A framework for large-scale characterization

Anthony Sena

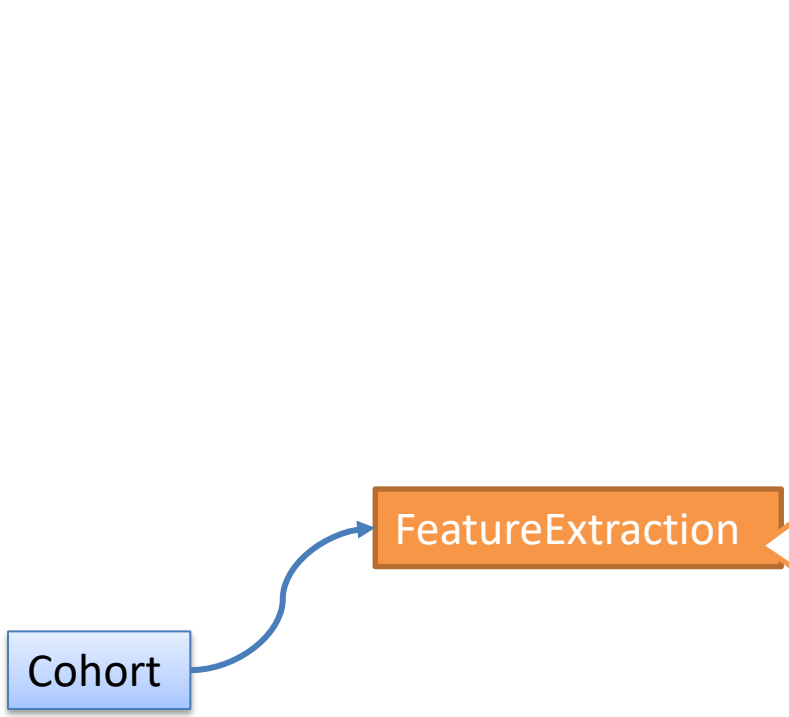
Janssen R&D, Erasmus MC

A journey through OHDSI's open source development

Cohort

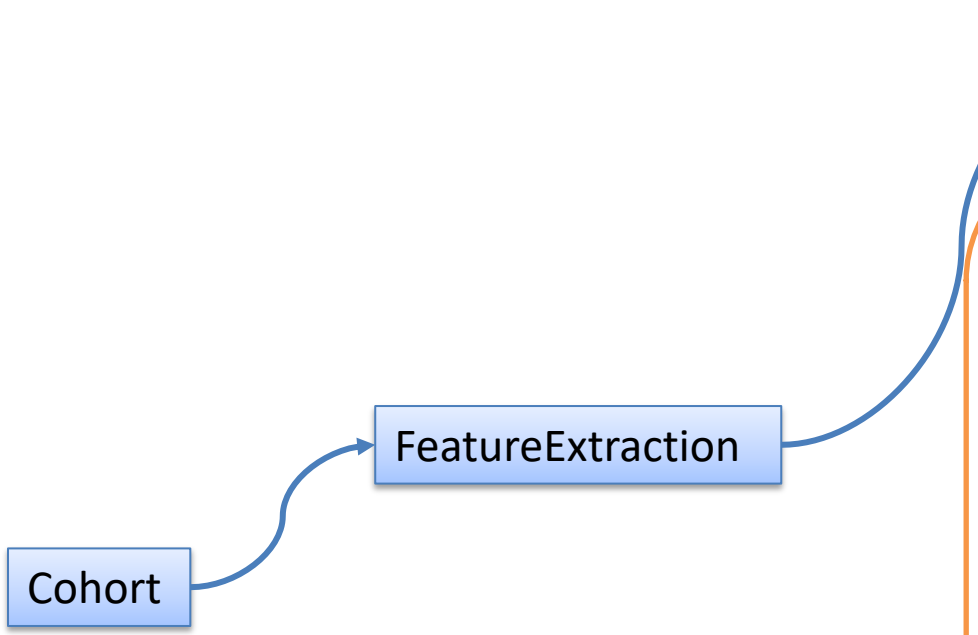
- Standardized representation (JSON) of set of persons satisfying a set of criteria for a period of time
- Standardized implementation (SQL) of specification of cohort entry event, inclusion criteria and exit event.
- Implemented as a criteria builder in the ATLAS interface
- Cornerstone of OHDSI study design powered by Circe

