CDC's Controlled Health Thesaurus and the UMLS

Edward Bunker, MPH

Mentor: Olivier Bodenreider, MD, PhD

July 28, 2005

Outline

- Introduction and Background
- Motivation and Questions
- Methods for Assessing Coverage
- Findings
- Limitations
- Future Directions
- Discussion

How is *Public Health* represented in the UMLS?

"...the use of controlled vocabulary in health care and public health systems is likely to increase the quality, effectiveness, and efficiency of health care and to facilitate clinical research, public health surveillance, and health services research."

(Humphreys, et al, 1997)

“The Controlled Health Thesaurus (CHT) is a public health view of pertinent concepts from the National Library of Medicine’s Metathesaurus. Additional concepts needed to cover the public health domain are being added and will be advanced to the NLM.”

(CHT Brochure, CDC)

Controlled Health Thesaurus

- Specifically developed to facilitate tagging of content on CDC's web site
- Has its roots in MeSH, CRISP, AOD…SNOMED-CT now being considered
- As of May 2005 contains 42,639 terms
- Part of a suite of vocabularies being promulgated by the Public Health Information Network (PHIN) at the CDC
- Kevric, Inc. is the CDC contractor on the CHT
- Apelon tools used for development

Sample Terms

1,2-Benzoquinones
Abdominal region
Absenteeism
AC protocol
Accident prevention
Accommodation and Food Services
Actinobacillosis
Acute Renal Failure
AIDS Public Information Data Set - Software
Aiken (County)
Air-purifying respirator
Allergen
Alligators and Crocodiles
American Samoa
Amusement and Theme Parks
Anterior Neck Pain
Aphasia, Graphomotor
Aspartic Acid, Calcium Salt

Sample Data Elements in CHT

Term	UMLS CUI	Parent
Adrenalin	C0014563	Epinephrine
Health status	C0018759	Demographic
Italian	NOT ASSIGNED	European
NIOSH Alert	NOT ASSIGNED	Publication

General features of the Controlled Health Thesaurus

- UMLS-based vocabulary
- “Living” and “co-evolving” vocabulary
- Specialty vocabulary
- “Publicly available” vocabulary

First Area of Inquiry

- Why do certain CHT terms not have an assigned UMLS CUI when some of them seem to relate to fairly common biomedical concepts?
  - Actual conceptual differences between the terms as they reside hierarchically in the CHT and UMLS?
  - Specificity?
  - Errors in concept assignment by the vocabulary developers?
  - Licensing/Proprietary considerations?

Second Area of Inquiry

- As the UMLS and other UMLS-*based* vocabularies evolve, what sorts of processes and tools might be useful for assessing coverage of “new” terms?
  - How can coverage and mapping information be fed back to vocabulary developers and domain experts?

First things first.

Methods & Procedures

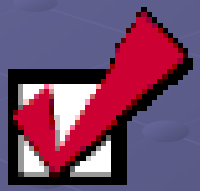
1. Identified and selected terms in CHT without UMLS CUIs

2. Mapped CHT terms to the UMLS using a normalized string index

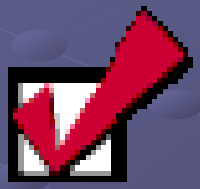
3. Performed semantic validation
   - Hierarchical mapping
   - Ancestor identification

Step #1: Identification of Terms

- ~~Anosmia (C0001326)~~



- Used Car Salesman (No CUI)



- Alpha radiation (No CUI)

- Italian (No CUI)

Step #2: UMLS Concept Mapping

- ~~Used Car Dealers [NO UMLS MATCHES]~~

- Alpha radiation → Alpha Particles (C0002217)

- "Italian"

  → Italian language (C0022275)

  → Italians (C0337810)

Step #3a: Hierarchical Mapping

(Unique Match Example)

*CHT Lineage for "*Alpha radiation*"*

CDC Controlled Health Thesaurus (Root)

>Processes and phenomena

>Physical phenomenon

>Ionizing radiation

> Alpha radiation

Step 3b: Ancestor Identification

(Unique Match Example)

CDC Controlled Health Thesaurus (Root)

>Processes and phenomena

>Physical phenomenon

>Ionizing radiation (C0034538)

>Alpha radiation

?

