**Analysis of consumer-health questions for development of question-answering technology**

**Bethany Harris**

**Jessi Van Der Volgen**

**Associate Fellows 2011-2012**

**August 2012**

**Project Sponsor: Dina Demner-Fushman, Lister Hill National Center for Biomedical Communications**

Table of Contents

[Abstract 3](#_Toc331693063)

[Introduction 4](#_Toc331693064)

[Methods 5](#_Toc331693065)

[Results. 8](#_Toc331693066)

[Discussion 12](#_Toc331693067)

[Recommendations 13](#_Toc331693068)

[Appendices 14](#_Toc331693069)

# Abstract

**Objectives:** The purpose of this project was to analyze the types of questions consumers submit to the National Library of Medicine, determine if the answers to the questions could be found in NLM resources, and create a taxonomy and annotation guidelines for consumer health questions.

**Methods:** A sample set of over 11,000 reference questions received by the National Library of Medicine customer service was examined and a subset of questions that could potentially be answered automatically was identified. We aligned questions with potential answer sources. Questions related to genetic conditions were initially determined to be the best candidates for automatic answering. A taxonomy of question types and indicators was created in a iterative process.

**Results:** A taxonomy and annotation guidelines for consumer health questions containing named diseases were created. The final schema notes the general type of question, as well as extraneous information and relevant misspellings. It annotates distinct entities, such as medical problems and genes, as well as words that indicate a particular clinical concept question type, such as prevention, symptoms, prognosis, or treatment. The guidelines were successfully used to annotate a set of 20 questions with good inter-annotator agreement among four annotators.

**Conclusion:** Consumer health questions are challenging to answer automatically because the questions may be complex, vague, or contain misspellings and it is difficult to understand the motivation for the question. The taxonomy and guidelines created in this project can be used for a machine learning task on questions containing named diseases and, with some modifications, may accommodate additional question types.

# Introduction

The National Library of Medicine (NLM) receives over 90,000 customer service requests per year. These requests cover many areas including questions about indexing policy, corrections to PubMed citations, and reference questions. The reference questions are further subcategorized and include questions about medications, diseases, lab and diagnostic tests, and many other topics. These questions can be submitted by users of NLM products, such as MedlinePlus, clinicaltrials.gov, or DailyMed, by completing a “contact us” form or by sending an email directly to NLM Customer Service.

NLM Customer Service responds to the questions with a stock reply, a customized stock reply, or a researched answer depending on what the question warrants. Responding to the request typically takes 4-10 minutes and as a result, it costs 8-11 dollars per question to respond. Because of the volume of questions NLM receives, a tool that could aid in automatically answering requests could reduce the workload of Customer Service and enable NLM to return an answer to the patron more quickly.

Question-answering technology is a type of information retrieval that processes a natural language request and returns a relevant section of text from a document in response. In order to automatically answer a question, the technology must be able to understand the question and have the appropriate resources from which to draw an answer. In this case, because the reference questions are primarily questions from the general public, it is important the resources used to answer the questions are understandable for this audience. NLM’s primary resources for consumer health information are  MedlinePlus ([http://www.nlm.nih.gov/medlineplus/)](http://www.nlm.nih.gov/medlineplus/%29) and Genetics Home Reference ([http://ghr.nlm.nih.gov/)](http://ghr.nlm.nih.gov/%29). An automatic question-answering technology could draw on these resources to supply answers to patrons. For example, NLM could respond to a question about the cause of a genetic disease with the relevant paragraph from Genetics Home Reference and link to the full source.

In order to train a question-answering system, it is important to have a framework for the types of questions that are submitted. Through question analysis taxonomy of question types can be created. Currently, no taxonomy for questions posed by consumers about health exists. Taxonomies have been created for generic clinical questions posed by family doctors regarding patient care, but these are unlikely to correspond sufficiently with the consumer health questions.

The objectives of this project were to determine if any of the customer service reference questions could potentially be automatically answered, analyze the question types, and to create a taxonomy of question types and guidelines for annotating the questions for a machine-learning task.

# Methods

The initial data set provided by NLM reference staff for exploratory analysis consisted of 11,000 reference requests submitted to NLM customer service between October 1, 2010-February 13, 2012 and categorized by staff as ‘reference questions’ in the Siebel information management system. The entire data set was examined through a combination of browsing performed by the project team and running the data through several programs to parse, categorize, and count key features within the data.