A journey through OHDSI's open source development



- R-package for constructing features at scale: all conditions, drugs, measurements, etc in a person's record
- Features are clinical observations relative to a cohort entry for a specified lookback time window
- Ability to aggregate statistics for concept-level feature reporting

A journey through OHDSI's open source development

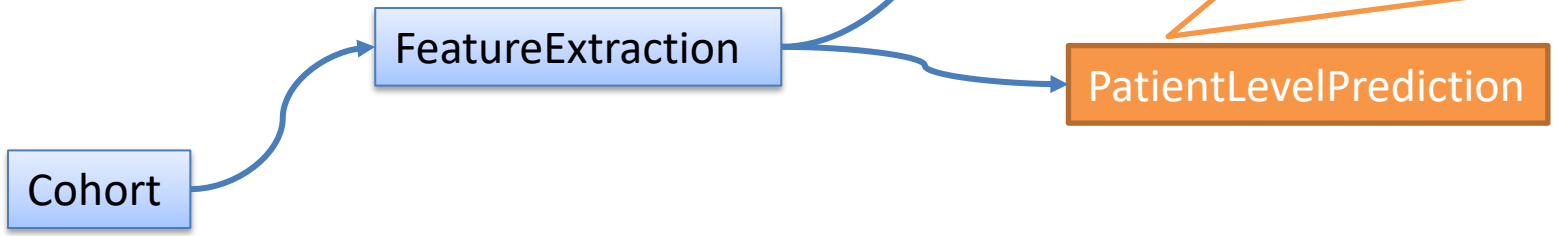


- Standardized framework utilizing cohorts as input to design population level estimation studies.
- Design questions defined using a target cohort (T), comparator cohort (C) and outcome cohort (O)
- Leverages FeatureExtraction for each target/comparator pair to compare features (standardize mean difference)
- Diagnostics built into the best practices of running an estimation study

HADES

A journey through OHDSI's open source development

- Standardized framework utilizing cohorts (T/Os) as input to design patient level prediction studies.
- Provides an R Shiny interface for interactive exploration of results
- R study package deployment via Hydra



A journey through OHDSI's open source development

- Subgroup criteria
- Custom feature based on criteria
- Compare results between >2 cohorts

