



Concept Prevalence Design diagnostics

Anna Ostropolets MD
Columbia University



Concept Prevalence

Network study: Concept

■ Researchers



aostropelets Anna Ostropelets

We want to announce a new network study on the OHDSI network. You can find the full protocol here: https://github.com/OHDSI/StudyProtocol_v0.1.docx 32

The full protocol can be found here: https://github.com/OHDSI/StudyProtocol_v0.1.docx 32

We want to study the usage patterns of concepts in the OHDSI network. This information is useful to answer many questions, such as: how granularly are codes captured in a network, how can they be coded as ICD9 code 585.2, how often are codes unspecified or even as 586 Renal. Currently, researchers have no way to get this information available for selection, or whether the information to define the cohort or the distribution of the concepts is limited. Studies are dependent on cohort selection.

In an ideal world, a cohort definition would be available to the community. We would like to make this information available to the community.

- Unique values in the *_concept_id
- Unique values in the *_source_id
- Mappings between them

As a side effect, we would also get a better understanding of the dynamics of that distribution over time, and we could draw conclusions about the impact of erroneous mappings.

Investigating Concept Heterogeneity and the OHDSI Network

Study Status Design Finalized

- Analytics use case(s): Characterization
- Study type: Methods Research
- Tags: -
- Study lead: Anna Ostropelets
- Study lead forums tag: [aostropelets](#)
- Study start date: March 15, 2019
- Study end date: -
- Protocol: [Word file](#)
- Publications: -
- Results explorer: -

This study aims to investigate the concept usage in the OHDSI network. It will investigate the cohort identification to facilitate cross-institutional research. In the OHDSI CDM, we will investigate if it is used at each site and how differences in corresponding source code are handled.

Additional information

In R, use the following commands to download and install:

```
install.packages("devtools") devtools::install_github("https://github.com/ohdsi-studies/ConceptPrevalence")
library("ConceptPrevalence")
```

Add inputs for the site:

No patient-
count

Counts of rec

```
select condition_concept_id, count(*), 'condition'
from condition_occurrence
group by condition_concept_id
```

concept_id	cnt	type
201286	12478349	condition
41647883	26477	drug
2083274	623845	procedure



Current results: 22 datasets from six countries



14 US
sources

Hospital
charge data
(3)

Claims
data
(8)

8 International
sources

EHR
(8)