Since the first goal was to characterize the reference requests and determine if questions could be clustered into specific topic categories, the Siebel categories (originally applied by reference staff) were considered for continued use in the question taxonomy to be developed. Counts and percentages of questions in each Siebel category were obtained using a simple Python script. The project sponsor developed her own scripts utilizing the Unified Medical Language System (UMLS) to identify key topics and terms by parsing and counting the entries in the Siebel record fields of Category, Summary, Message, Anatomy, Drug, Intervention, and Problem. These categories and counts were useful in determining what consumer question topics were most prevalent and organizing them into meaningful clusters, whether or not they fit within the original Siebel categories. Questions that had been categorized as Disease/Condition/Therapy in Siebel were determined to often be suitable and decent candidates on which to train a new question-answering (QA) system.

The next step was to collect a sample of questions that appeared answerable with QA technology (as opposed to answerable by human staff). Specifically, the research team and library administration had an interest in determining if a significant proportion of reference questions had discrete information requests that could be satisfied with information contained in existing NLM consumer health resources, such as MedlinePlus and Genetics Home Reference. Questions were further browsed, with an emphasis on questions in the Disease/Condition/Therapy category, and examples were collected that contained obvious requests for consumer health information. Example questions were organized into handouts and categorized according to the mechanism by which it was sent, such as the contact form on MedlinePlus or Genetics Home Reference, direct contact with NLM Customer Service, or any other mechanism.

During the iterative process of question analysis and the first stages of question categorization (*i.e*. taxonomy development), several handouts containing example reference questions, their mechanism of origin, and potential answers and/or answer resources for the questions were developed by the associates and discussed with the research team (including library administration). Major points of discussion included: what and how is the consumer asking, what is the likely motivation behind the questions, where was the consumer when they submitted the question (what webpage), and what kinds of question patterns were a) common and b) structured in such a way to be suitable for QA technology to process.

Handouts were often composed to include possible answers and which resources contained those answers. The associates researched the answers to the consumer questions as they would any reference question as if they had received it themselves at the reference desk. The potential answer resource was documented in the handout and specific areas of the web page (where the answer was located) were highlighted to note the position of the useful information. Notes were taken and discussed with the research team as to any issues or concerns with the available (or lack of) information in the NLM consumer health resources and other possible web resources, such as the Genetic Testing Registry (GTR) and various NIH-produced consumer resources that were considered.

Through the research team discussions over the handouts, familiarity with the questions set increased and specific question types emerged as appearing to be frequently answerable, with potential answers being derived from NLM consumer health resources. Questions containing specific mentions of a disease or medical problem were found to often have answerable question structures (*e.g.*, ‘What is [medical problem]?’ or ‘What are the symptoms of [disease name]?’). It also became apparent that questions seeking genetic information featured answerable question structures (*e.g.*, ‘Does [disease name] run in families?’) and were well-suited to be answered by a combination of Genetics Home Reference and MedlinePlus materials.

It was decided that a formal data set consisting of genetics questions containing a named disease or medical problem would be used for formal taxonomy development. The criteria used for extraction of a genetics question set from the original set of 11,000 questions included: 1) a disease or condition contained in the UMLS, and 2) any of these selected words or partial words that likely indicate a genetics question: GARD (the acronym for the Genetic and Rare Diseases Information Center, from which several genetics questions were sourced), herit, hered, genetic, gene, mutation, DNA, chromosome, and/or famil. Using the partial words allowed us to capture variations of related words (*e.g.*, ‘inherit’ and ‘heritable’).

The project sponsor extracted the genetics question sets based on our criteria and compiled one set that accounted for ‘famil’ and one that did not, since we wanted to capture questions like ‘Is this a familial disease?’, but many reference requests also include the word ‘family’ without relating to  a request for genetics information. The two sets were then filtered by the associates to compile a final set of questions for eventual annotation. Questions were eliminated if they did not meet our criteria, such as if they did not contain a medical problem, did not have a genetic information component, or did not seem to contain a discrete information need that might possibly be answerable by QA technology. Eventually, it was later decided to build upon the unique genetics question set by adding non-genetics questions with a named medical problem and those questions were compiled by two new student colleagues working under the supervision of the project sponsor.