CohortMethod

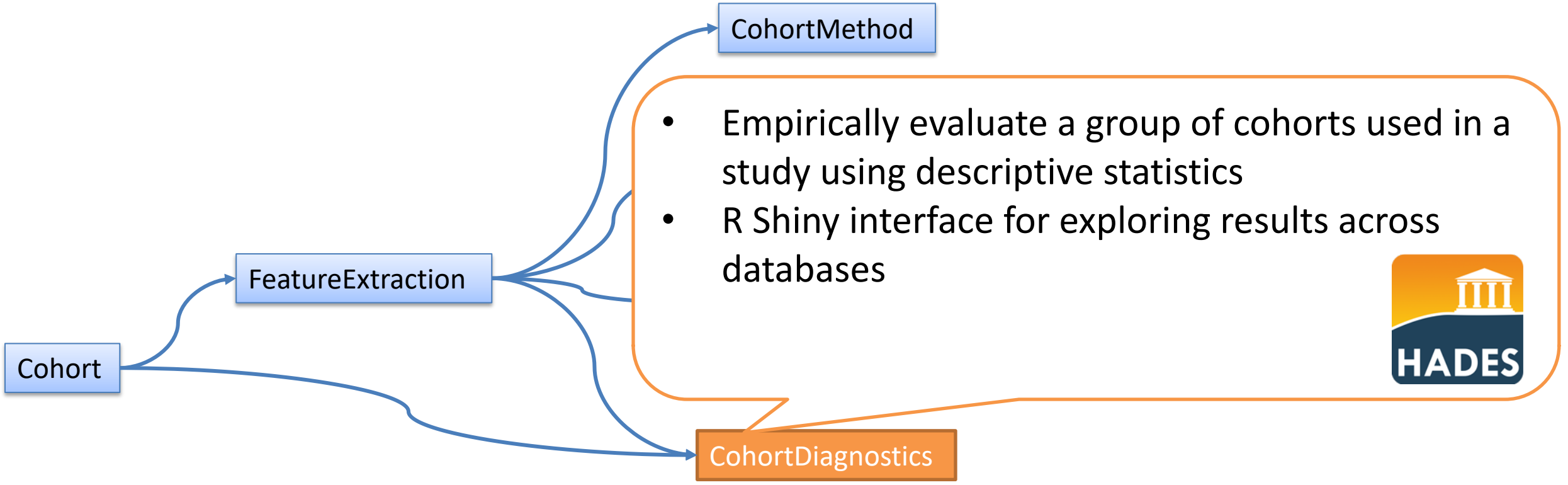
ATLAS/Characterization

FeatureExtraction

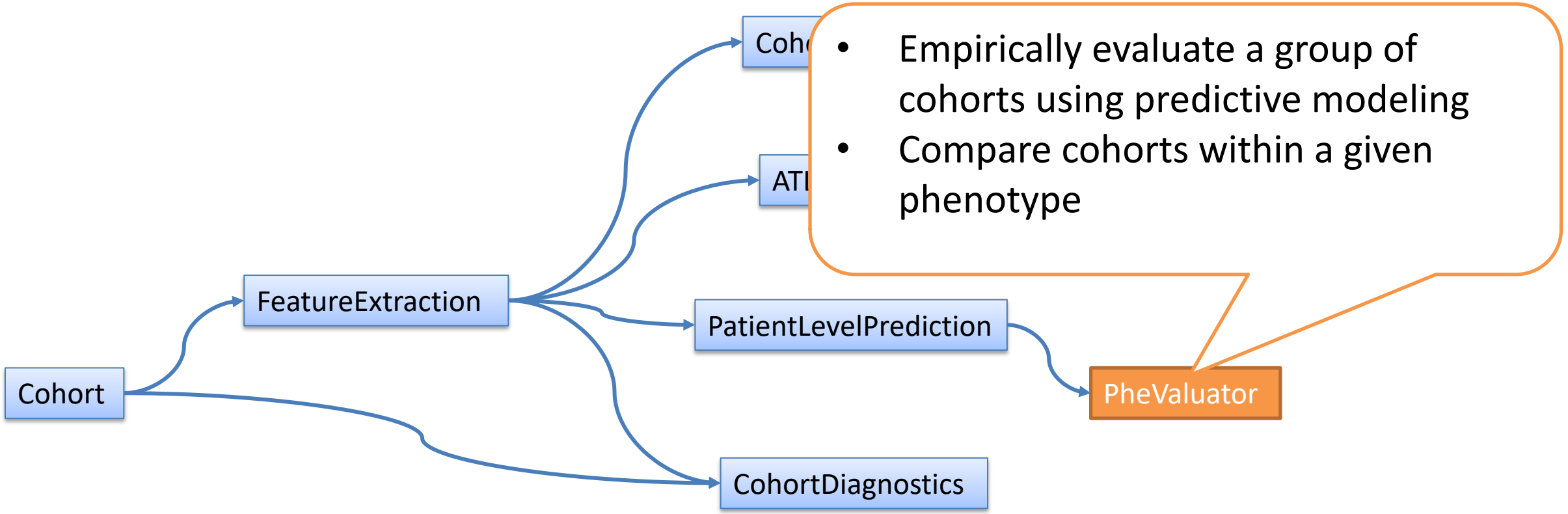