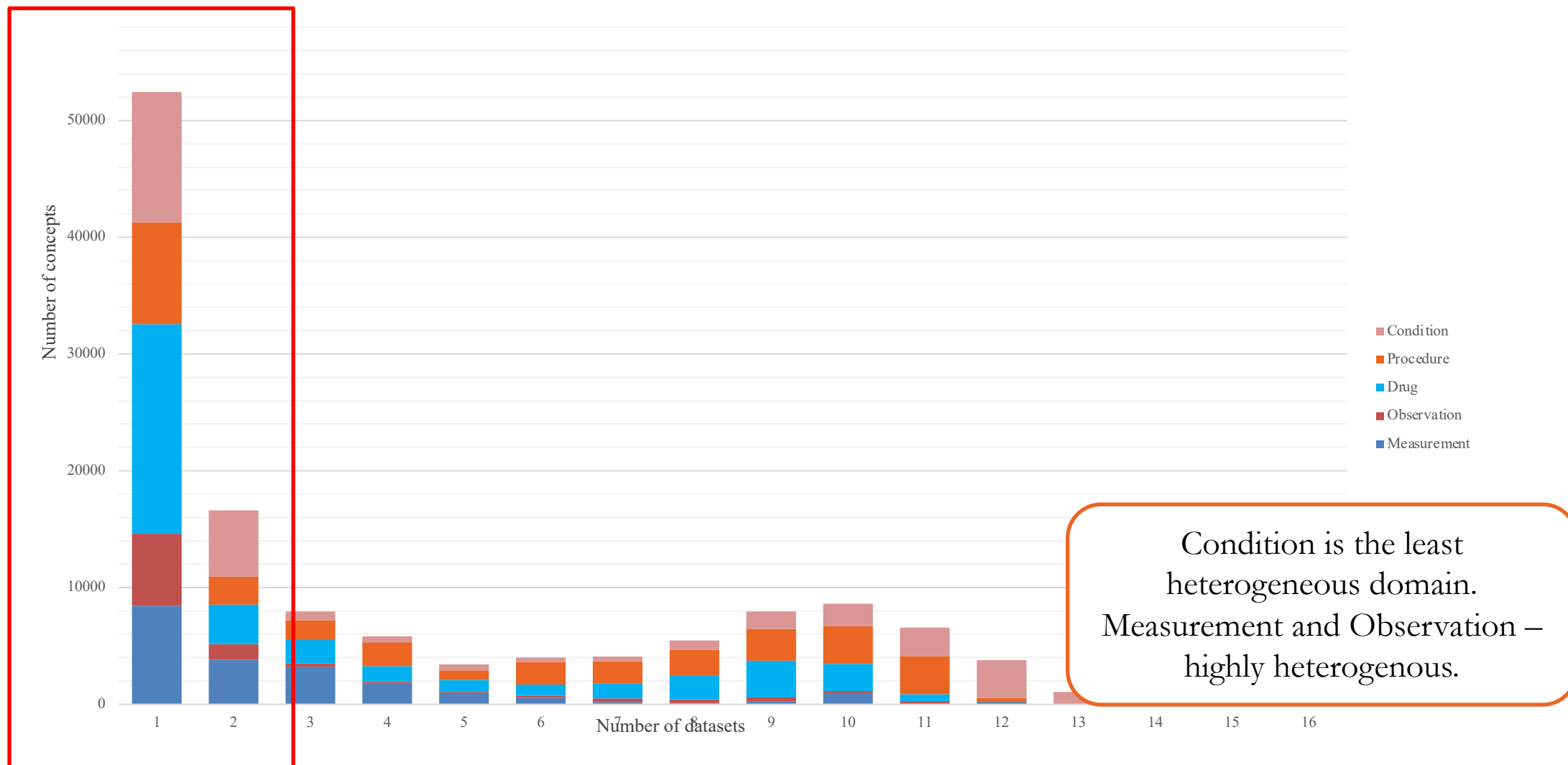
~375 billion
records

14.6%
of all databases within the OHDSI

> 1 million
distinct concepts

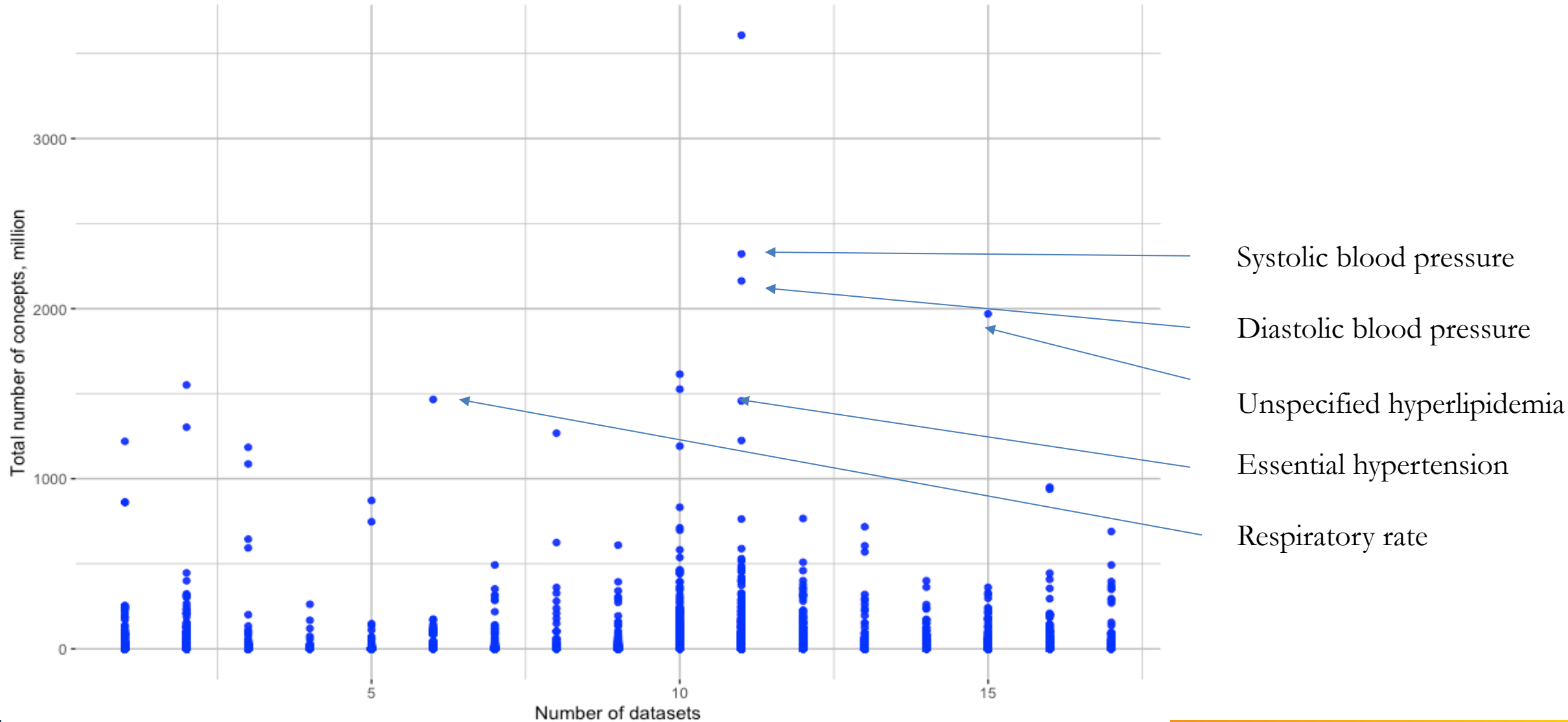


Most of the concepts can be found only in 1 dataset





Even common concepts do not appear in all datasets





Different databases have different granularity

Levels of separation	DA France	JMDC	LPD AU	AmbEMR	CUMC	MDCD	Average
0; Renal impairment			23%	4%	0.01%		2.2%
1; Chronic kidney disease	94%	90%	32%	17%	25%	13%	24.7%
2; Chronic kidney disease stage 5	5%	9%	45%	69%	64%	68%	59.4%
3; Chronic kidney disease stage 5 due to hypertension	0.04%	1%		10%	9%	19%	13.1%
4; Malignant hypertensive chronic kidney disease stage 5	0.04%	0.1%		0.2%	1%	1%	0.5%
5; Malignant hypertensive end stage renal disease on dialysis				0.01%			0.00017%

Database with broad concepts

Database with specific concepts

On average, US sources are more granular than international and US claims – more granular than EHR