Alpha Particles (C0002217)

Ancestors of C0002217

(Searching for Ionizing radiation “C0034538”)

C0002217|C0013878;C0028585;C0031816;C0031837;C0034538;C0036397;
C0080022;C0085772;C0336529;C0336996;C0337029;C0338065;C0338370;
C0347997;C0439062;C0439861;C0449234;C0541459;C0542479;C0563221;
C0567414;C0586397;C0596702;C0597237;C0678530;C0678531;C0681949;
C0681951;C0729601;C0729759;C0851346;C0935479;C0935523;C1135584;
C1140093;C1140116;C1140118;C1140124;C1140129;C1140162;C1254345;
C1256739;C1256741;C1275493;C1285164

Step 3b: Ancestor Identification

(Unique Match Example)

CDC Controlled Health Thesaurus

>Processes and phenomena

>Physical phenomenon

>Ionizing radiation (C0034538)

>Alpha radiation

YES

Alpha Particles (C0002217)

Step #3a: Hierarchical Mapping

(Multiple Match Example)

CHT Lineage for "Italian"

Population Group By Race

>White

>European

>Italian

Step #3b: Ancestor Identification

(Multiple Match Example)

Mapping "Italian" as *"Italian Language"*

Population Group By Race
>White (C0043157)
>European
>Italian: Candidate "Italian Language" (C0022275)

NO

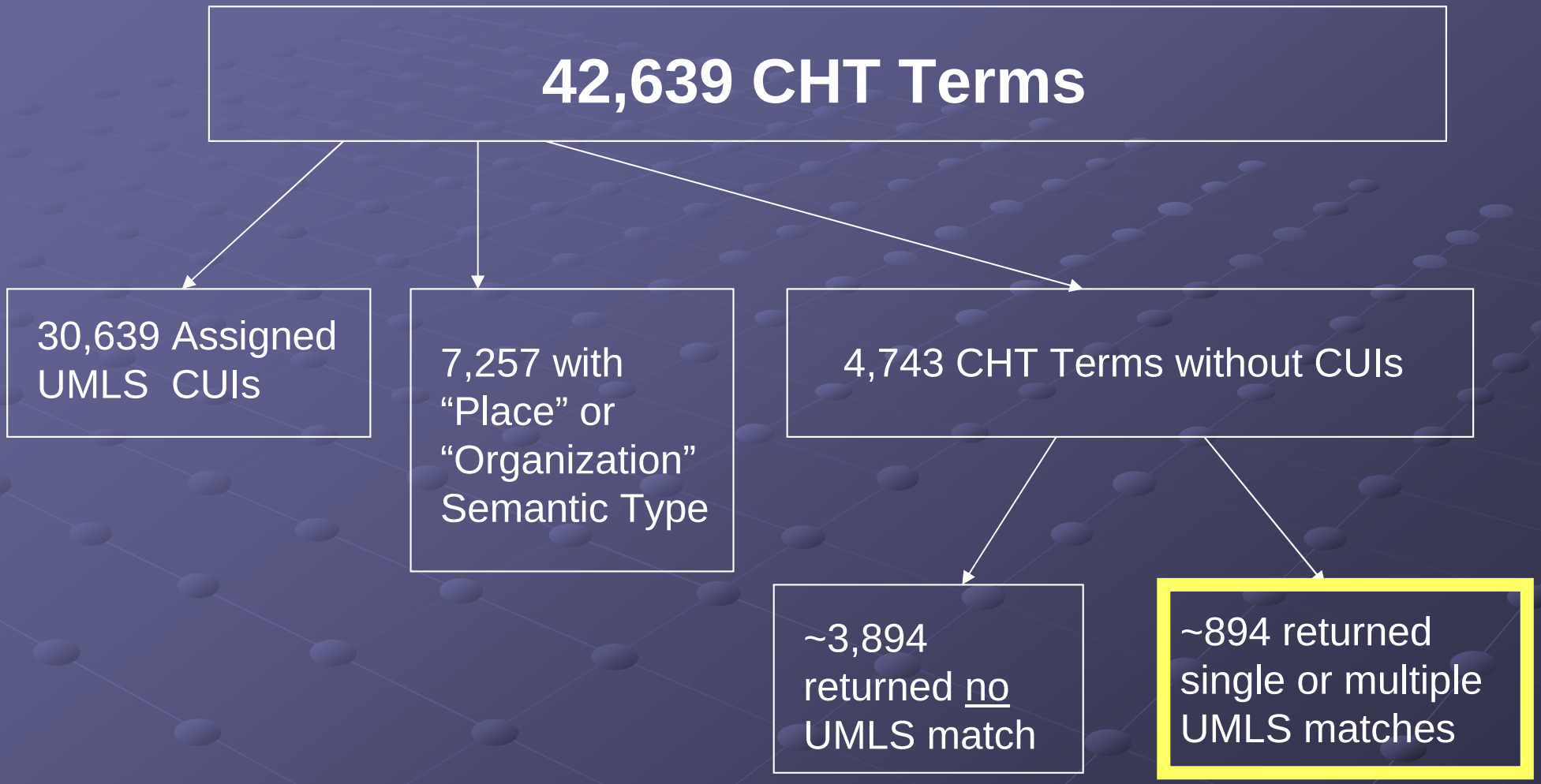
Mapping "Italian" as *"Italians"*

Population Group By Race
>White (C0043157)
>European
>Italian: Candidate "Italians" (C0337810)