As certain question features and indicators emerged as useful indicators for training QA technology, formal taxonomy and annotation guideline development began. Handouts used during research team discussions typically included a working version of the annotation schema (*e.g.*, yellow highlighting = disease, pink highlighting = genetics information) with guidelines to help explain how to use the annotation schema correctly and consistently, as well as numerous samples of questions annotated with the current annotation schema. This process of developing the questions taxonomy and annotation guidelines, annotating example questions, and discussing whether the schema worked or not on real consumer health questions was repeated for several cycles (biweekly meeting with library administration) and the taxonomy and schema were adapted according to feedback.

Originally, questions were annotated in Microsoft Word, using features such as text highlighting and various font colors and/or underlining. This was useful for discussion in group meetings. Later, the research team reviewed a handful of annotation software tools that would be used to perform the actual annotation that would become the input for the QA system in development. After testing out a few tools, the associates decided upon the Stanford annotator, a manual annotation tool. The decision was based upon the tool’s ease of use, intuitive interface and annotation features, the ability to apply multiple annotation tags to the same text, and the fact that the annotation tags were made of color-coded XML tags, which facilitated both human and machine readability of the annotated questions.

Once a semi-final annotation schema was agreed upon, the associates and the two student colleagues proceeded with annotating 20 example questions from the genetics question set to evaluate inter-annotator agreement using the current guidelines. All four annotators used both MS Word and the Stanford annotator tool to independently annotate the same 20 questions. The four annotated versions were compared and differences were discussed with the research team. Consensus annotation was determined for each of the 20 questions and this set of questions became the gold-standard annotated set to which all future annotators may refer. The annotation guidelines were finalized based on the inter-annotator agreement discussions. The final annotation schema along with the gold-standard annotated 20 questions can be used to train new annotators and research team members as well as be used for input in programming the QA system being developed.

Results

The primary outcomes of this process are a taxonomy with guidelines for applying annotation schema and a set of twenty consensus-annotated questions. Several intermediate products were valuable in the development and are described here.

The Python script output (Appendix A) generated some useful counts and characteristics about the original data set of 11,000 questions. The Siebel category ‘Disease/Condition/Therapy (General Public)’ was found to be the largest category of questions and contained 3,690 reference requests. Counts of the topics that occurred in the subject lines also yielded some interesting information, such as the number of reference requests that originated from MedlinePlus (841), which indicated plenty of consumers were utilizing the consumer health resource yet still had unanswered questions.

Approximately 12 handouts were constructed for review by the research team (including library administration), not including the 13 iterations of the annotation guidelines. The handouts served to share progress and challenges in developing the taxonomy and guidelines, to generate discussion among all parties, and to make decisions to move forward. Table 1 describes several handouts that were influential in the development of the taxonomy and guidelines and the decisions that resulted from each.

Typically, the handouts featured a subset of consumer health questions along with the service the patron was using, such as DailyMed or MedlinePlus. If available, the handouts also included the specific webpage the user was viewing when they submitted their questions to NLM. As the development of the taxonomy began, consumer questions were grouped and categorized according to proposed types. Later handouts added potential answers from MedlinePlus, Genetics Home Reference, and other NIH resources. An example of a handout with questions, from pages, and potential answers can be found in Appendix B.

| Handout | Description | Decisions and Next Steps |
| --- | --- | --- |
| 1 | 20 questions from the Disease/Condition/Therapy Siebel category identified as potentially answerable, organized by from service (*e.g*., MedlinePlus vs. direct email) | Genetic questions look interesting. Should examine more genetics questions.  |
| 2 | 14 questions from the entire question set containing the word “genetic” and identified as potentially answerable, organized by from service.  | Genetics questions appear promising. Pursue taxonomy development around genetics questions to begin.  |
| 3  | 19 questions with the word “genetic” or clearly about genetics, grouped by three types:1. Known disease: is it genetic? How is it inherited?
2. Known disease: general info
3. Unknown disease: what is it?
 | Revised genetic questions into two types:1. What is it? Is it genetic?
2. How is it inherited? Will I get it?

