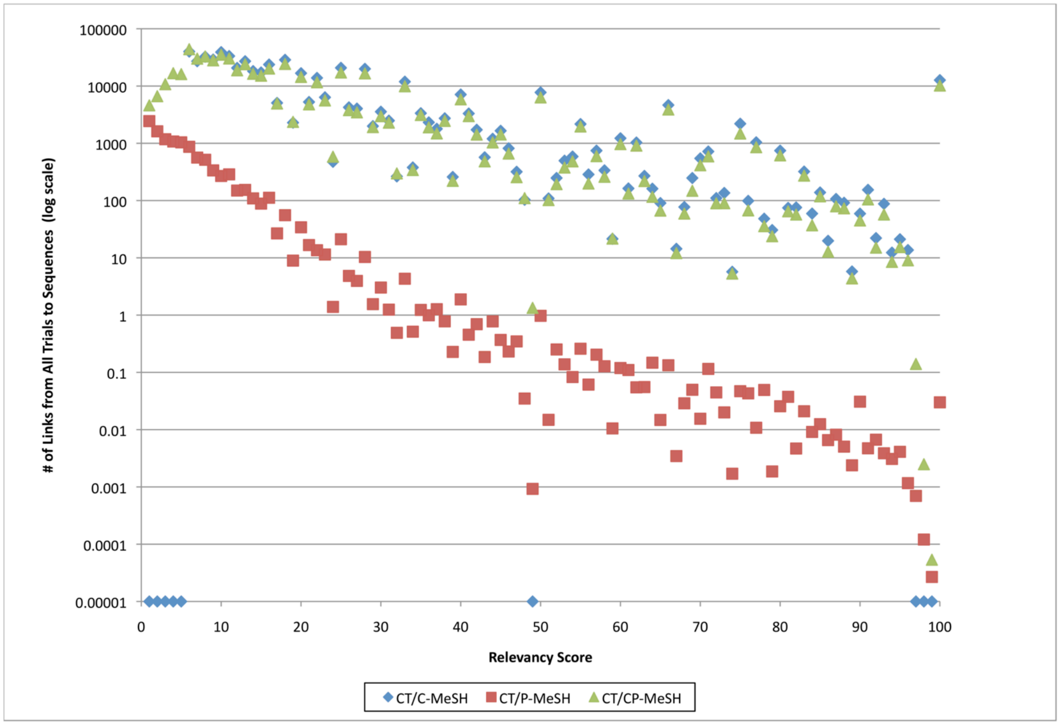
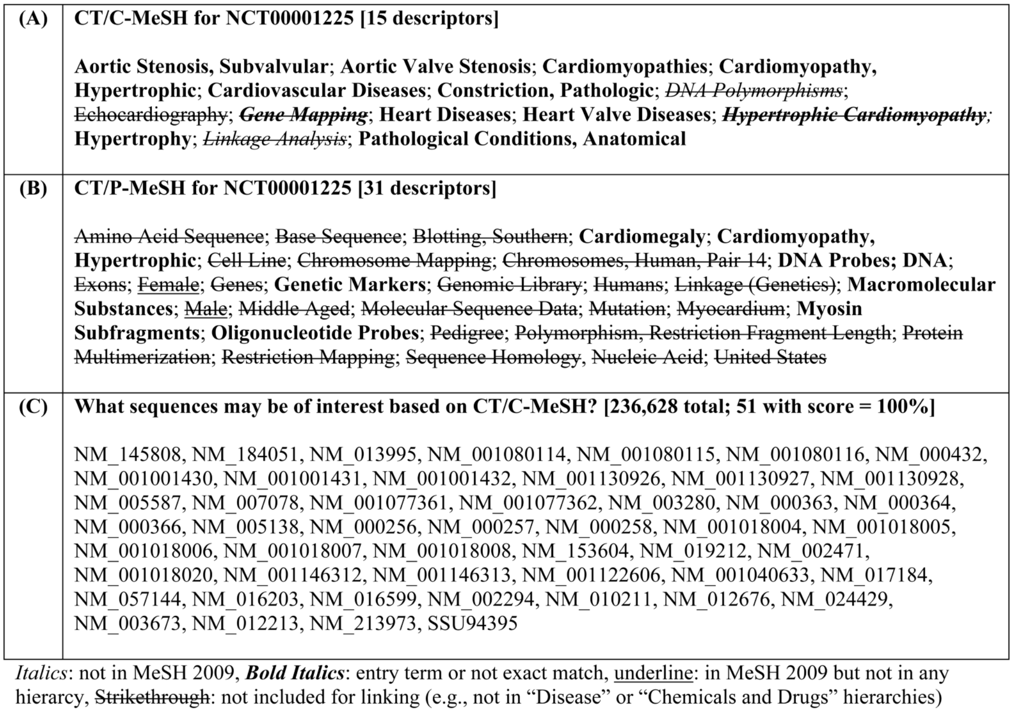


