Guidelines for Annotating Named Entities in Consumer Health Questions:

The goal of this study is to annotate all named entities that are relevant to question understanding tasks, such as frame extraction and focus recognition, in a corpus of consumer health questions. The corpus consists of 1548 questions submitted by consumers to NLM in 2014 and 2015. The selected questions focus on disease/condition/therapy and drug/supplement information. They were selected based on whether they were answerable using authoritative NIH resources. Personal health information was removed from the questions before annotation.

Based on a test annotation of 20 questions by four annotators, we determined the following entity categories to be relevant for our goals. Some UMLS semantic groups that may be relevant for the categories are also given[1]:

Anatomy	Includes organs, body parts, and tissues. Entities at the cellular and molecular level should not be considered for this type. Examples include *head*, *neck*, *gum*. UMLS: *Body System, Anatomical Structure, Body Part, Organ, or Component, Body Space or Junction.*
CellularEntity	Includes anatomical entities at the molecular or cellular level in the human body. Examples: *giant cell, hemoglobin, chromosome 4.* UMLS: *Cell, Cell Component, Embryonic Structure*
DiagnosticProcedure	Includes tests and procedures used for diagnosis. Examples: *biopsy, hemoglobin, iron levels.* *UMLS: Diagnostic Procedure, Laboratory Procedure*
Measurement	A quantity that is a core attribute of a named entity, such as dosage of a drug. Example: *10mg, 2%.*
DrugSupplement	Includes substances used for therapeutic purposes. Examples: *atenolol, atenolol 50 mg, campho-phenique, campho-phenique treatment*. UMLS: *Clinical Drug, Pharmacologic Substance, Steroid, Vitamin*
Food	Refers to specific nutritional substances. Examples: *eggs, breads, meat*. UMLS: *Food*
GeneProtein	Includes specific genes and gene products. Examples: *BRCA1, BRCA1 gene, GLUT4 protein*.

[1] UMLS semantic types are given as examples only, and should not be used as the sole criterion in determining the type.

	UMLS: *Gene or Genome, Amino Acid, Peptide, or Protein*
GeographicLocation	Includes countries, cities, etc. Examples: *India, Singapore, etc.* UMLS: *Geographic Area*
Lifestyle	Refers to daily and recreational activities. Examples: *smoking, yoga, etc.* UMLS: *Daily or Recreational Activity*
Organization	Includes institutions as well as their subparts. Examples: *navy, hospital, California hospitals, etc.* UMLS: *Organization*
PersonPopulation	Includes individuals (gender, age group, etc.) and population groups. Examples: *daughter, grandfather, female, 16 year old, elderly, war veteran, citizens of China, etc.* UMLS: *Age Group, Family Group, Population Group, Human*
Problem	Includes disorders, symptoms, abnormalities, complications, etc. Examples: *autoimmune disease, rheumatoid arthritis, broke, allergic, problem with cholesterol, cholesterol, HIV, lithotripsy complications* UMLS: *Disease or Syndrome, Pathologic Functions, Neoplastic Process*
ProcedureDevice	Refers to procedures or medical devices used for therapeutic purposes. It also includes non-specific interventions that do not involve drug therapy. Examples: *shingles treatment, nephrolithotomy, implants* UMLS: *Therapeutic or Preventive Procedure, Medical Device*
Profession	Includes mentions referring to an occupation, discipline, or expertise. Examples: *dermatologist, dr, surgeon, etc.* UMLS: *Professional or Occupational Group, Occupation or Discipline*
Substance	Includes chemicals, hazardous substances, and bodily substances. Examples: *iron, blood, cholesterol, alcohol* UMLS: *Biologically Active Substance, Carbohydrate, Lipid, Chemical, Inorganic Chemical, Hazardous or Poisonous Substance, Body Substance*
OTHER	Includes entities that are relevant to question understanding, but do not neatly fit in one of the categories above. These should be marked, so that if there is a critical mass of certain entity types, we can add those categories, as well.

A mention can be annotated with multiple types when it is ambiguous. For example, *hemoglobin* can be annotated both as CellularEntity and DiagnosticTest.

