Natural Language Processing in the Division of Vital Statistics

APRIL 16, 2019

PATRICK DRAKE

NATIONAL CENTER FOR HEALTH STATISTICS

Presentation Objectives

1. Highlight selected Natural Language Processing (NLP) approaches being used by the National Center for Health Statistics (NCHS)

2. Discuss NLP projects within the Division of Vital Statistics with realworld examples
   - Finding and classifying drug related infant deaths
   - Automating the classification of cause of fetal deaths

Use cases for Natural Language Processing

1. Searching for a topic through large volumes of text
2. Cleaning and homogenizing language prior to analysis
   a) Stemming and lemmetization
   b) Abbreviation handling
   c) Correcting misspellings
3. Learning about the language being used
   a) Finding a word’s synonyms, antonyms
   b) Are deaths from novel drugs appearing in our data ? (Both legal and illicit drug use is of interest here)
4. Assigning cause of death to death certificates

Pattern Matching

Terms

- Regular expressions – a sequence of characters that define a search pattern
- Tokenization - preprocessing step where text is segmented into plausible units (i.e., tokens).
- Token – can be words, acronyms, abbreviations, numbers, punctuation symbols, etc.

Challenges

Abbreviations (MD = doctor, state?), apostrophes, hypens, varying formats (e.g., acetyl-fentanyl, acetyl fentanyl, acetylfentanyl), varying boundary demarcations (e.g., The oil prices fell in the U.S.).

Regex: Dealing with Abbreviations

Replacing abbreviations in text with their meaning during data cleaning and processing can improve the performance of any text analysis or algorithm

Medical data (death certificates, and health records) in particular contains a wide variety of abbreviations:

- Diseases and syndromes (e.g. CM = Chiari malformation, dm = diabetes mellitus, . . .)
- Short hand (e.g. fx = fracture, hb = hemoglobin, . . .)

Abbreviation handling:

Input text	Pattern	Quality	Output text
Gestational dm	“dm”	Worst	Gestational diabetes mellitus
NAS	“[Nn][Aa][Ss]”	Better	Neonatal abstinence syndrome
h.s.v	“[Hh][-.]*[Ss][-.]*[Vv]”	Best	Herpes simplex virus

Special Characters for use with *Regular Expressions and their meaning*: (in R)

Quantifiers:

- * match at least 0 times
- \+ match at least 1 times
- ? Match at most 1 time
- {n} match n times

Specifying position:

- ^ match at start of string
- $ match at end of string
- \b “word boundary” matches at end/beginning of word

Spell-checking literal text fields

Spelling errors are common in text describing health conditions, medical jargon, and descriptions of deaths.

Without handling errors in some way, a model will treat different spellings of a word as entirely unrelated.

Example:

Does "*gestation iabetes and placental abrupton*"

equal

"*gestational diabetes and placental abruption*"?

Spell-checking literal text fields

A good spell checker has three main components:

1. Dictionary
2. A method of measuring the “distance” between two strings
3. Language model or decision rules about which word from the dictionary was misspelled in the text

NOTE: The quality of all three parts corresponds to the overall quality of the spell checker. A bad dictionary, poor choice of distance metric, or an improper language model will cause poor results even if the other elements are well implemented.

Spell-checking literal text fields

"This sentence contains a *misspleling*."

Words/tokens	**In dictionary?**
This	True
sentence	True
contains	True
a	True
Misspleling	False

Similar words in dictionary
Misspelling
Misspellings
Misspelled
Mrs. Speiling
. . .

Replace "misspleling" with the most sensible similar word

"This sentence contains a *misspelling*."

Determining word associations (e.g. finding novel drugs)

Basic steps:

1. Use regular expressions to match known words of interest
2. Define a context within which to consider each word
   - N-grams (similar to ‘neighborhood’ in real analysis)
   - Bag of words
   - Punctuation based (e.g. Which words were used in the same sentence)
3. Find other occurrences of contexts of interest:
   - Synonyms/antonyms – words that appear in similar contexts
   - Modifiers/adjectives – words that commonly appear around a word of interest are typically describing a characteristic of that word

Demonstration: An automated approach for classifying cause of fetal death

NLP FOR AUTOMATED ICD-10 CODING ASSIGNMENT

An automated approach for classifying cause of fetal death

Background

- NCHS provides cause of death coding for all death records in the United States including fetal deaths
- This predominantly manual approach takes time and resources to complete
- Upon receipt by NCHS, cause of death coding for fetal deaths ~~s~~ can take years to complete

Objective: To create an automatic rule-based procedure for assignment of multiple cause ICD-10 codes for fetal death records at the national level

- Automating the classification of cause of death for fetal death records would provide an immediate benefit to research and surveillance efforts.