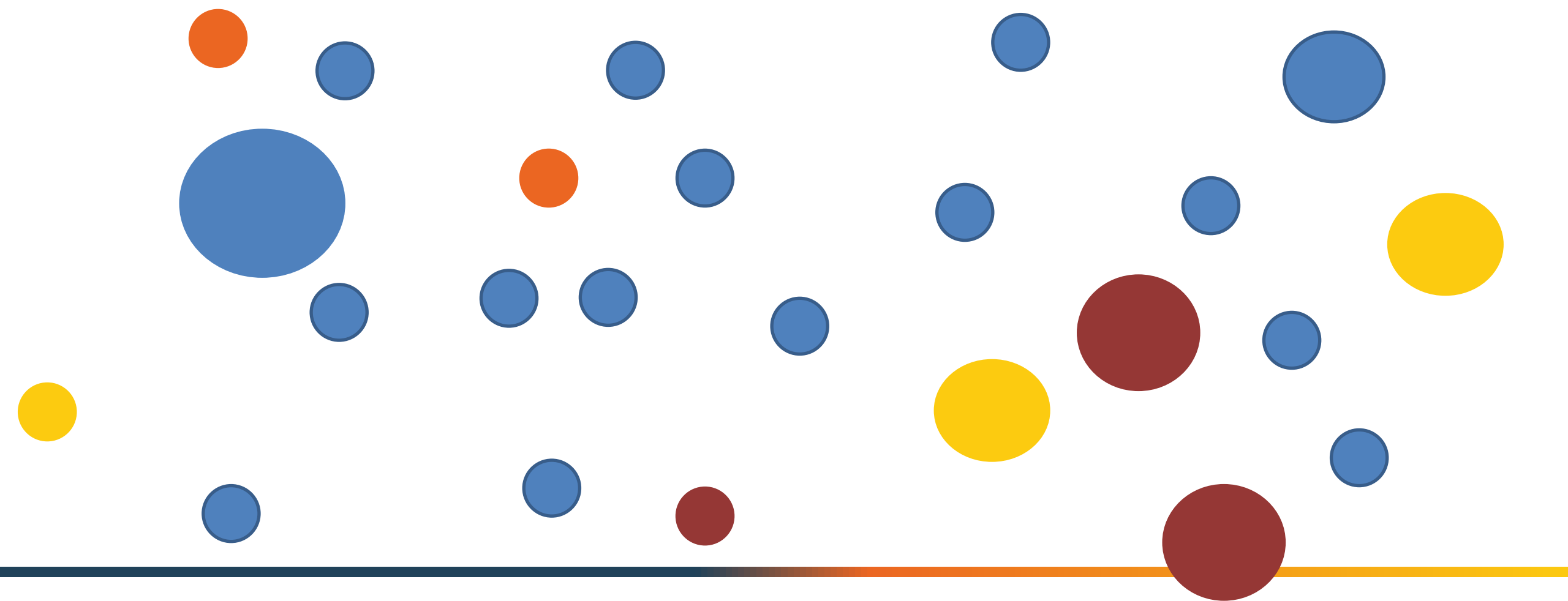


OHDSI's standardized vocabularies


- 153 Vocabularies across 41 domains
 - MU3 standards: SNOMED, RxNorm, LOINC
 - Disparate sources: ICD9CM, ICD10(CM), Read, NDC, Gemscript, CPT4, HCPCS...
- >9 million concepts
 - >3.3 million standard concepts
 - >5.1 million source codes
 - >629,000 classification concepts
- >55 million concept relationships
- >84 million ancestral relationships



Given variety and heterogeneity of the data where do we start to create phenotypes?



Trying to select diabetes

ATHENA

SEARCH BY KEYWORD

"diabetes type 2"

SEARCH

DOWNLOAD

LOGIN

?

"diabetes type 2" x


DOWNLOAD RESULTS

Show by 15 items Total 59 items

1 2 3 4 >

DOMAIN	ID	CODE	NAME	CLASS	CONCEPT	VALIDITY	DOMAIN	VOCAB
CONCEPT	45877606	LA10552-0	Diabetes Type 2	Answer	Standard	Valid	Meas Value	LOINC
CLASS	45465547	66Ao.00	Diabetes type 2 review	Read	Non-standard	Valid	Observation	Read
VOCAB	4341452	N0000000954	Diabetes Mellitus, Type 2	Ind / CI	Non-standard	Valid	Drug	NDFRT
VALIDITY	45611690	D003924	Diabetes Mellitus, Type 2	Main Heading	Non-standard	Valid	Condition	MeSH
	3187674	10656271000119102	Diabetes type 2 with diabetic ulcer of toe, skin breakdown	Clinical Finding	Non-standard	Valid	Condition	Nebraska Lexicon
	4221495	420756003	Cataract due to diabetes mellitus type 2	Clinical Finding	Standard	Valid	Condition	SNOMED
	3329005	44054006	Diabetes mellitus type 2	Clinical Finding	Non-standard	Valid	Condition	Nebraska Lexicon
	46274058	10656271000119102	Skin ulcer of toe due to diabetes mellitus type 2	Clinical Finding	Standard	Valid	Condition	SNOMED
	45757474	1481000119100	Diabetes mellitus type 2 without retinopathy	Clinical Finding	Standard	Valid	Condition	SNOMED
	3338973	359642000	Diabetes mellitus type 2 in nonobese	Clinical Finding	Non-standard	Valid	Condition	Nebraska Lexicon
	45928487	155914	diabetes mellitus type 2 in nonobese	Diagnosis	Non-standard	Valid	Condition	CIEL
	45929270	142472	Diabetes Mellitus Type 2 in Obese	Diagnosis	Non-standard	Valid	Condition	CIEL

Trying to select diabetes

ATHENA

SEARCH BY KEYWORD

"type 2 diabetes mellitus"

SEARCH

DOWNLOAD

LOGIN

?

"type 2 diabetes mel..."

DOWNLOAD RESULTS

Show by 15 items Total 541 items

1 2 3 4 5 ... 37 >

DOMAIN	ID	CODE	NAME	CLASS	CONCEPT	VALIDITY	DOMAIN	VOCAB
CONCEPT	45420114	C109.12	Type 2 diabetes mellitus	Read	Non-standard	Valid	Condition	Read
CLASS	45420119	C10F.00	Type 2 diabetes mellitus	Read	Non-standard	Valid	Condition	Read
VOCAB	45571656	E11	Type 2 diabetes mellitus	ICD10 Hierarchy	Non-standard	Valid	Condition	ICD10
VALIDITY	1567956	E11	Type 2 diabetes mellitus	3-char nonbill code	Non-standard	Valid	Condition	ICD10CM
	1409150	E11	Type 2 diabetes mellitus	ICD10 Hierarchy	Non-standard	Valid	Condition	ICD10CN
	201826	44054006	Type 2 diabetes mellitus	Clinical Finding	Standard	Valid	Condition	SNOMED
	4063043	199230006	Pre-existing type 2 diabetes mellitus	Clinical Finding	Standard	Valid	Condition	SNOMED
	3470656	445353002	Brittle type 2 diabetes mellitus	Clinical Finding	Non-standard	Valid	Condition	Nebraska Lexicon
	3270614	199230006	Pre-existing type 2 diabetes mellitus	Clinical Finding	Non-standard	Valid	Condition	Nebraska Lexicon
	3438725	443694000	Type 2 diabetes mellitus uncontrolled	Clinical Finding	Non-standard	Valid	Condition	Nebraska Lexicon
	37082034	E11	Type 2 diabetes mellitus	ICD10 Hierarchy	Non-standard	Valid	Condition	ICD10GM
	40483315	445353002	Brittle type 2 diabetes mellitus	Clinical Finding	Standard	Valid	Condition	SNOMED



Trying to select diabetes

SEARCH

DOWNLOAD

LOGIN



SEARCH BY KEYWORD

"type II diabetes"



"type II diabetes" ×

DOWNLOAD RESULTS

Show by 15 items

Total 163 items

1

2

3

4

5

...