Unspecific mentions or anaphoric mentions should not be annotated. For example, *symptoms, this medication*, and *my illness* are not valid named entities. A principle to use to determine whether a mention is unspecific is to check whether the mention is in the name of an entity category (e.g., *problem, device, procedure, profession, substance, drug, gene, protein* would all be considered unspecific). Another useful criterion may be to check whether the word appears as the head of a UMLS semantic type name (e.g., *activity, process, disease, syndrome, function, enzyme, organization, abnormality*). On the other hand, some more specific terms that can still be considered quite general can be annotated (such as, *surgery, operation, cancer, trauma*). A list of overly general Problem terms is provided in Appendix. During the course of this study, the annotators will also collaboratively develop lists of general terms and non-general terms (those that might be confusing) based on their reconciliation. These lists will be in Appendix, as well.

Nested annotations are allowed. The criterion for a nested annotation is that it should refer to an entity different from but related to that referred to by the top-level annotation, excluding general terms. Some relations that can be considered between the top level and inner entity are specialization, attribute, and location, among other types of semantic modification.

- In "left *hand ring finger*", "*left hand", "ring finger", "finger",* and *"hand"* can be annotated in addition to the top-level "*left hand ring finger*", since they indicate different entities.
- In "*left ankle bone", "bone", "ankle", ''left ankle'', "ankle bone",* and "*left ankle bone"* can be annotated.
- In "*TDAP injection"*, both "*TDAP"* and "*TDAP injection"* can be annotated, since the latter refers to a specialization of the former.
- In "*BCG treatment"*, on the other hand, "*BCG"* should not be annotated, since it essentially refers to the same concept as "*BCG treatment"* and *"treatment"* is a general term.
- If "*blood test*" is annotated as DiagnosticProcedure, then it is not necessary to annotate "*blood*" as a DiagnosticProcedure, since both annotations would refer to the same entity. However, "*blood*" can still be annotated as a nested entity with the type Substance.

Syntactic form of the mention should be taken into account. A simple noun phrase (including pre-modifiers and head word) is often a better candidate for an entity annotation than a noun phrase with attached prepositional phrases (a macro-NP). For example, "*lithotripsy complication"* is a good candidate for annotation. On the other hand, "*itches in my leg"* is not. Some macro NPs, especially involving the preposition *of*, can be good candidates if the top-level annotation refers to an established entity (e.g., *50mg of atenolol)*. We think the prepositions *of* and *for* are generally good, while *in*, *to*, and *at* are generally not. Avoid annotating expressions crossing macro-NP boundaries. For example, do not annotate *menstruation was heavy* as a Problem in its entirety.

Named entities do not have to correspond to nouns. In the following, *sensitive* can be annotated as a Problem, for example: *I am still sensitive in my back and head*.

Non-contiguous named entities can be annotated (using brat's Add Fragment feature). For example, in *diabetes mellitus type 1 and 2*, four annotations can be created: three contiguous annotations (*diabetes mellitus type 1*, *diabetes mellitus*, and *diabetes*) and the non-contiguous entity *diabetes mellitus type…2*.

Ensure that you do not include determiners in the annotation. For example, annotate *insurance company* instead of *the insurance company*.

Ensure that Measurement annotations are core attributes to named entities. For example, annotate *20mg* in *20mg of atenolol*. But, skip *4&8 hours* in *Onset of autoimmune disease between 4 &8 hrs. later*.

Normal organism functions should not be annotated. For example, *menstruation*. However, note that such a mention could also indicate a Problem, and it should be annotated if it clearly does. For example, *heavy menstruation* can be annotated as a Problem.

Ignore whether the entity is negated or is stated as uncertain. For example, in *is not ruptured,* you can annotate *ruptured* as a Problem.

Do not annotate de-identified portions of questions, such as '[LOCATION]' or '[PROFESSION]'.

If you see PHI that is not de-identified, notify Halil immediately.

Using Pre-Annotations:

A subset of questions has been pre-annotated using several systems. The annotation type name includes the tool used: SE (Essie), MM (MetaMap Lite), Lex (lexical lookup), DB (Dbpedia), and i2b2 (CRF-based).

SE and MM annotations use UMLS semantic type abbreviations. For full names, see http://ii.nlm.nih.gov/Publications/Docs/SRDEF_2014AA.txt.

These pre-annotations serve as a reference only, and the annotator should create a new annotation for each mention deemed relevant. One of the entity types described earlier (Anatomy, Problem, etc.) should be used for the new annotations. There is no need to remove the pre-annotations.

Brat Notes:

If an entity spans multiple lines (line numbers are displayed on the left), annotate it as a multi-part annotation, using brat's Add Fragment feature.