Data Source and Software

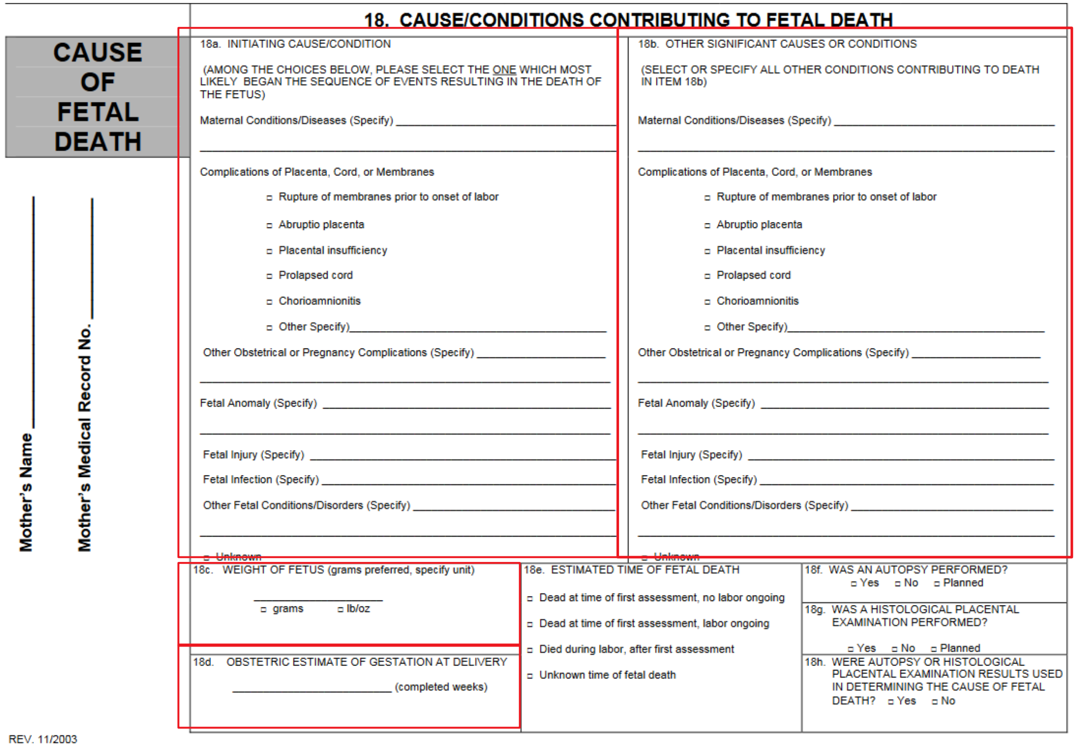
Data Source: 2014 – 2015 Fetal death reports

Literal text refers to the information written by the death certifier on the death report/certificate:

- Maternal Conditions/Diseases
- Complications of placenta, cord, or membranes
- Fetal Anomalies
- Injuries
- Infections
- Other field

Non-Literal information from the record includes:

- Weight of the fetus
- Plurality
- Length of gestation
- Sex



CAUSE OF FETAL DEATH

Mother's Name

Mother's Medical Record No.

18. CAUSE/CONDITIONS CONTRIBUTING TO FETAL DEATH

18a. INITIATING CAUSE/CONDITION

(AMONG THE CHOICES BELOW, PLEASE SELECT THE ONE WHICH MOST LIKELY BEGAN THE SEQUENCE OF EVENTS RESULTING IN THE DEATH OF THE FETUS)

Maternal Conditions/Diseases (Specify)

Complications of Placenta, Cord, or Membranes

- Rupture of membranes prior to onset of labor
- Abruptio placenta
- Placental insufficiency
- Prolapsed cord
- Chorioamnionitis
- Other Specify)

Other Obstetrical or Pregnancy Complications (Specify)

Fetal Anomaly (Specify)

Fetal Injury (Specify)

Fetal Infection (Specify)

Other Fetal Conditions/Disorders (Specify)

- Unknown

18b. OTHER SIGNIFICANT CAUSES OR CONDITIONS

(SELECT OR SPECIFY ALL OTHER CONDITIONS CONTRIBUTING TO DEATH IN ITEM 18b)

Maternal Conditions/Diseases (Specify)

Complications of Placenta, Cord, or Membranes

- Rupture of membranes prior to onset of labor
- Abruptio placenta
- Placental insufficiency
- Prolapsed cord
- Chorioamnionitis
- Other Specify)

Other Obstetrical or Pregnancy Complications (Specify)

Fetal Anomaly (Specify)

Fetal Injury (Specify)

Fetal Infection (Specify)

Other Fetal Conditions/Disorders (Specify)

- Unknown

18c. WEIGHT OF FETUS (grams preferred, specify unit)

- grams
- lb/oz

18d. OBSTETRIC ESTIMATE OF GESTATION AT DELIVERY

(completed weeks)

18e. ESTIMATED TIME OF FETAL DEATH

- Dead at time of first assessment, no labor ongoing
- Dead at time of first assessment, labor ongoing
- Died during labor, after first assessment
- Unknown time of fetal death

18f. WAS AN AUTOPSY PERFORMED?
- Yes
- No
- Planned

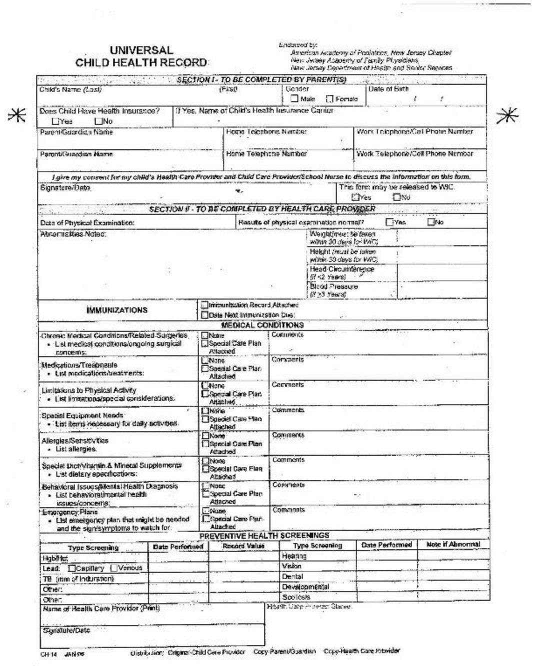
18g. WAS A HISTOLOGICAL PLACENTAL EXAMINATION PERFORMED?
- Yes
- No
- Planned