YES

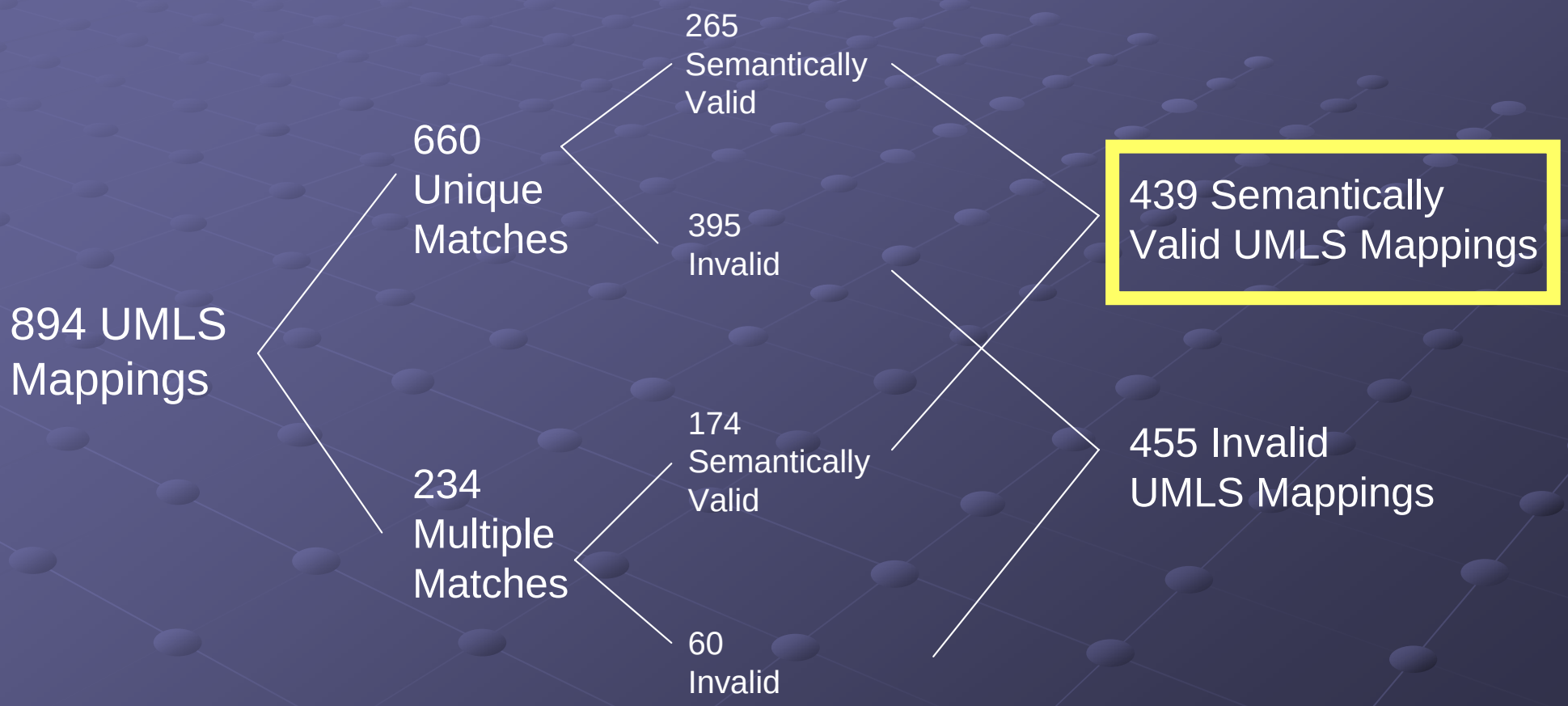
Now *that's* Italian!