Questions with unnamed diseases are too difficult for now.  |
| 4  | 17 questions about genetics without the word “genetic,” with a screenshot of the from page where available.  | Many genetics questions can be identified without the word “genetic.” Sometimes the patron is on the page that appears to answer their question.  |
| 5 | 15 genetics questions with a named disease grouped by five types:1. What is it? Is it genetic?
2. How is it inherited? Will I get it?
3. Is X related to Y?
4. Is there a test for it?
5. What is the treatment or prevention?
 | Added question types and began annotation schema by highlighting indicator words in different colors.  |
| 6 | 23 genetics questions grouped by type | Questions were independently annotated, then reconciled to consensus. Additional types and indicators added.  |
| 7 | 21 genetics questions annotated according to annotation schema with 13 indicators for question types. | Revised annotation schema and guidelines. |
| 8 | 53 potentially-answerable, non-genetics questions  | Many questions with named diseases fit the existing schema. Questions about drugs and lab values are good candidates for adding to the schema.  |
| 9 | 22 annotated genetics questions with screenshots of both the from page and potential answer sources | Some questions may require a follow-up question from customer service in order to be answered. |
| 10 | Annotation schema and guidelines consisting of general guidelines, annotations for entities, and annotations for indicators.  | Tested on a set of 20 questions and revised to form final schema and guidelines.  |

Table 1. Representative handouts and resulting observations and decisions.

The extraction of possibly-answerable genetics questions (based on our criteria of containing a named medical problem and including one or more words from our list of genetics-related terms often used by consumers) yielded a total of 585 questions. After the associates filtered the questions down to just those that were likely-answerable by QA technology, we were left with a data set of 217 questions to be used in final annotation. This genetics questions data set was also heavily sampled from for the remaining handouts, in the form of selected questions annotated according to whichever was the most current annotation schema at the time the handout was created.

Approximately 13 versions of the annotation schema were created, resulting in a final schema to be used in future annotation (Appendix C). As a result of the iterative process of creating and refining the question taxonomy and annotation guidelines, we think that it is robust and scalable and can be used for most questions in which there is a clear question and has a named disease. The final annotation guidelines are comprised of general guidelines, guidelines for annotating distinct entities, and guidelines for annotating indicators for questions types and examples of each type.

The general guidelines consist of general principles and annotations that can apply to an entire question or large portions of a question. Extraneous information is annotated as such, relevant misspellings are annotated and corrected, and anaphora for relevant nouns and noun phrases are annotated. An entire question or segment can be tagged as not answerable, genetic, non-genetic, or follow-up. Annotation for genetic or non-genetic provides information for machine learning as to which answer resources are likely to be most useful. Annotating a question for follow-up indicates that the question could be answered if clarifying information could be obtained from the person submitting the question.

Guidelines for annotating distinct entities include instructions for annotating medical problems, causal or exposure agents, genes, proteins, or mutations, and tests (diagnostic or genetic). The third part of the guidelines is for annotating indicators for question type. In this case, the annotated words are those that indicate a need for general information or a particular clinical concept. Clinical concepts covered include diagnosis, prevention, treatment, probability of developing a disease, etiology, prognosis, symptoms, and complications. For example, ‘Is there a way to prevent’ would be annotated as an indicator that the patron is seeking information about prevention.

The final result is a question taxonomy with useable guidelines, as indicated by two new annotators (not involved in the creation of the guidelines) being able to immediately apply them while annotating sample questions. The guidelines are usable for genetic questions, but scalable to other types of questions, so that the taxonomy may address most questions with named diseases that ask a clear health-related question. For now we have counted questions about lab values, drugs, unnamed diseases (usually asking for a diagnosis) as unanswerable, but feel optimistic that the guidelines could eventually be used with many of those questions with slight modifications.

The guidelines were successfully used to annotate a set of 20 questions with good inter-annotator agreement among four annotators (see Appendix D). The four annotators collaborated to create a gold-standard of annotated questions and this process did not identify any need for new indicators, but did add clarifying statements to the guidelines. This process allowed us to resolve inconsistencies in how guidelines were applied and add minor details to the established guidelines, but did not make any substantial changes to the final annotation schema.