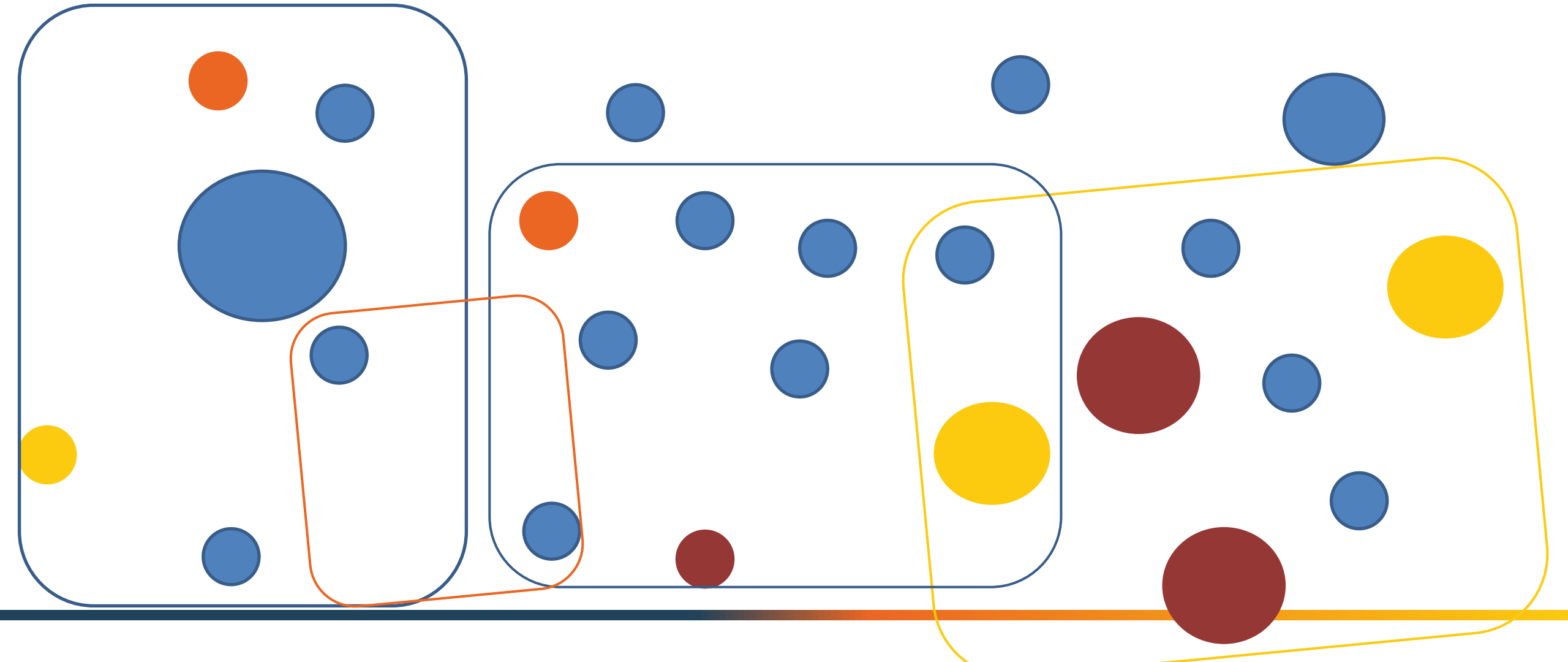
11



DOMAIN	ID	CODE	NAME	CLASS	CONCEPT	VALIDITY	DOMAIN	VOCAB
CONCEPT	45453110	C109.13	Type II diabetes mellitus	Read	Non-standard	Valid	Condition	Read
CLASS	45486691	C10F.11	Type II diabetes mellitus	Read	Non-standard	Valid	Condition	Read
VOCAB	40386778	190384004	Type II diabetes mellitus	Clinical Finding	Non-standard	Invalid	Condition	SNOMED
VALIDITY	3121806	190384004	Type II diabetes mellitus	Clinical Finding	Non-standard	Invalid	Condition	Nebraska Lexicon
	40482801	443694000	Type II diabetes mellitus uncontrolled	Clinical Finding	Standard	Valid	Condition	SNOMED
	45443104	C10FJ11	Insulin treated Type II diabetes mellitus	Read	Non-standard	Valid	Condition	Read
	45473337	C109J12	Insulin treated Type II diabetes mellitus	Read	Non-standard	Valid	Condition	Read
	45503184	C10P100	Type II diabetes mellitus in remission	Read	Non-standard	Valid	Condition	Read
	45463265	C10FG11	Type II diabetes mellitus with arthropathy	Read	Non-standard	Valid	Condition	Read
	45499864	C109G11	Type II diabetes mellitus with arthropathy	Read	Non-standard	Valid	Condition	Read
	45423316	C10F511	Type II diabetes mellitus with gangrene	Read	Non-standard	Valid	Condition	Read