18h. WERE AUTOPSY OR HISTOLOGICAL PLACENTAL EXAMINATION RESULTS USED IN DETERMINING THE CAUSE OF FETAL DEATH?
- Yes
- No

REV. 11/2003

Software: R statistical language, focused on using base scripting language without additional resources

Project’s data flow



UNIVERSAL
CHILD HEALTH RECORD

Endorsed by:
American Academy of Pediatrics, New Jersey Chapter
New Jersey Academy of Family Physicians
New Jersey Department of Health and Senior Services

SECTION I - TO BE COMPLETED BY PARENT(S)

Child's Name (Last)	(First)	Gender ☐ Male ☐ Female	Date of Birth / /
Does Child Have Health Insurance? ☐Yes ☐No	If Yes, Name of Child's Health Insurance Carrier		
Parent/Guardian Name	Home Telephone Number	Work Telephone/Cell Phone Number	
Parent/Guardian Name	Home Telephone Number	Work Telephone/Cell Phone Number	

I give my consent for my child's Health Care Provider and Child Care Provider/School Nurse to discuss the information on this form.

Signature/Date	This form may be released to WIC. ☐Yes ☐No

SECTION II - TO BE COMPLETED BY HEALTH CARE PROVIDER

Date of Physical Examination:	Results of physical examination normal? ☐Yes ☐No

Abnormalities Noted:	Weight (must be taken within 30 days for WIC)	
	Height (must be taken within 30 days for WIC)	
	Head Circumference (if <2 Years)	
	Blood Pressure (if >3 Years)	

IMMUNIZATIONS	☐Immunization Record Attached ☐Date Next Immunization Due:

MEDICAL CONDITIONS

Chronic Medical Conditions/Related Surgeries • List medical conditions/ongoing surgical concerns:	☐None ☐Special Care Plan Attached	Comments
Medications/Treatments • List medications/treatments:	☐None ☐Special Care Plan Attached	Comments
Limitations to Physical Activity • List limitations/special considerations:	☐None ☐Special Care Plan Attached	Comments
Special Equipment Needs • List items necessary for daily activities:	☐None ☐Special Care Plan Attached	Comments
Allergies/Sensitivities • List allergies:	☐None ☐Special Care Plan Attached	Comments
Special Diet/Vitamin & Mineral Supplements • List dietary specifications:	☐None ☐Special Care Plan Attached	Comments
Behavioral Issues/Mental Health Diagnosis • List behavioral/mental health issues/concerns:	☐None ☐Special Care Plan Attached	Comments
Emergency Plans • List emergency plan that might be needed and the sign/symptoms to watch for:	☐None ☐Special Care Plan Attached	Comments

PREVENTIVE HEALTH SCREENINGS

Type Screening	Date Performed	Record Value	Type Screening	Date Performed	Note if Abnormal
Hgb/Hct			Hearing		
Lead: ☐Capillary ☐Venous			Vision		
TB (mm of Induration)			Dental		
Other:			Developmental		
Other:			Scoliosis		

Name of Health Care Provider (Print)	Health Care Provider Stamp
Signature/Date	

CH-14 JAN 98 Distribution: Original-Child Care Provider Copy-Parent/Guardian Copy-Health Care Provider

Spellchecking functionality

1. Dictionary
2. Distance metric
3. Language model

Abbreviation lookup table

Abbreviation 1 -> meaning 1
Abbreviation 2 -> meaning 2
Abbreviation 3 -> meaning 3
. . .