# Discussion

In examining the question set, several challenges emerged. Many of the requests include misspellings, unclear grammar or ambiguous abbreviations. While some could be resolved and understood, many others could not. This is likely to be improved with spell-check capabilities and further machine learning. Secondly, the information need can be difficult parse from the request. Several of the requests included lengthy paragraphs that contained numerous details regarding family history, past medical problems, and other personal but irrelevant information. Because of this, an annotation for extraneous information was added to the schema to help train a question-answering system regarding which parts of the question are pertinent in understanding the question and providing a relevant resource.

As the full question set was explored, we realized that there are a large number of questions submitted to NLM in which the user lists symptoms or personal health history in hopes of being told what disease they might have. This realization helped us weed out questions with unknown or unnamed diseases, since medical librarians are not in the field of diagnosing library users. There are also several requests in which the question is slightly unclear, but is potentially answerable with some clarifying information. Because of this, an annotation for ‘follow-up’ was created. This notation could also be used in situations where the user appears to viewing a webpage that addresses the question that they have submitted.

Finally, many of the questions ask about complex relationships. For example, a question might ask about a connection between a childhood illness and adult disease. This type of question was originally recognized as a distinct class and included in early versions of the schema, but it became clear that our resources do not satisfactorily answer these consumer questions.

The next steps of the project are to use the guidelines to annotate a training corpus of 300 questions to aid in further development of the question-answering system. The performance of the QA system can then be evaluated using an additional corpus of 100 reference questions. Additional training and testing of the system will highlight any problems with the current taxonomy and inform further modifications or additions that are needed.

As work progresses, the taxonomy and guidelines can be expanded to address questions other than those with a named disease. Questions regarding a particular gene, lab value, or drug appear to have patterns that indicate they may be good candidates for adding to the existing schema, which would vastly expand the number of questions on which the question-answering system could subsequently be trained and tested.

# Recommendations

Matching the questions posed by patrons to NLM’s available resources provided a unique look into how NLM resources are currently meeting or not meeting the needs of these consumers. Often the question would be posed from a page that seemed to answer the question. This provides NLM with an opportunity to use the reference questions as a form of feedback for evaluating and improving the existing resources. For example, Genetics Home Reference has a wealth of information but several questions evidenced that it may need more consumer-friendly language to be used in automatically answering the variety of questions that are received by NLM Customer Service.

As additional question types are incorporated into the question-answering technology, other answer resources will also need to be indexed for retrieval. Continued development on the project should consider which current NLM and NIH resources should be added to the knowledge base from which to draw. Finally, once the taxonomy and guidelines are refined and tested, making them publicly available to the research community could to contribute to question-answering technology particularly for questions in the consumer health domain.

Appendices

Appendix A: Python script results

Counts from Siebel subcategories in chs.csv file (highest to lowest frequency):

[(3690, 'Dis/Cond/Ther (Gen Public)\n'), (2494, 'Drug Information\n'), (1399, 'Research Request\n'), (1341, 'Other Reference Ques\n'), (390, 'Find a Doctor or Expert\n'), (388, 'Free/Discounted Med Tx Sources\n'), (284, 'Herb/Nutritional Dietary Sup\n'), (275, 'Lab or Diagnostic Tests\n'), (240, 'Consumer Health Information\n'), (221, 'Medical Devices\n'), (193, 'Statistics/Cost Expend\n'), (188, 'Hospital/Facility/Health Org.\n'), (116, 'Dis/Cond/Ther (Health Prof)\n'), (107, 'Film/Video/Stock\n'), (78, 'Career Guidance Information\n'), (61, 'GARD\n'), (53, 'Psych/Educ Eval Tests\n'), (32, 'Biographical Info\n'), (31, 'Animal/Pet/Veterinary Ques\n'), (26, 'Patient Records\n'), (19, 'Codes\n'), (11, 'Acronyms & Initials\n'), (3, '\n')]

Counts from Summary topics in chs.csv file (only Top 30 subjects):