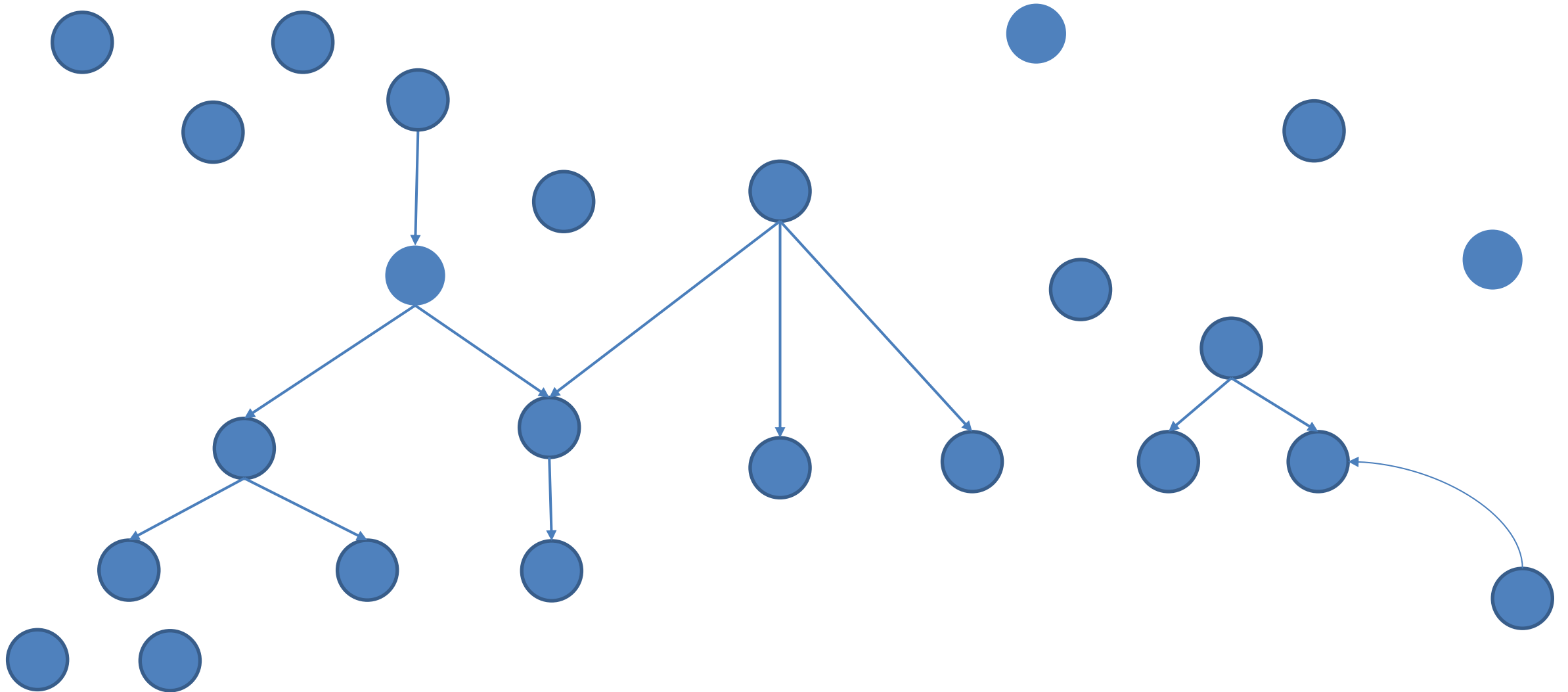


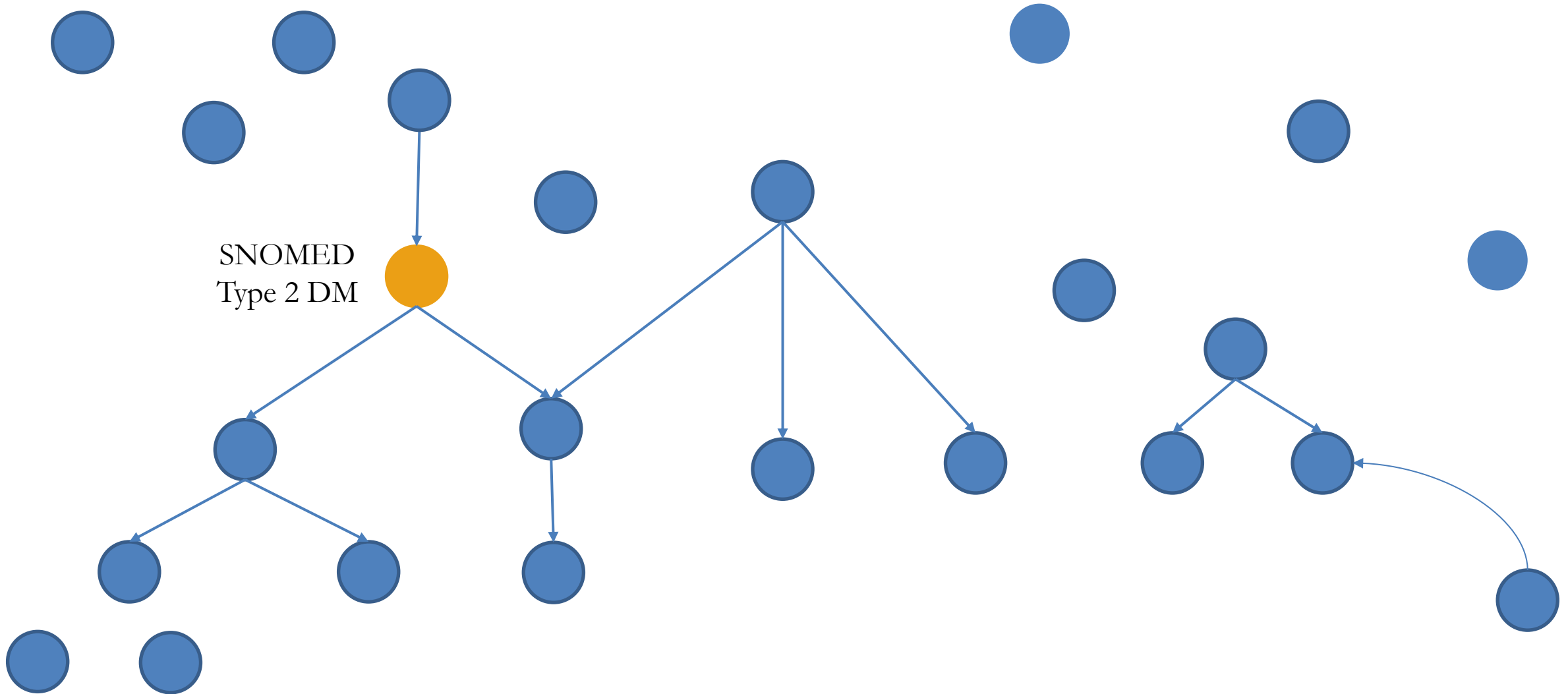
Even if we got the starting point, how do we select all the concepts that are relevant?