Input text

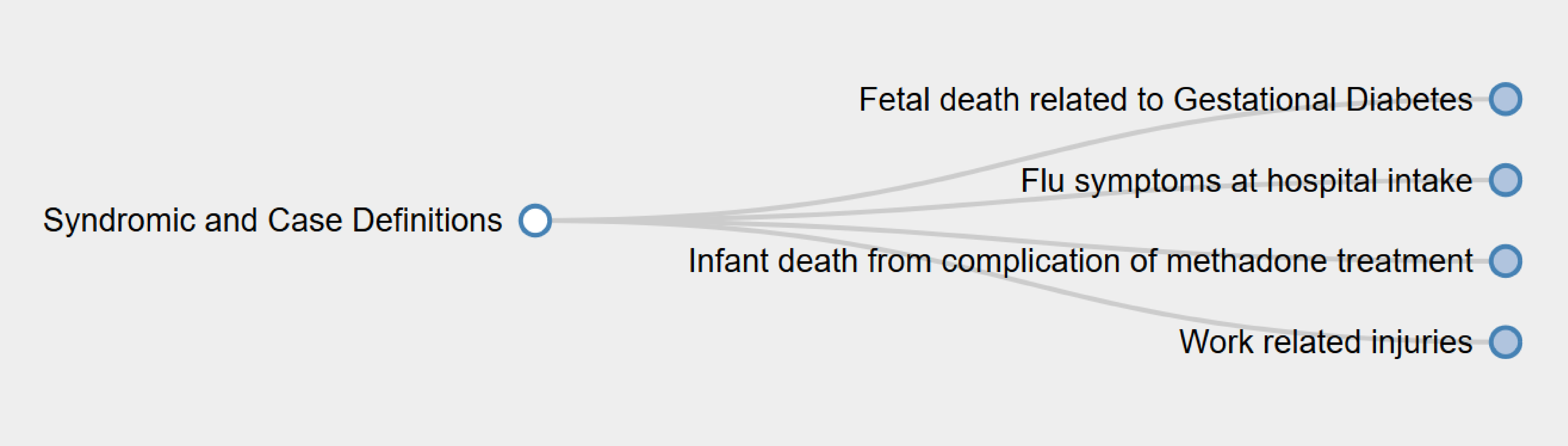
Output text

Classification Algorithm

COD Classification algorithm

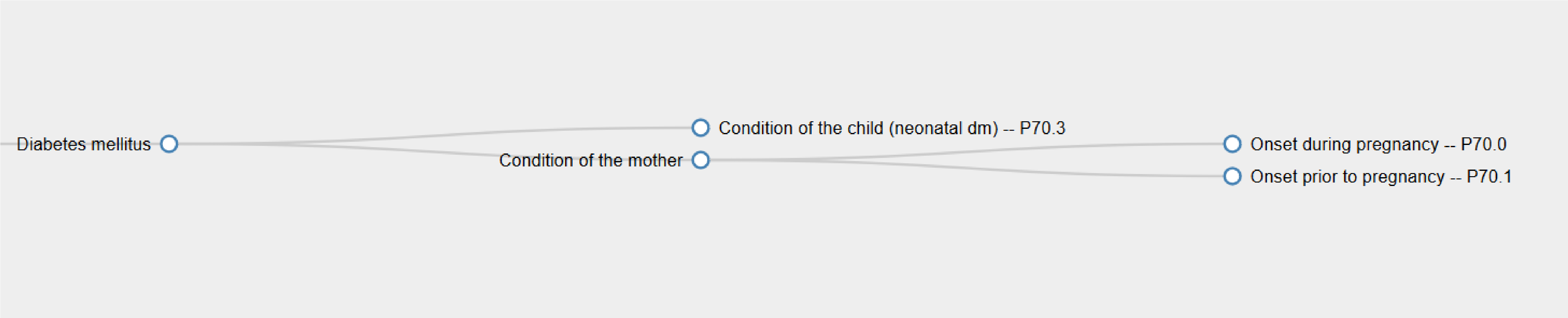
see D3 and Rshiny visualization (or screenshots)

Defining topics for use by an algorithm



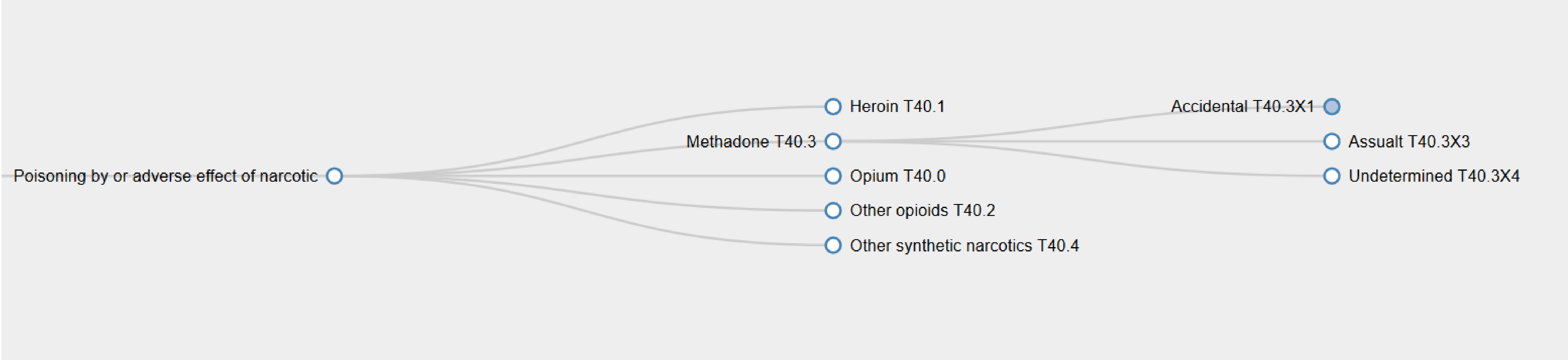
Defining a topic:

Fetal death associated with gestational diabetes of the mother



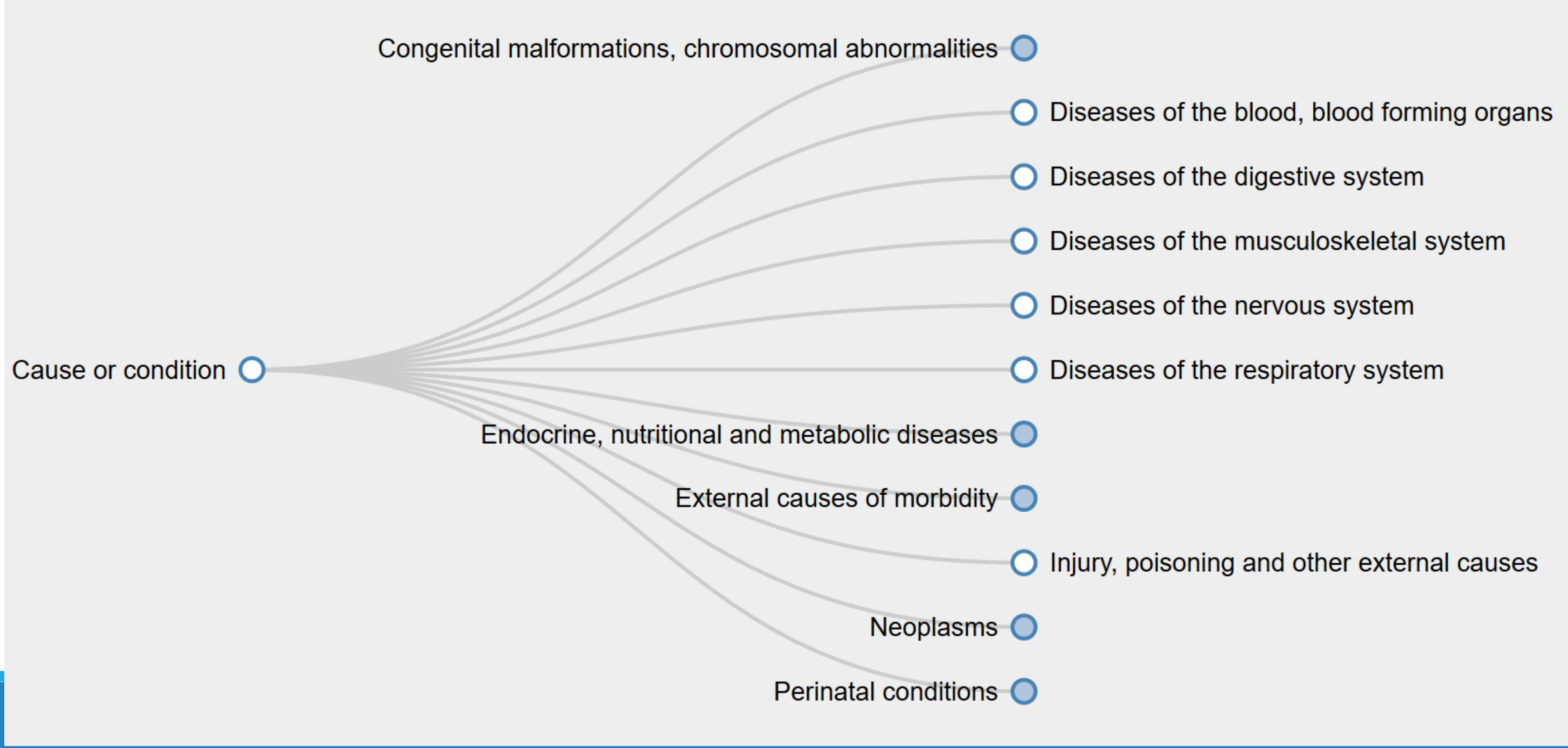
Defining a topic:

Fetal death associated with complications of methadone therapy



Representing all causes of death:

International Classification of Diseases



Text mining the coding manual to define all ICD codes

SECTION I - Alphabetical index to diseases and nature of injury

A

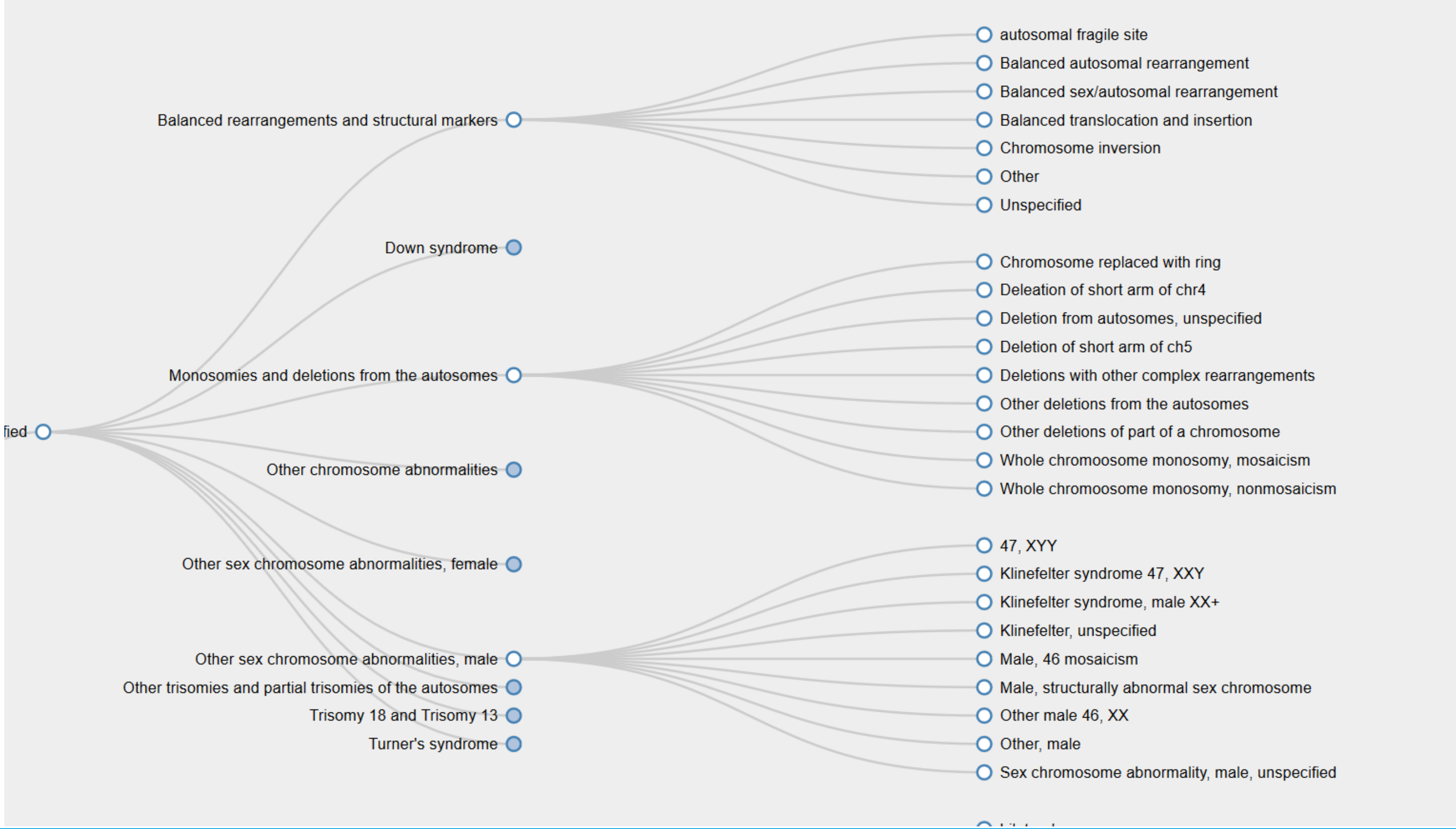
Aarskog's syndrome Q87.1 2a
Abandonment T74.0
Abasia(-astasia) (hysterical) F44.4
Abdomen, abdominal - *see also condition*
- acute R10.0 2b
- convulsive equivalent G40.8
- muscle deficiency syndrome Q79.4
Abdominalgia R10.4 2b
- periodic E85.9 2b
Abduction contracture, hip or other joint - *see* Contraction, joint
Aberrant (congenital) - *see also* Malposition, congenital
- adrenal gland Q89.1
- artery (peripheral) NEC Q27.8 2a 2b
- breast Q83.8
- endocrine gland NEC Q89.2
- hepatic duct Q44.5
- pancreas Q45.3
- parathyroid gland Q89.2
- pituitary gland Q89.2
- sebaceous glands, mucous membrane, mouth, congenital Q38.6
- spleen Q89.0
- subclavian artery Q27.8 2a 2b
- thymus (gland) Q89.2
- thyroid gland Q89.2
- vein (peripheral) NEC Q27.8 2a 2b
Aberration, mental F99 2a 2b
Abetalipoproteinemia E78.6
Abiotrophy R68.8 2a 2b
Ablatio, ablation
- pituitary (gland) E23.0 2b
- placentae (*see also* Abruptio placentae) O45.9 2b
- - affecting fetus or newborn P02.1 2a
- retinae (*see also* Detachment, retina) H33.2
- uterus 'Z90.7
Ablepharia, ablepharon Q10.3
Abnormal, abnormality, abnormalities - *see also* Anomaly
- acid-base balance (mixed) E87.4
- - due to