Overall Findings

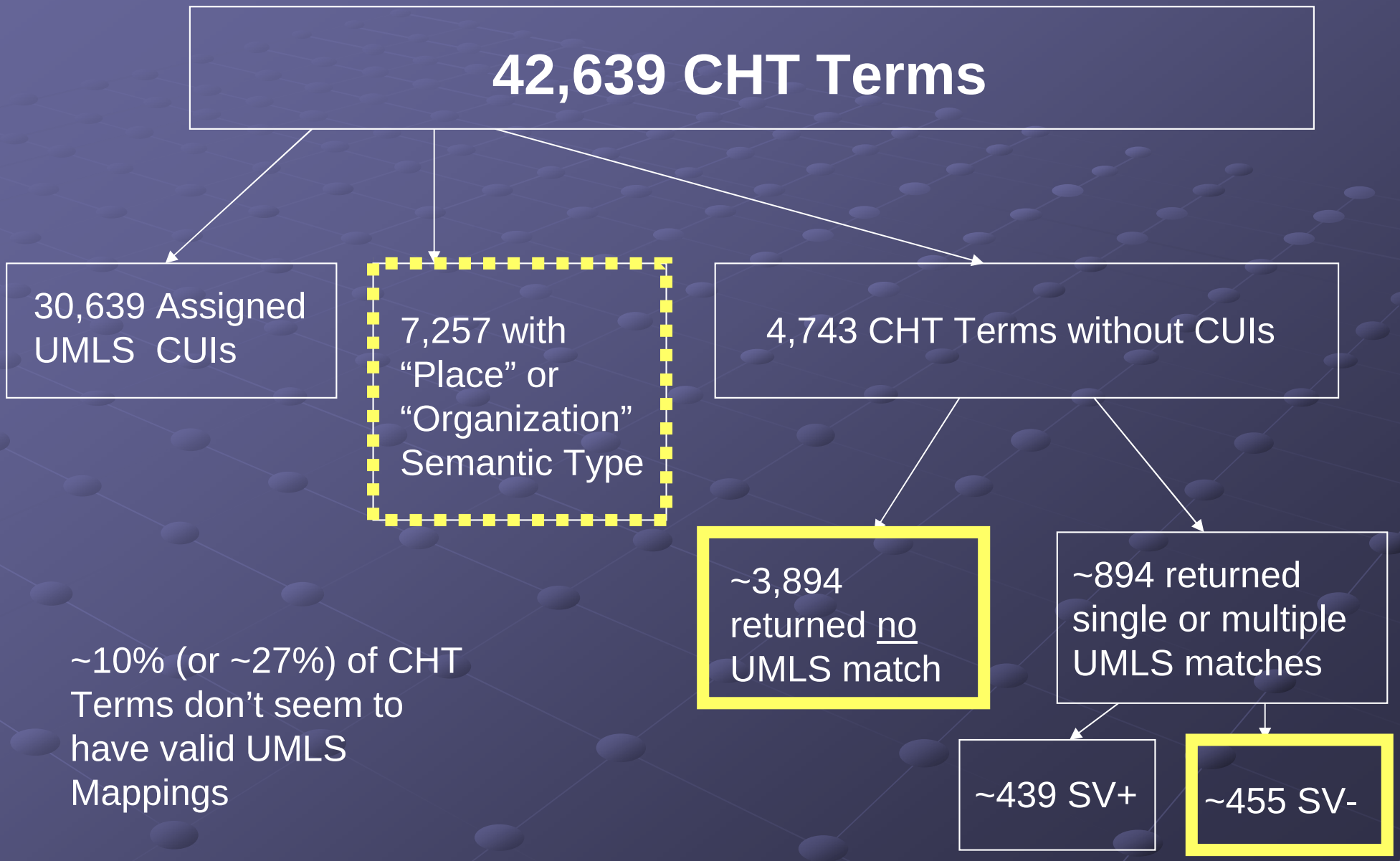


Specific Findings

(Estimates due to a slight bug in the works)



Overall Findings: Revisited



Of the ~439 Semantically Valid UMLS Matches…

- ~ 34% were covered by standard MeSH
- ~ 75% were covered by SNOMED-CT
- ~90% were covered by some combination of MeSH, SNOMED-CT, CRISP, Library of Congress, MDR, NCBI or MTH

Limitations

- Ancestry mapping was done with 2004 data, while normalized string mapping was done with 2005 data
- Data handling errors resulted in ~150 terms not being mapped

Future Directions

- Examination of those terms that didn’t map to (or didn’t validly map to) a UMLS CUI
- Validation of existing CUI assignments
- Evaluation of whether our findings will be of practical use to Kevric and the CDC, i.e., Can our report be used for manual curation?

Discussion

- Normalized string matching was found to be a useful method for our purposes because it increased coverage.
- Formatted reports were helpful for manually reviewing and validating proposed CUI assignments.

Conclusions

Producing “resynchronization” reports is technically feasible and leverages existing resources of the NLM. Providing periodic feedback to vocabulary developers may be one way to enhance collaboration.

Acknowledgements

- Olivier deserves both thanks and hazard pay.

- Tom provided advice and guidance during the initial phases of this project.

Questions?