Figure 5. MeSH-based Relevancy of Molecular Sequences Linked to Clinical Trials

The number of molecular sequences that were linkable across all clinical trials (Y-Axis) is shown according to their relevancy (X-axis, on a scale of 0[not relevant]–100[highly relevant]).



(A)	**CT/C-MeSH for NCT00001225 [15 descriptors]** **Aortic Stenosis, Subvalvular**; **Aortic Valve Stenosis**; **Cardiomyopathies**; **Cardiomyopathy, Hypertrophic**; **Cardiovascular Diseases**; **Constriction, Pathologic**; ~~*DNA Polymorphisms*~~; ~~Echocardiography~~; ~~***Gene Mapping***~~; **Heart Diseases**; **Heart Valve Diseases**; ~~***Hypertrophic Cardiomyopathy***~~; **Hypertrophy**; ~~*Linkage Analysis*~~; **Pathological Conditions, Anatomical**
(B)	**CT/P-MeSH for NCT00001225 [31 descriptors]** ~~Amino Acid Sequence~~; ~~Base Sequence~~; ~~Blotting, Southern~~; **Cardiomegaly**; **Cardiomyopathy, Hypertrophic**; ~~Cell Line~~; ~~Chromosome Mapping~~; ~~Chromosomes, Human, Pair 14~~; **DNA Probes; DNA**; ~~Exons~~; ~~Female~~; ~~Genes~~; **Genetic Markers**; ~~Genomic Library~~; ~~Humans~~; ~~Linkage (Genetics)~~; **Macromolecular Substances**; ~~Male~~; ~~Middle Aged~~; ~~Molecular Sequence Data~~; ~~Mutation~~; ~~Myocardium~~; **Myosin Subfragments**; **Oligonucleotide Probes**; ~~Pedigree~~; ~~Polymorphism, Restriction Fragment Length~~; ~~Protein Multimerization~~; ~~Restriction Mapping~~; ~~Sequence Homology~~, ~~Nucleic Acid~~; ~~United States~~
(C)	**What sequences may be of interest based on CT/C-MeSH? [236,628 total; 51 with score = 100%]** NM_145808, NM_184051, NM_013995, NM_001080114, NM_001080115, NM_001080116, NM_000432, NM_001001430, NM_001001431, NM_001001432, NM_001130926, NM_001130927, NM_001130928, NM_005587, NM_007078, NM_001077361, NM_001077362, NM_003280, NM_000363, NM_000364, NM_000366, NM_005138, NM_000256, NM_000257, NM_000258, NM_001018004, NM_001018005, NM_001018006, NM_001018007, NM_001018008, NM_153604, NM_019212, NM_002471, NM_001018020, NM_001146312, NM_001146313, NM_001122606, NM_001040633, NM_017184, NM_057144, NM_016203, NM_016599, NM_002294, NM_010211, NM_012676, NM_024429, NM_003673, NM_012213, NM_213973, SSU94395

Italics: not in MeSH 2009, ***Bold Italics***: entry term or not exact match, underline: in MeSH 2009 but not in any hierarcy, ~~Strikethrough~~: not included for linking (e.g., not in "Disease" or "Chemicals and Drugs" hierarchies)