O

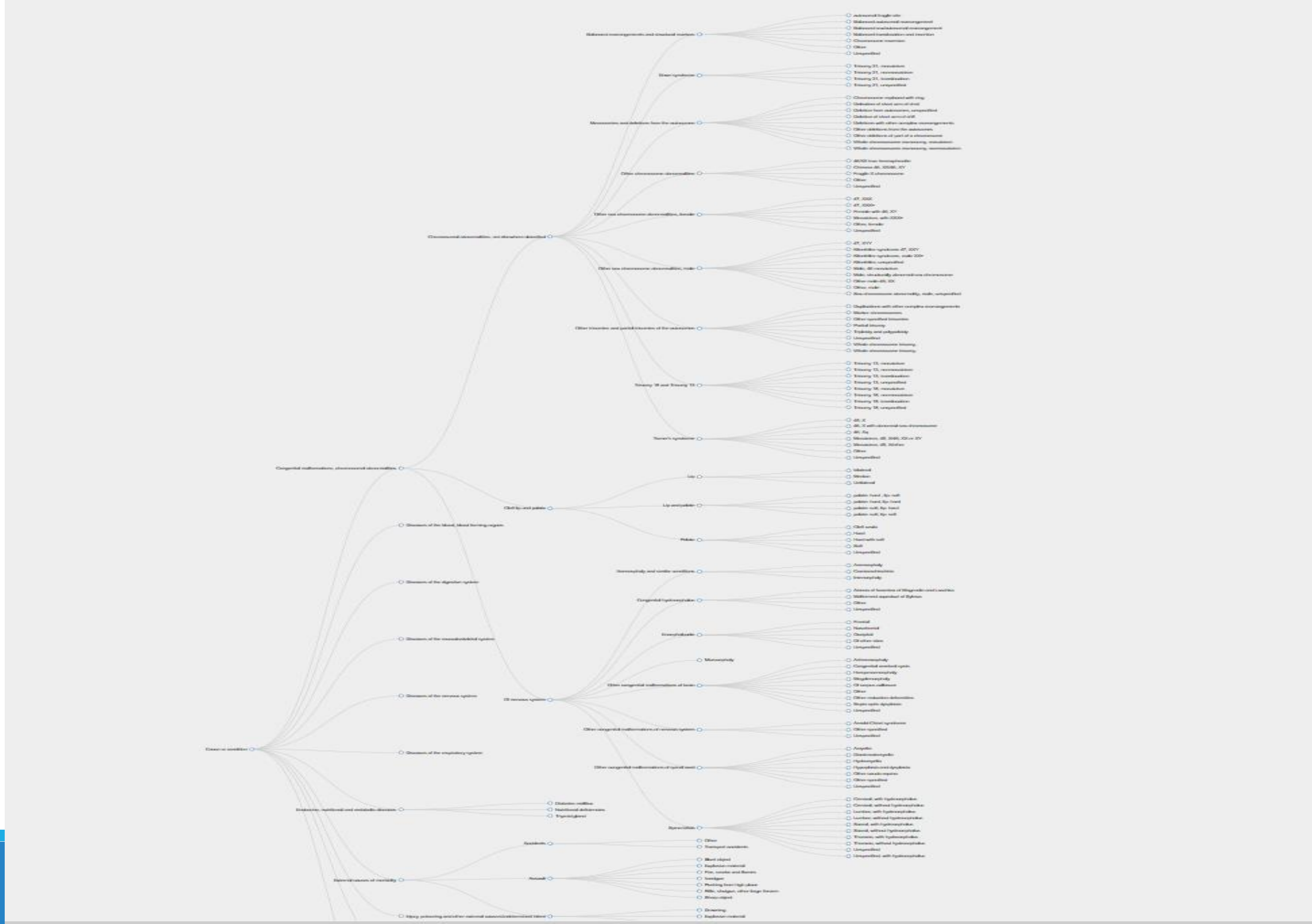
Substance	Chapter XIX	Accidental	Intentional self-harm	Undetermined intent	Adverse effect in therapeutic use
			Poisoning		
Obidoxime chloride		T50.6	X44-	X64-	Y14-
Octafonium (chloride)		T49.3	X44-	X64-	Y14-
Octamethyl pyrophosphoramide		T60.0 2b	X48-	X68-	Y18-
Octanoin		T50.9 2a 2b	X44-	X64-	Y14-
Octatropine methylbromide		T44.3	X43-	X63-	Y13-
Octotiamine		T45.2	X44-	X64-	Y14-
Octoxinol (9)		T49.8	X44-	X64-	Y14-
Octreotide		T38.9	X44-	X64-	Y14-
Oestradiol		T38.5	X44-	X64-	Y14-
Oestriol		T38.5	X44-	X64-	Y14-
Oestrogen		T38.5	X44-	X64-	Y14-
Oestrone		T38.5	X44-	X64-	Y14-
Ofloxacin		T36.8	X44-	X64-	Y14-
Oil (of)					
- bitter almond		T62.8	X49-	X69-	Y19-
- cloves		T49.7	X44-	X64-	Y14-
- colors		T65.6	X49-	X69-	Y19-
- fumes		T59.8 2a 2b	X47-	X67-	Y17-
- lubricating		T52.0	X46-	X66-	Y16-
- Niobe		T52.8 2b	X46-	X66-	Y16-
- vitriol (liquid)		T54.2	X49-	X69-	Y19-
- - fumes		T54.2	X47-	X67-	Y17-
Oily preparation (for skin)		T49.3	X44-	X64-	Y14-
Ointment NEC		T49.3	X44-	X64-	Y14-
Olanzapine		T43.5 2b	X41-	X61-	Y11-
Oleander		T62.2	X49-	X69-	Y19-
Oleandomycin		T36.3	X44-	X64-	Y14-
Oleandrin		T46.0 2a 2b	X44-	X64-	Y14-
Oleic acid		T46.6	X44-	X64-	Y14-
Oleovitamin A		T45.2	X44-	X64-	Y14-
Oleum ricini		T47.2	X44-	X64-	Y14-
Olivomycin		T45.1	X44-	X64-	Y14-
Olsalazine		T47.8	X44-	X64-	Y14-
Omeprazole		T47.1	X44-	X64-	Y14-
OMPA		T60.0 2b	X48-	X68-	Y18-
Ondansetron		T45.0 2b	X44-	X64-	Y14-
Ophthalmological drug NEC		T49.5	X44-	X64-	Y14-

Poisoning Adverse

International Classification of Diseases (zoomed in)