PatientLevelPrediction

Cohort

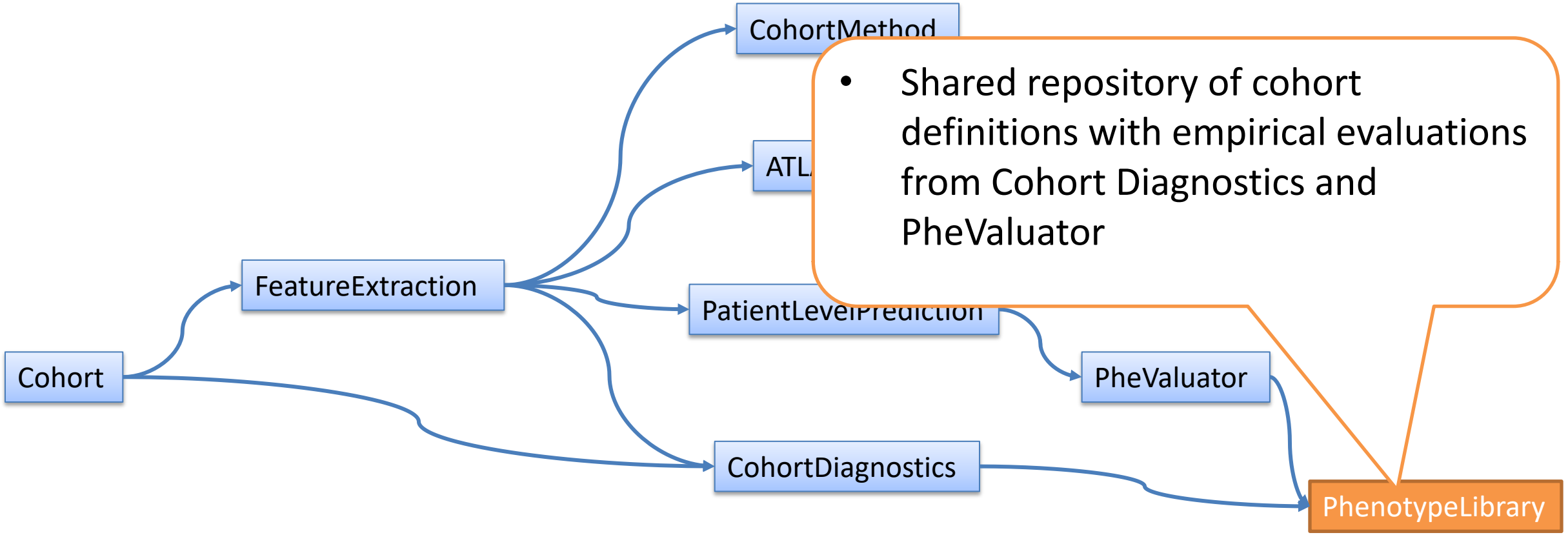
A journey through OHDSI's open source development