[(841, 'MEDLINEPLUS SERVICE REQUEST\n'), (344, 'QUESTION - GENERAL INFORMATION\n'), (148, 'CUSTOMER SERVICE REQUEST\n'), (121, 'QUESTION - SPECIFIC TRIAL\n'), (120, '\n'), (45, 'DRUG INFO\n'), (42, 'SIEBEL BACKUP PROCESS\n'), (42, 'HANGUP\n'), (39, 'SUGGESTION\n'), (31, 'QUESTION\n'), (21, 'RESEARCH HELP\n'), (21, 'HELP\n'), (18, 'GENERAL COMPLAINT\n'), (18, 'CONSUMER HEALTH INFO\n'), (17, 'SHINGLES\n'), (17, 'DAILYMED FEEDBACK\n'), (16, 'FIND A DOCTOR\n'), (14, 'INFORMATION\n'), (13, '< NO SUBJECT >\n'), (11, 'RESEARCH\n'), (11, 'NEED HELP\n'), (11, 'COMPLIMENT\n'), (10, 'STATISTICS\n'), (10, 'RE: QUESTION - GENERAL INFORMATION\n'), (10, 'QUESTION - HOW TO USE SITE\n'), (10, 'HI\n'), (10, 'DISEASE INFO\n'), (10, 'DIABETES\n'), (9, 'PLEASE HELP\n'), (9, 'MSDS\n')]

### Appendix B: Sample section of a handout.

Includes the consumer’s question, annotated according to the schema version current at that time, the from service, the from page with screenshot (if available), and a potential answer.

4.Subject: Diabetes

**Is it possible to get** diabetes even if it **does not run in your family**?

From Service: MedlinePlus

From Page: <http://vsearch.nlm.nih.gov/vivisimo/cgi-bin/query-meta?v%3Aproject=medlineplus&amp;query=asthma>



**Answer**: <http://diabetes.niddk.nih.gov/dm/pubs/overview/#who>

Diabetes is not contagious. People cannot “catch” it from each other. However, certain factors can increase the risk of developing diabetes.

Type 1 diabetes occurs equally among males and females but is more common in whites than in nonwhites. Data from the World Health Organization’s Multinational Project for Childhood Diabetes indicate that type 1 diabetes is rare in most African, American Indian, and Asian populations. However, some northern European countries, including Finland and Sweden, have high rates of type 1 diabetes. The reasons for these differences are unknown. Type 1 diabetes develops most often in children but can occur at any age.

Type 2 diabetes is more common in older people, especially in people who are overweight, and occurs more often in African Americans, American Indians, some Asian Americans, Native Hawaiians and other Pacific Islander Americans, and Hispanics/ Latinos.

Tutorial: <http://www.nlm.nih.gov/medlineplus/tutorials/diabetesintroduction/htm/index.htm>

5. Subject: bipolar disorder

**Is** bipolar disorder **genetic?** **My husbands mother** has a very strong case of bipolar disorder and there other minor but **present cases in his family. He and his sister are** both diagnosed **without the disorder**. Though **do** my **children have the chance of developing** it in the future? **What are the likely chances?** **My family doesn't have any cases of** the disorder.

From Service: MedlinePlus

From Page: <http://www.nlm.nih.gov/medlineplus/bipolardisorder.html>



Answer: <http://www.nimh.nih.gov/health/publications/bipolar-disorder/complete-index.shtml>

* Bipolar disorder tends to run in families, so researchers are looking for genes that may increase a person's chance of developing the illness. Genes are the "building blocks" of heredity. They help control how the body and brain work and grow. Genes are contained inside a person's cells that are passed down from parents to children.

Children with a parent or sibling who has bipolar disorder are four to six times more likely to develop the illness, compared with children who do not have a family history of bipolar disorder. However, most children with a family history of bipolar disorder will not develop the illness.

6. Subject: mitral valve

**is** mitral valve prolapse **inherited**

From Service: MedlinePlus

From Page: <http://www.nlm.nih.gov/medlineplus/ency/article/000180.htm>



Answer: <http://www.nlm.nih.gov/medlineplus/ency/article/000180.htm>

* Some forms of mitral valve prolapse seem to be passed down through families (inherited). Mitral valve prolapse has been associated with [Graves disease](http://www.nlm.nih.gov/medlineplus/ency/article/000358.htm).