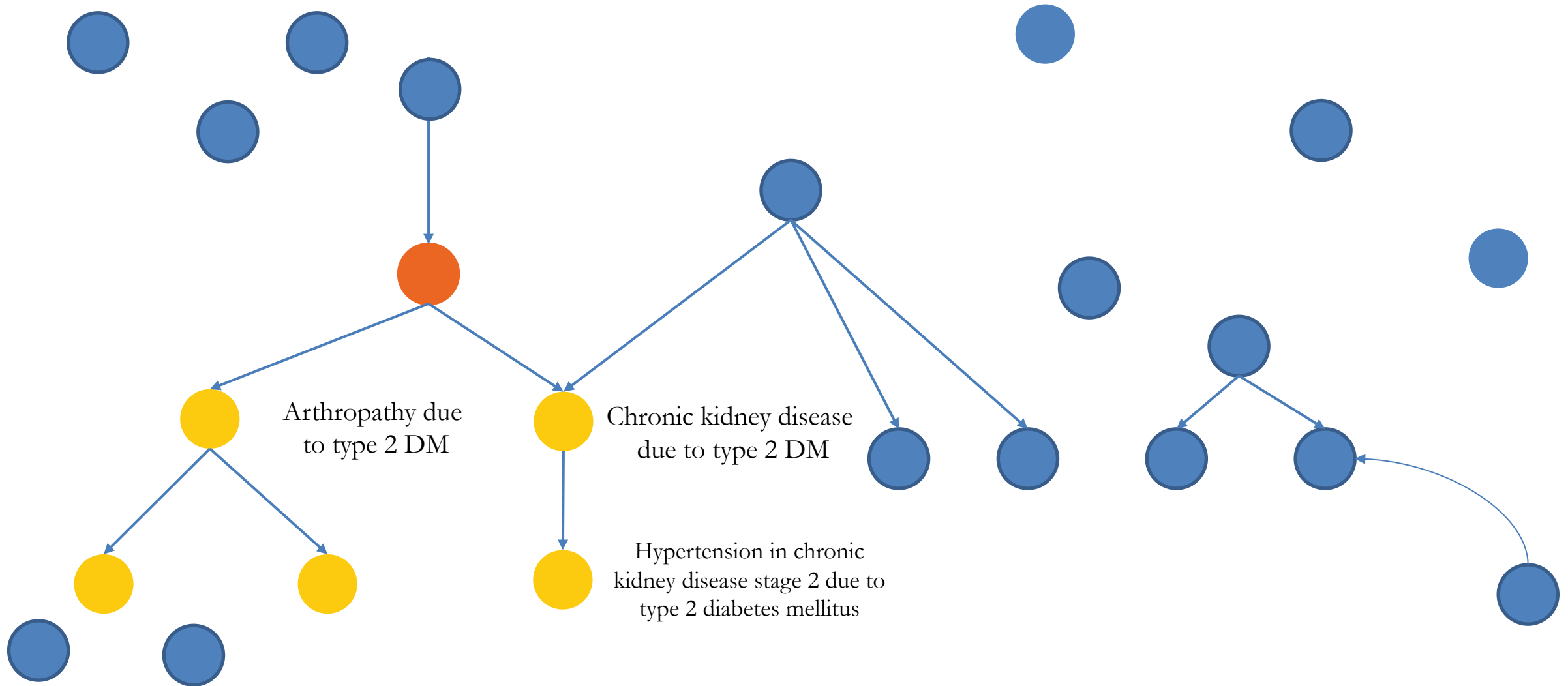
Select the starting point





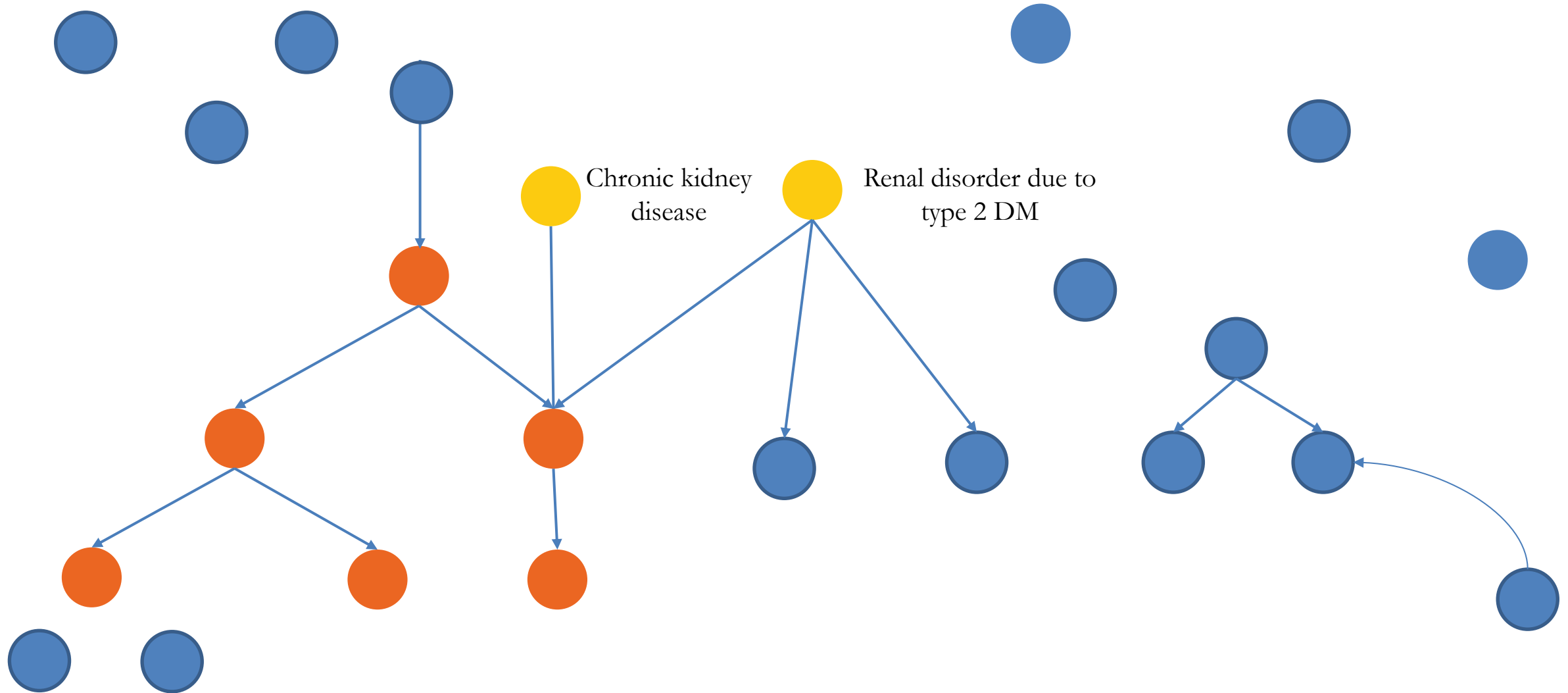


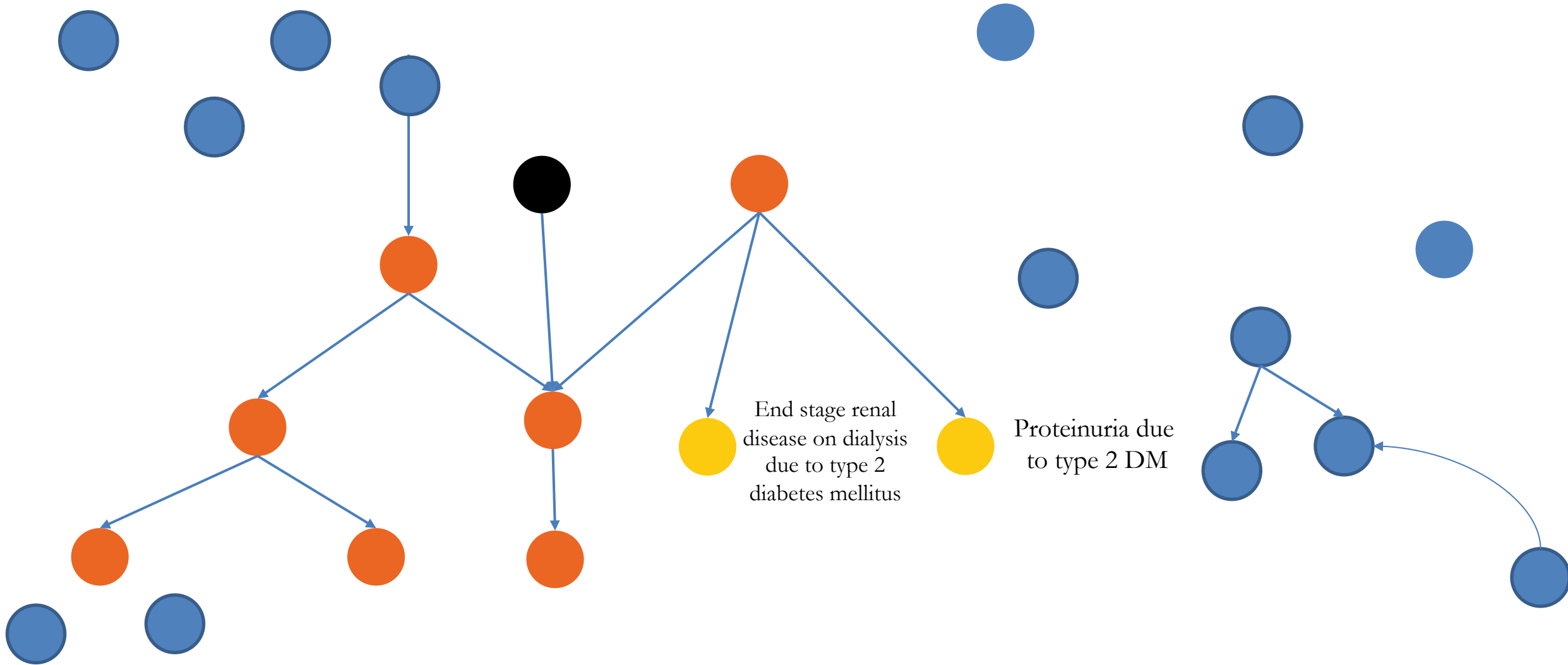
Select descendants

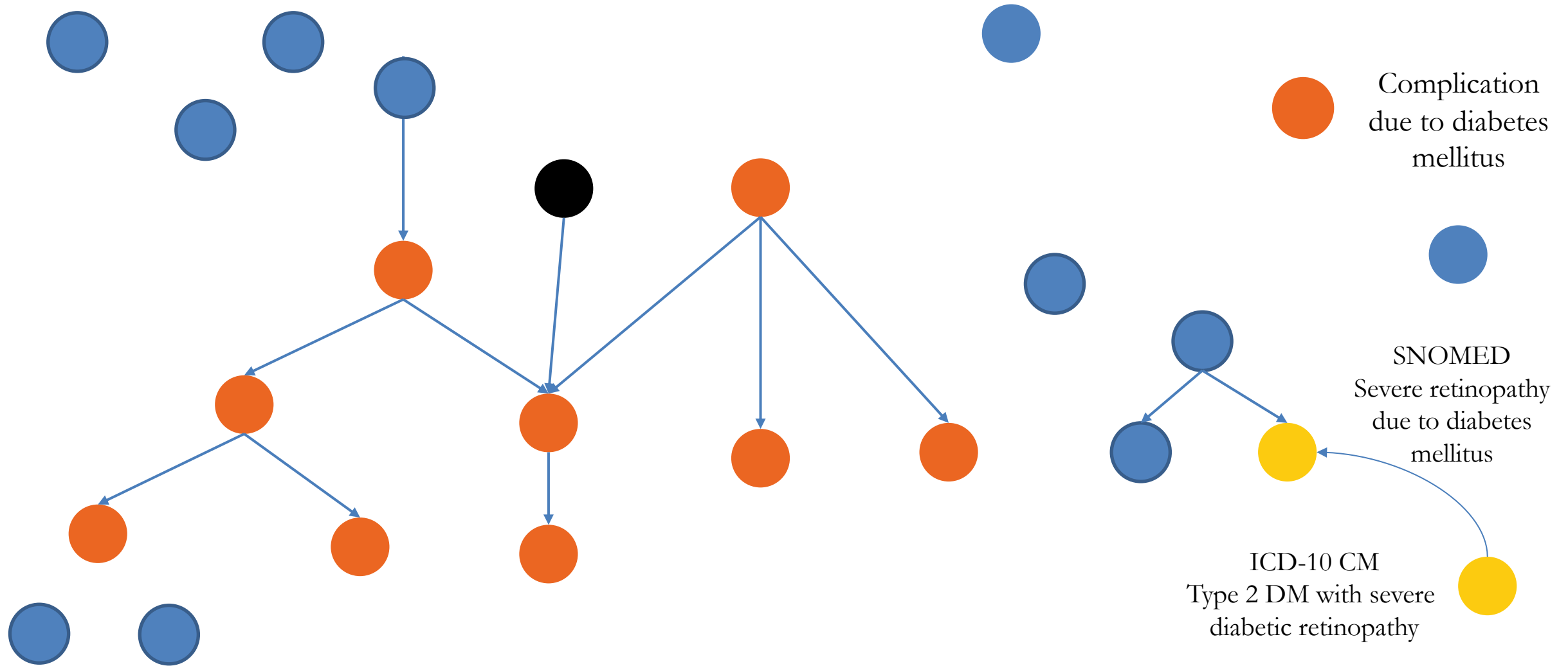