Ensure that you capture the boundaries of the named entities correctly. It can be a bit tricky to get the boundaries right, if the mention spans multiple tokens.

Appendix

Overly Generic Problem Terms from MetaMap annotations:

Aberrations
Abnormal development
Abnormalities
Abnormality
Active disease

Anomalies
Anomaly
Associated disease
Clinical findings
Complication
Complications
Condition
Conditions
Course disease
Critically ill
Damage
Defect
Defects
Deficiencies
Deficiency
Deficienty
Deficit
Deficits
Deformity
Degeneration
Disease
Disease transmission
Diseases
Disorder
Disorders
Distortion
Dysfunction
Emergencies
Finding
Findings
Health problems
Ill
Illness
Impairment
Lesions
Malformations
Pathogenesis
Pathology
Pathologies
Progression
Proliferation
Rare condition
Recurrent disease
Relapse

Signs and symptoms
Symptom
Symptoms
Syndrome
Syndromes
Toxicity
Transmission

<u>Other</u> Overly Generic Terms (to be filled incrementally by annotators)

Chromosome
Cure
Device
Discharge
Drug
Enzyme
Exon
Family
Family member
Gene
Organization
People
Person
Problem
Procedure
Profession
Protein
Poison
Side effect
Substance
Test
Tissue
Therapy
Treatment

<u>Other</u> Not Overly Generic Terms (to be filled incrementally by annotators)

Active
Contagious
Fraction
Friend
Genetic disease
Inactive
Infection

Injection
Itch
Mutation
Operation
Pain
Sensitive
Shot
Surgery
Thin
Tired
Vaccine
Weak

UMLS Semantic Types

Acquired Abnormality
Activity
Age Group
Amino Acid Sequence
Amino Acid, Peptide, or Protein
Amphibian
Anatomical Abnormality
Anatomical Structure
Animal
Antibiotic
Archaeon
Bacterium
Behavior
Biologic Function
Biologically Active Substance
Biomedical Occupation or Discipline
Biomedical or Dental Material
Bird
Body Location or Region
Body Part, Organ, or Organ Component
Body Space or Junction
Body Substance
Body System
Carbohydrate
Carbohydrate Sequence
Cell
Cell Component
Cell Function
Cell or Molecular Dysfunction
Chemical
Chemical Viewed Functionally
Chemical Viewed Structurally
Classification
Clinical Attribute
Clinical Drug
Conceptual Entity
Congenital Abnormality
Daily or Recreational Activity
Diagnostic Procedure
Disease or Syndrome
Drug Delivery Device
Educational Activity
Eicosanoid
Element, Ion, or Isotope
Embryonic Structure
Entity
Environmental Effect of Humans
Enzyme
Eukaryote
Event
Experimental Model of Disease
Family Group
Finding
Fish
Food
Fully Formed Anatomical Structure
Functional Concept
Fungus
Gene or Genome
Genetic Function
Geographic Area
Governmental or Regulatory Activity
Group
Group Attribute
Hazardous or Poisonous Substance
Health Care Activity
Health Care Related Organization
Hormone
Human
Human-caused Phenomenon or Process
Idea or Concept
Immunologic Factor

Indicator, Reagent, or Diagnostic Aid
Individual Behavior
Injury or Poisoning
Inorganic Chemical
Intellectual Product
Laboratory or Test Result
Laboratory Procedure
Language
Lipid
Machine Activity
Mammal
Manufactured Object
Medical Device
Mental or Behavioral Dysfunction
Mental Process
Molecular Biology Research Technique
Molecular Function
Molecular Sequence
Natural Phenomenon or Process
Neoplastic Process
Neuroreactive Substance or Biogenic Amine
Nucleic Acid, Nucleoside, or Nucleotide
Nucleotide Sequence
Occupation or Discipline
Occupational Activity
Organ or Tissue Function
Organic Chemical
Organism
Organism Attribute
Organism Function
Organization
Organophosphorus Compound
Pathologic Function
Patient or Disabled Group
Pharmacologic Substance
Phenomenon or Process
Physical Object
Physiologic Function
Plant
Population Group
Professional or Occupational Group
Professional Society
Qualitative Concept
Quantitative Concept
Receptor
Regulation or Law
Reptile
Research Activity
Research Device
Self-help or Relief Organization
Sign or Symptom
Social Behavior
Spatial Concept
Steroid
Substance
Temporal Concept
Therapeutic or Preventive Procedure
Tissue
Vertebrate
Virus
Vitamin