Figure 6. MeSH-based Links for Trial in ClinicalTrials.gov, NCT00001225 (Family Studies of Inherited Heart Disease)

Fifteen MeSH descriptors are explicitly associated with this study (CT/C-MeSH) where nine descriptors (highlighted in bold) are in the "Disease" or "Chemicals and Drugs" MeSH hierarchies (A). Thirty-one descriptors are associated with the study through PMIDs (CT/P-MeSH) where eight are in the required hierarchies (B). Links to over 236,000 sequences are identified where MeSH descriptors for 51 of these sequences completely overlap (relevancy score = 100%) with respect to CT/C-MeSH (C).

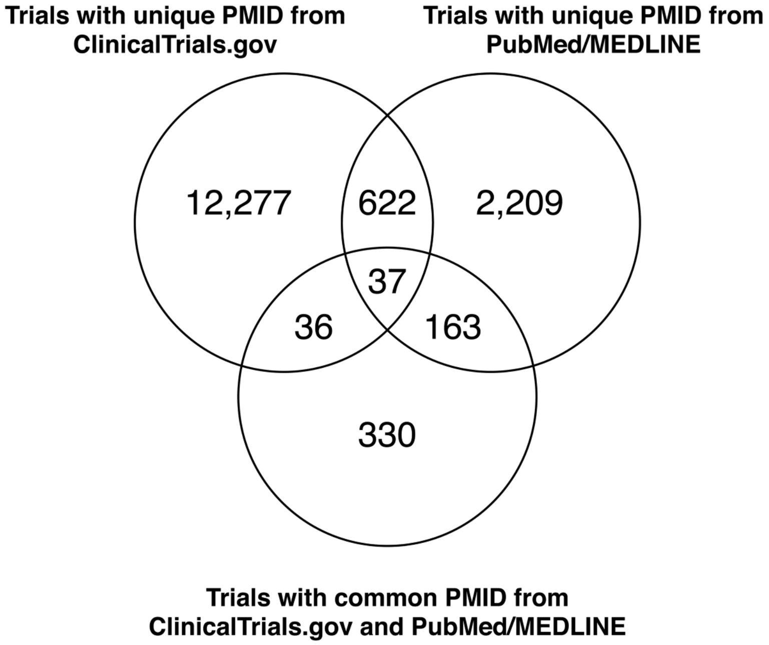

Trials with unique PMID from ClinicalTrials.gov
Trials with unique PMID from PubMed/MEDLINE
12,277
622
2,209
37
36
163
330
Trials with common PMID from ClinicalTrials.gov and PubMed/MEDLINE

Figure 7. Source of PMIDs for Clinical Trials

The contribution of PMIDs from the sources (ClinicalTrials.gov and PubMed/MEDLINE) is shown for clinical trials examined in the present study. The upper-left circle represents the number of trials with *unique* PMIDs contributed *only* by ClinicalTrials.gov; the upper-right circle shows the number of trials with *unique* PMIDs contributed *only* by PubMed/MEDLINE; the bottom circle depicts the number of trials with *common* PMIDs contributed by *both* ClinicalTrials.gov and PubMed/MEDLINE.

Table 1

Elements and Sources for Literature References Associated with Studies in ClinicalTrials.gov

Element	**XML Element**	**Public Web Site Headers**	**Source**
Background Reference	reference	Publications	ClinicalTrials.gov XML
Results Reference	results_reference	Publications	ClinicalTrials.gov XML
MeSH Descriptor	N/A	“Additional relevant MeSH terms”	ClinicalTrials.gov Web page
Additional Reference	N/A	“Additional publications automatically indexed to this study by National Clinical Trials Identifier (NCT ID)”	PubMed/MEDLINE (E-Utilities)

Table 2

Total Number of Literature References, Unique PMIDs or Citations, and Unique Trials