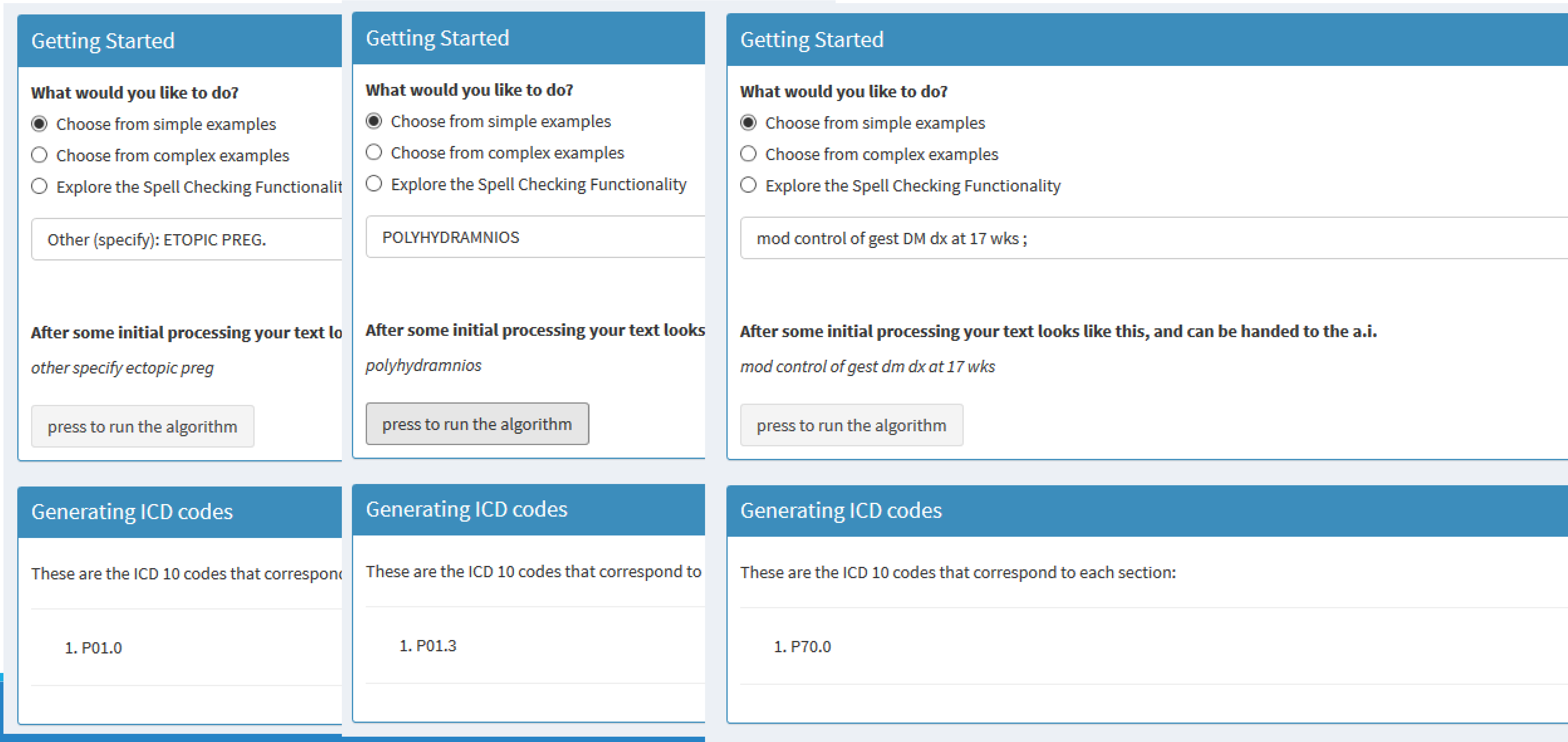


International Classification of Diseases (high level view)



Shown here is just one subsection of one chapter of the ICD-10

shinyApp screenshots: *Fetal COD Classification algorithm*



Getting Started

What would you like to do?

- ( ) Choose from simple examples
- (•) Choose from complex examples
- ( ) Explore the Spell Checking Functionality

ectopic pregnancy polyhydramnios true knot in cord

After some initial processing your text looks like this, and can be handed to the a.i.

ectopic pregnancy polyhydramnios true knot in cord

press to run the algorithm

Understanding sentence structure

These examples are more complex sentences. The algorithm has broken your choice into the following sections:

1. ectopic pregnancy
2. polyhydramnios
3. true knot in cord

Generating ICD codes

These are the ICD 10 codes that correspond to each section:

1. P01.4
2. P01.3
3. P02.5

Getting Started

What would you like to do?

- ( ) Choose from simple examples
- (•) Choose from complex examples
- ( ) Explore the Spell Checking Functionality

gestaton iabetes and placetal abrupton

After some initial processing your text looks like this, and can be handed to the a.i.

gestational diabetes and placental abruption

press to run the algorithm

Understanding sentence structure

These examples are more complex sentences. The algorithm has broken your choice into the following sections:

1. gestational diabetes
2. placental abruption

Generating ICD codes

These are the ICD 10 codes that correspond to each section:

1. P70.0
2. P02.1

Getting Started

What would you like to do?

- ( ) Choose from simple examples
- (•) Choose from complex examples
- ( ) Explore the Spell Checking Functionality

premature preterm rupture of membranes, placental abruption none none

After some initial processing your text looks like this, and can be handed to the a.i.

premature preterm rupture of membranes placental abruption none none

press to run the algorithm

Understanding sentence structure

These examples are more complex sentences. The algorithm has broken your choice into the following sections:

1. premature preterm rupture of membranes
2. placental abruption

Generating ICD codes

These are the ICD 10 codes that correspond to each section:

1. P01.1
2. P02.1

Questions?

PATRICK DRAKE, STATISTICIAN

DIVISION OF VITAL STATISTICS | 301-458-4848 | NNH9@CDC.GOV