7. Subject: Leber hereditary optic

**Can a Male pass** Leber hereditary optic neuropathy **to** his son? If so **can the female** offspring of the son **pass** Leber hereditary optic neuropathy **to** his daughter? In my case **my paternal grandfather** has Leber hereditary optic neuropathy and I want to know the **chances of passing this on to** my offspring. **I am a female.**

From Service: CustServ

From Page: <http://ghr.nlm.nih.gov/condition/leber-hereditary-optic-neuropathy>



Answer: <http://ghr.nlm.nih.gov/condition/leber-hereditary-optic-neuropathy>

* **This condition has a mitochondrial pattern of inheritance**, which is also known as maternal inheritance. This inheritance pattern applies to genes contained in mitochondrial DNA. Because egg cells, but not sperm cells, contribute mitochondria to the developing embryo, only **females pass mitochondrial conditions to their children**. Mitochondrial disorders can appear in every generation of a family and can affect both males and females, but **fathers do not pass mitochondrial traits to their children.**

Appendix C: Final annotation schema and guidelines

**General Guidelines**

1. Extraneous (light gray)

Annotate words or segments with entity and/or indicator tags if they are relevant to the question. Only annotate summary if it is offering new or otherwise unstated information pertinent to the question at hand. Use the minimal amount of cues when annotating. Punctuation is never annotated. The rest can be tagged (only at the level of whole sentences or paragraphs) as extraneous to indicate what should be ignored.

1. Repeated and Nonspecific (anaphora) Mentions

Annotate anaphora for all relevant nouns and noun phrases. A repeated mention of a noun or noun phrase is only considered relevant if it adds to the comprehension of the question at hand. In other words, do not annotate rephrased sentences.

1. Misspelling (turquoise)

When annotations of relevant text include misspelled words, apply a secondary annotation to each misspelled word indicating a misspelling. If spelling is not clearly known to be correct or incorrect, look it up. If the correct spelling of the intended word is unambiguous, add COR=“[correctly spelled word(s)]” at the very end of the opening tag in the Stanford Annotator.

1. Annotate question record number or relevant segment of a question with:

(a) **Not answerable** (*If entire question can be marked ‘not answerable, do not annotate any segments of the message*)

(b) **Genetic**

(c) **Not-genetic**

(d) **Follow-up** (*Indicates that the question could possibly be answered with additional information from a follow-up question*)

***Note:*** *Questions may be annotated with one or more of the above.*

**Distinct Entities**

1. Medical problem (yellow)

Annotate each mention of a relevant disease, condition, gene, finding or specifically named symptom including all abbreviations, acronyms, and anaphora (e.g. “this disease”, as in example below).

 Example: **What** exactly **is** Ectrodactyly **What causes** this disease?

 ***Note:*** *Annotate mentions that directly pertain to the question. Ignore unessential mentions of disease, such as in excessive medical/family history descriptions. Annotate distinct mentions of entity separately especially if they are in close proximity to each other.*

 Causal/Exposure agents (**underlined green**)

Annotate all mentions of the presence or absence of the pertinent disease in the family (for genetic questions) and relevant exposure context (for environmental questions, e.g. being exposed to a sick person).

Example: **Is it possible to get** diabetes even if it **does not run in your family**? (genetic)

Example: Can you get shingles after the singles vaccine with no blisters? (environment)

1. Genes or proteins (**underlined black**), mutations (**underlined blue**)

Annotate specific mentions of genes, proteins, and mutations.

Example: Also can you please try to explain **what the** **ASPA gene** **does** in 8th grade terms.

***Note:*** *If gene name is part of the disease name just annotate as a disease.*

1. Tests (**bold red**), location of a test (**underlined red**)

Annotate words pertaining to the existence or location of any diagnostic or genetic tests.

Example: i just had a question about the **ALDH gene** **testing** (**patch test**). i heard that it's done in japan. i was wondering **if they had the test** in **any other countries**, if so, **is there a list of places** where i can look it up.

1. Treatments (**aqua green**)

Annotate specific mentions of named treatments.

Example: Is **iontophoresis** safe for cancer patients.

1. From Page (**bluish grey**)

Annotate from page cite if one is given.

Example: **FROM: http://vsearch.nlm.nih.gov/vivisimo/cgi-bin/query- meta?v%3Aproject=medlineplus&query=Desmoid+Fibroma&x=15&y=16**

**Indicators for Question Types**

1. General information or definition (**bold light blue**)

 Annotate only essential words expressing a need for general information.

 Example: Please **email me** any **info** you may have **on** this disease.(Ehlers Danlos Disease).