	All References			References with PMID			References with no PMID		
ClinicalTrials.gov	***Total***	***# Unique Citations***	***# Unique Trials***	***Total***	***# Unique PMID***	***# Unique Trials***	***Total***	***# Unique Citations***	***# Unique Trials***
Background References	60,840	53,135	10,937	57,132	49,513	10,539	3,708	3,516	1,474
Results References	9,467	9,235	4,179	7,537	7,322	3,581	1,930	1,913	1,173
All References	**70,307**	**61,695**	**14,320**	**64,669**	**56,167**	**13,465**	**5,638**	**5,419**	**2,513**
PubMed/MEDLINE	**3,790**	**3,790**	**3,397**	**3,790**	**3,790**	**3,397**	**N/A**	**N/A**	**N/A**
Total (CT/PMID)					**59,234**	**15,674**			

Table 3

Total and Unique Number of MeSH Descriptors Associated with Sequences and Trials

	# Sequences or Trials	# Total MeSH	# Unique MeSH (%)[*]
GB/P-MeSH	34,717,804	761,211,303	18,801 (74.65)
CT/C-MeSH	69,196	954,985	5,105 (20.27)
CT/P-MeSH	15,366	948,756	14,611 (58.01)
CT/CP-MeSH	70,213	1,903,741	16,063 (63.78)

[*] Percent of descriptors in MeSH 2009 (n = 25,186)

Table 4

Number of MeSH Descriptors Associated with Sequences and Trials in Each Hierarchy

	Sequences		Clinical Trials					
	GB/P-MeSH		CT/C-MeSH		CT/P-MeSH		CT/CP-MeSH	
Top-Level Hierarchy	***Total***	***%***	***Total***	***%***	***Total***	***%***	***Total***	***%***
Analytical, Diagnostic and Therapeutic Techniques and Equipment [E]	91,840,523	10.61	4,325	0.45	211,541	20.02	215,866	10.75
Anatomy [A]	28,317,599	3.27	237	0.02	39,980	3.78	40,217	2.00
Anthropology, Education, Sociology and Social Phenomena [I]	2,110,324	0.24	55	0.01	14,673	1.39	14,728	0.73
Chemicals and Drugs [D]	**109,383,265**	**12.63**	**391,588**	**41.15**	**145,454**	**13.77**	**537,042**	**26.74**
Disciplines and Occupations [H]	27,335,274	3.16	0	0.00	5,758	0.55	5,758	0.29
Diseases [C]	**11,253,228**	**1.30**	**511,887**	**53.80**	**120,240**	**11.38**	**632,127**	**31.48**
Geographicals [Z]	9,664,560	1.12	0	0.00	12,637	1.20	12,637	0.63
Health Care [N]	30,093,428	3.48	191	0.02	138,840	13.14	139,031	6.92
Humanities [K]	71,915	0.01	0	0.00	2,359	0.22	2,359	0.12
Information Science [L]	82,607,619	9.54	0	0.00	20,317	1.92	2,0317	1.01
Named Groups [M]	5,052,400	0.58	0	0.00	121,045	11.46	121,045	6.03
Organisms [B]	133,595,928	15.43	0	0.00	86,481	8.19	86,481	4.31
Phenomena and Processes [G]	322,307,090	37.23	7,622	0.80	94,470	8.94	102,092	5.08
Psychiatry and Psychology [F]	1,363,488	0.16	32,391	3.40	36,922	3.49	69,313	3.45
Publication Characteristics [V]	10,744,794	1.24	3,248	0.34	5,768	0.55	9,016	0.45
Technology, Industry, Agriculture [J]	91,840,523	10.61	4,325	0.45	211,541	20.02	215,866	10.75

Table 5

Number of PMID- and MeSH-based Links between Sequences and Trials

GenBank	ClinicalTrials.gov	# Sequences	# Trials	# Links
GB/PMID	CT/PMID	64,181	426	78,259
GB/P-MeSH	CT/C-MeSH	34,717,804*	67,672	39,109,150,340
GB/P-MeSH	CT/P-MeSH	34,717,804*	14,010	823,561,641
GB/P-MeSH	CT/CP-MeSH	34,717,804*	68,567	39,791,577,560

*estimated upper bound