Check non selected parents

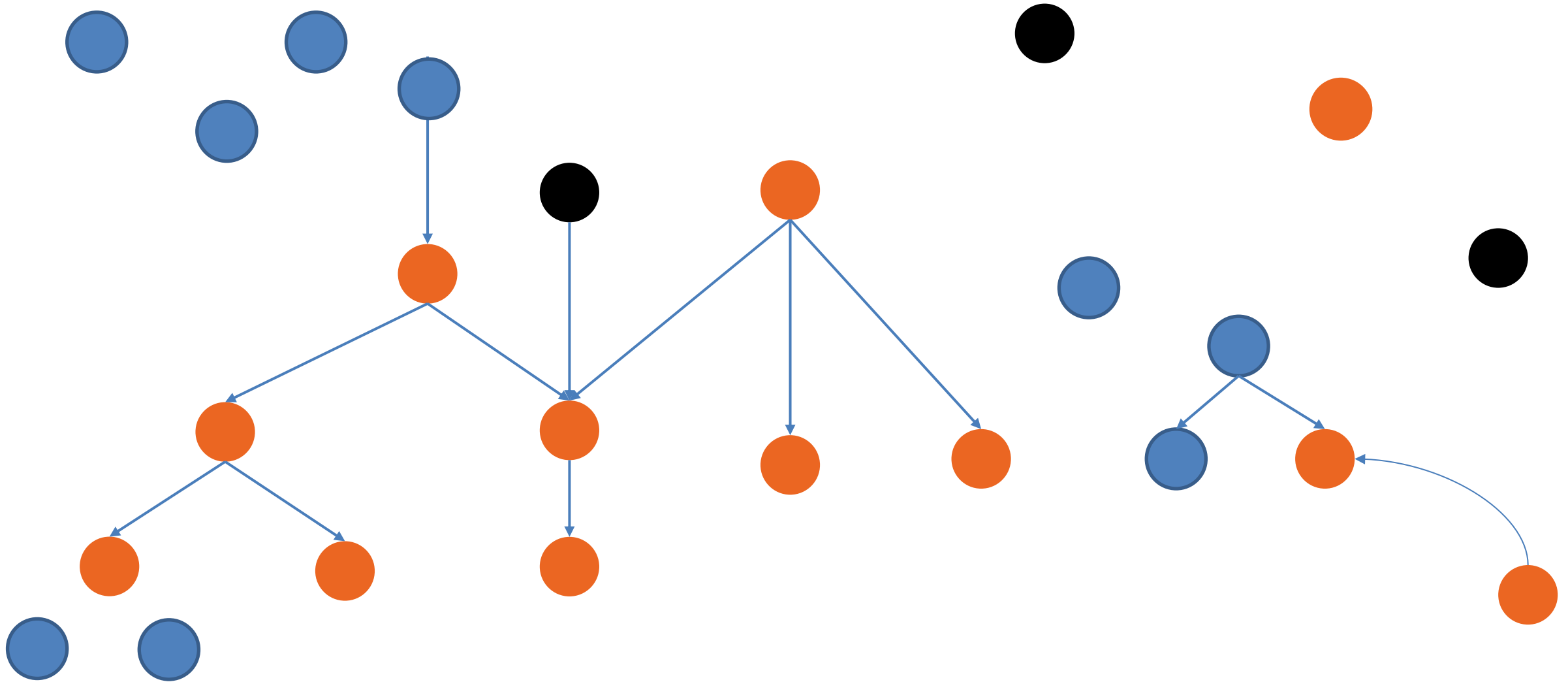








Iterating until a comprehensive concept set





A principled algorithm for concept selection

1. Pick initial concept
2. Review and add descendants
3. Review and add parents
4. Review lexically similar concepts
5. Review and remove included concepts
6. Create concept set
7. Repeat steps 2 - 6 until the concept set is unchanged



PHeNOType Observed Entity Baseline Endorsements (PHOEBE)