***Clinical Concepts***

1. Diagnosis (**bold red**)

Annotate all words pertaining to the diagnosis of a disease or condition (analogous to the **red** used in diagnostic tests, but here it is an indicator rather than distinct entity).

Example: **How is** NAGS deficiency **diagnosed** and what are the management options available if you are **diagnosed** with this deficiency

1. Probability of developing disease or condition (**bold pink**)

Annotate all essential words pertaining to the probability of developing or contracting a disease or condition (e.g. ‘Will I or [a person] get it?’).

Example: **Is** bipolar disorder **genetic**? My husbands mother has a very strong case of bipolar disorder and there other minor but present cases in his family. He and his sister are both diagnosed without the disorder. Though do my children have the chance of developing it in the future? What are the likely chances? My family doesn't have any cases of the disorder.

1. Prevention (**bold purple**)

Annotate all words and phrases pertaining to the prevention of diseases.

Example: What would be or is there a way ro prevent spina bifida from occuring in an unborn child?

1. Treatment (gray)

Annotate all words pertaining to the treatment, management, or possible cures for a disease.

Example: How can found the articles regarding treatment or any development of the treatment on desmoid fibroma on this website?

1. Etiology /exposure (**bold blue**)

Annotate all words pertaining to the cause of a disease, whether genetic or environmental.

Example: The poject requires help from an expert and I would like to know facts about **what** **chromosome number** is affected **to cause** Rett Syndrome and **what testing and or screening is used to diagnose** Rett Syndrome. If you could reply in the next few days that would be great!

1. Prognosis (bright green)

Annotate all words pertaining to prognosis, life expectancy, and expectations for daily living.

Example: amyopathic dermotomyositis  Have recently been diagnosed with the above disease. What is the prognosis ?

1. Consequences/complications (**underlined brown**)

Annotate all words pertaining to complications

Example: What are the complications of vulvoginitis?

1. Symptoms/manifestations: (**bold orange**)

Annotate all words pertaining to the presence or absence of symptoms and how the disease manifests or presents itself. If a list of specific symptoms is provided, annotate each symptom separately. In relevant cases, symptoms should be double annotated as a problem as well.

Example: I am an 8th grade student at [school] doing a research project on Canavan Disease. I was wondering **what some of the internal and external symptoms of** this disease

Example: I have been diagnosed with Hereditary cerebellum ataxia. My grandfather - mother - 2 brothers and I have this disease. My grandfather, mother and 1 brother have died as a result and I am living with it. My balance, co-ordination, eye movement, speech are all deterioorating rapidly. I need help to control this disease or to improve all these prblems I am experiencing at the moment.

Appendix D: Examples of independent and consensus-annotated questions

1. <SR>1-46997525</SR> Please email me any info you may have on this disease.(Ehlers Danlos Disease).

Annotator #1
<SR>1-46997525</SR> Please **email me any info you may have on** this disease.(Ehlers Danlos Disease).

Annotator #2
<SR>1-46997525</SR> Please email me **any info** you may have on this disease.(Ehlers Danlos Disease).

Annotator #3
<SR>1-46997525</SR> Please email me **any info** you may have **on** this disease.(Ehlers Danlos Disease).

Annotator #4
<SR>1-46997525</SR> Please email me **any info** you may have **on** this disease.(Ehlers Danlos Disease).

Consensus annotation: <SR QT=”GENETIC”>1-46997525</SR> Please email me any info you may have on this disease.(Ehlers Danlos Disease).

1. <SR>1-47614348</SR> When you have been told you have a Urachal Cyst what does that mean And how can it be fixed

Annotator #1
<SR>1-47614348</SR> **When you have been told you have** a Urachal Cyst **what does that mean** And how can it be fixed

Annotator #2
<SR>1-47614348</SR> When you have been told you have a Urachal Cyst **what does that mean** And how can it be fixed

Annotator #3
<SR>1-47614348</SR> When you have been told you have a Urachal Cyst what does that mean And how can it be fixed

Annotator #4
<SR>1-47614348</SR> When you have been told you have a Urachal Cyst **what does that mean** And how can it be fixed

Consensus annotation: <SR QT=”NON-GENETIC”>1-47614348</SR> When you have been told you have a Urachal Cyst **what does that mean** And how can it be fixed