A journey through OHDSI's open source development

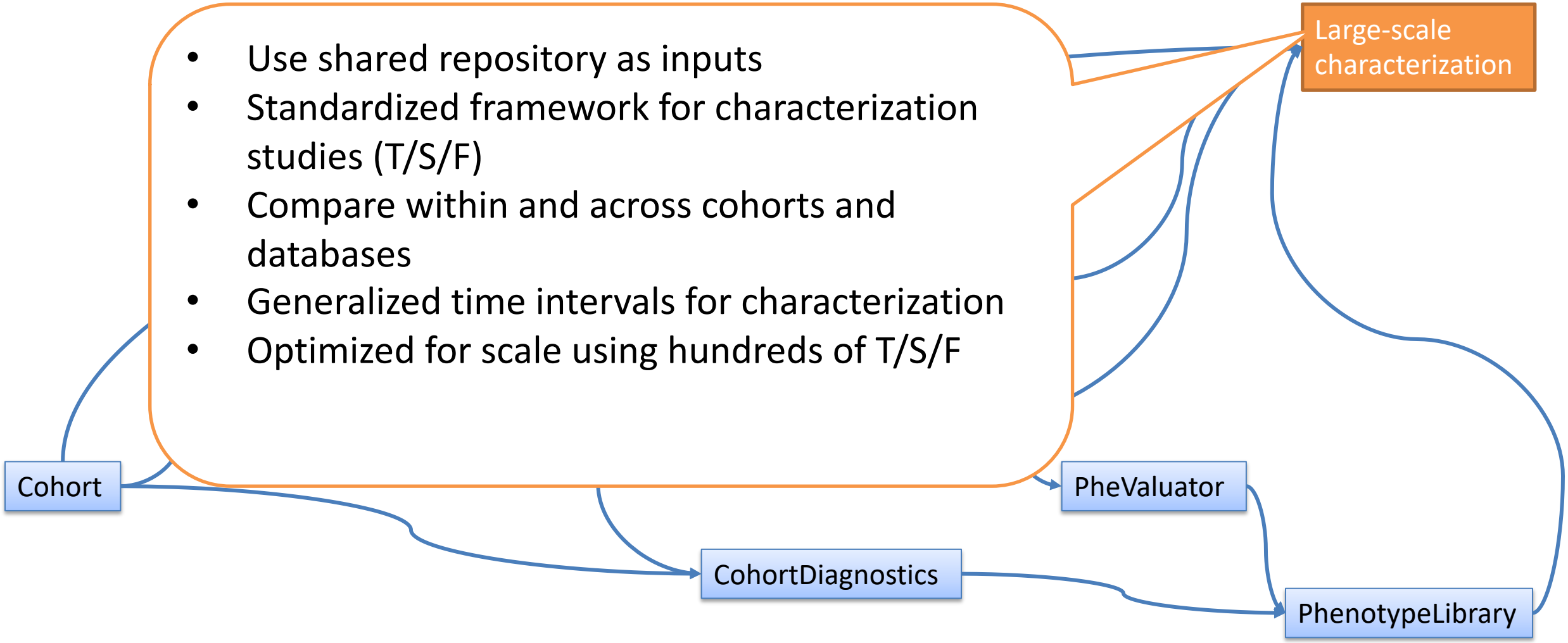


A journey through OHDSI's open source development

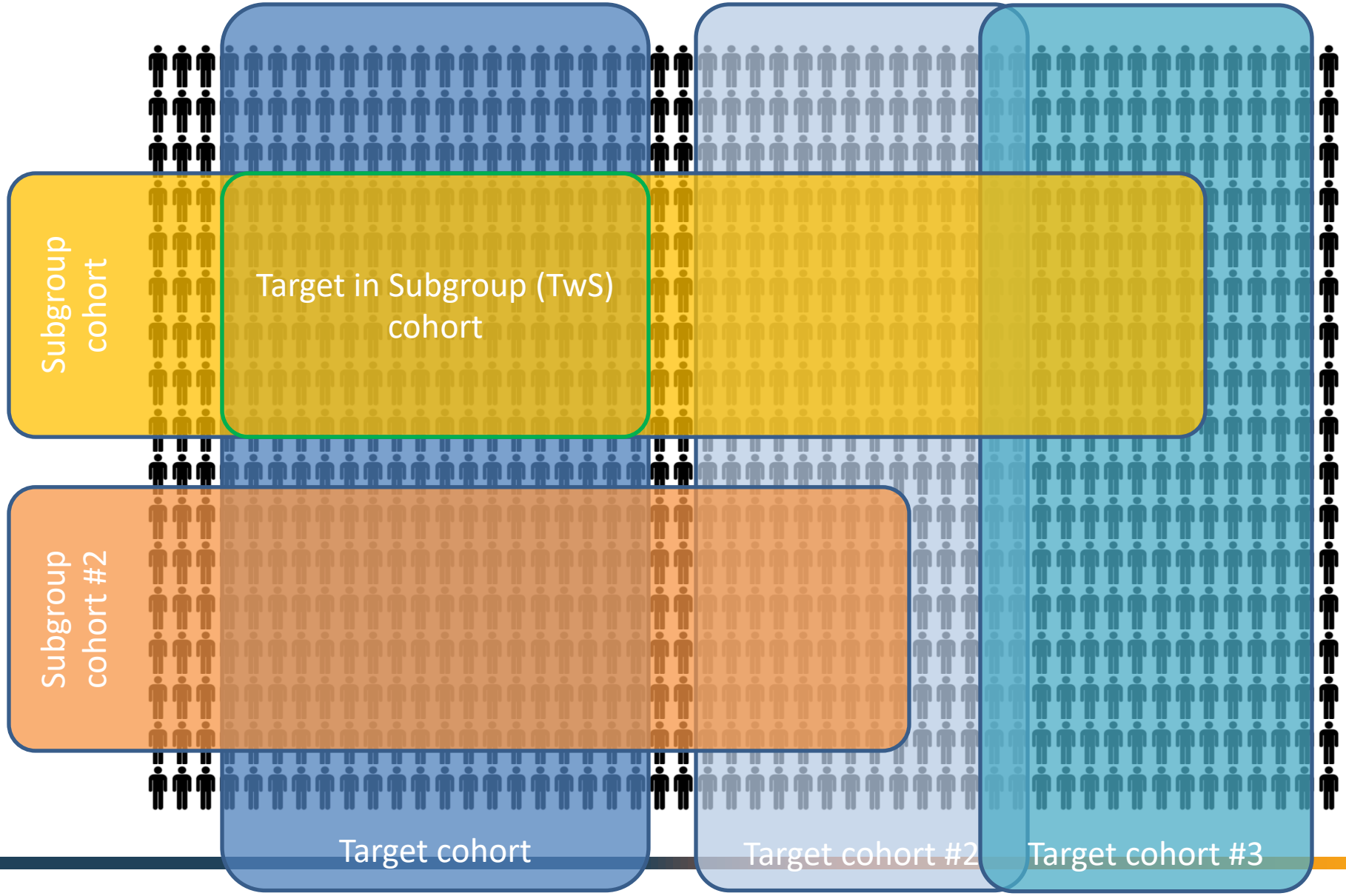


A journey through OHDSI's open source development

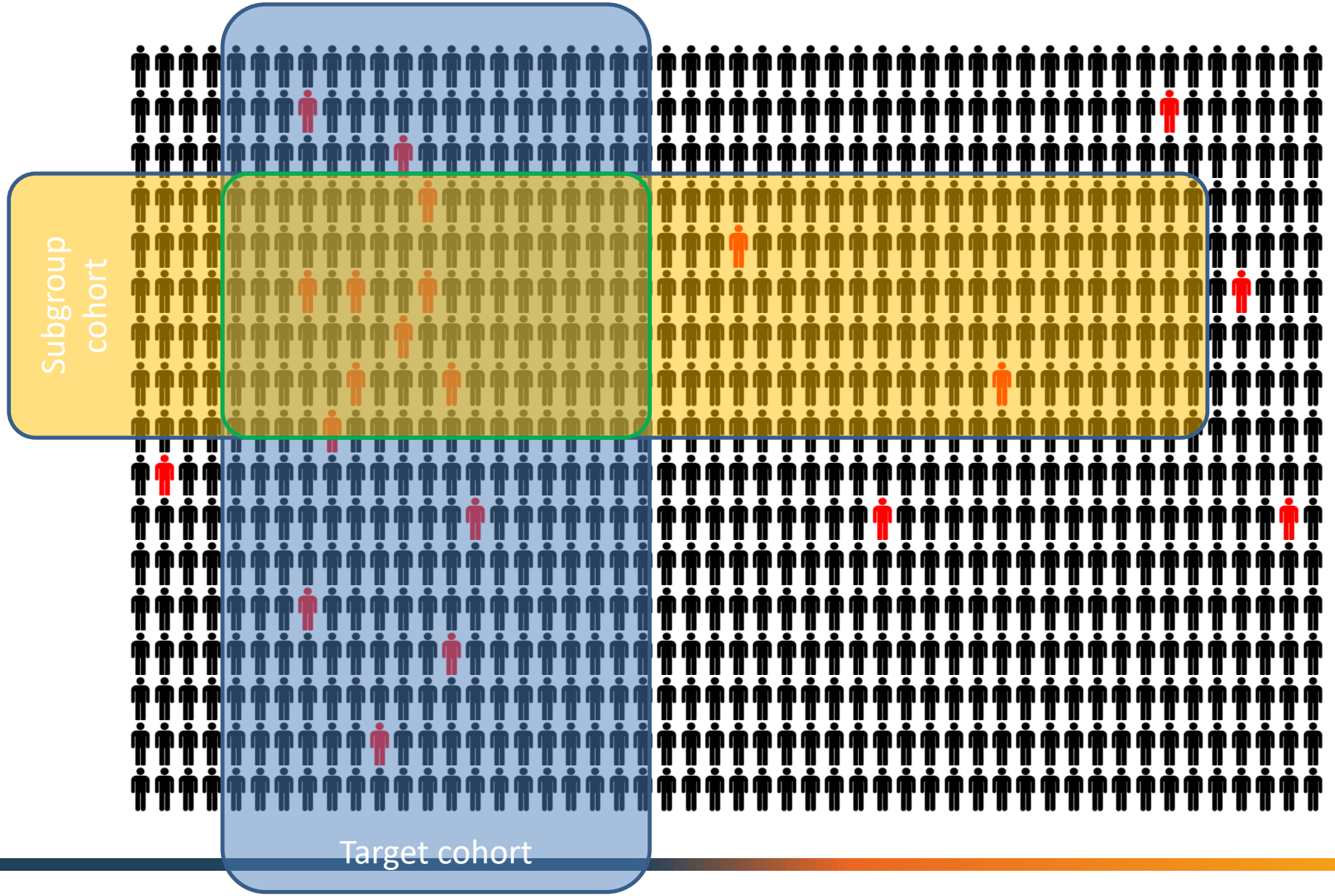
- Use shared repository as inputs
- Standardized framework for characterization studies (T/S/F)
- Compare within and across cohorts and databases
- Generalized time intervals for characterization
- Optimized for scale using hundreds of T/S/F



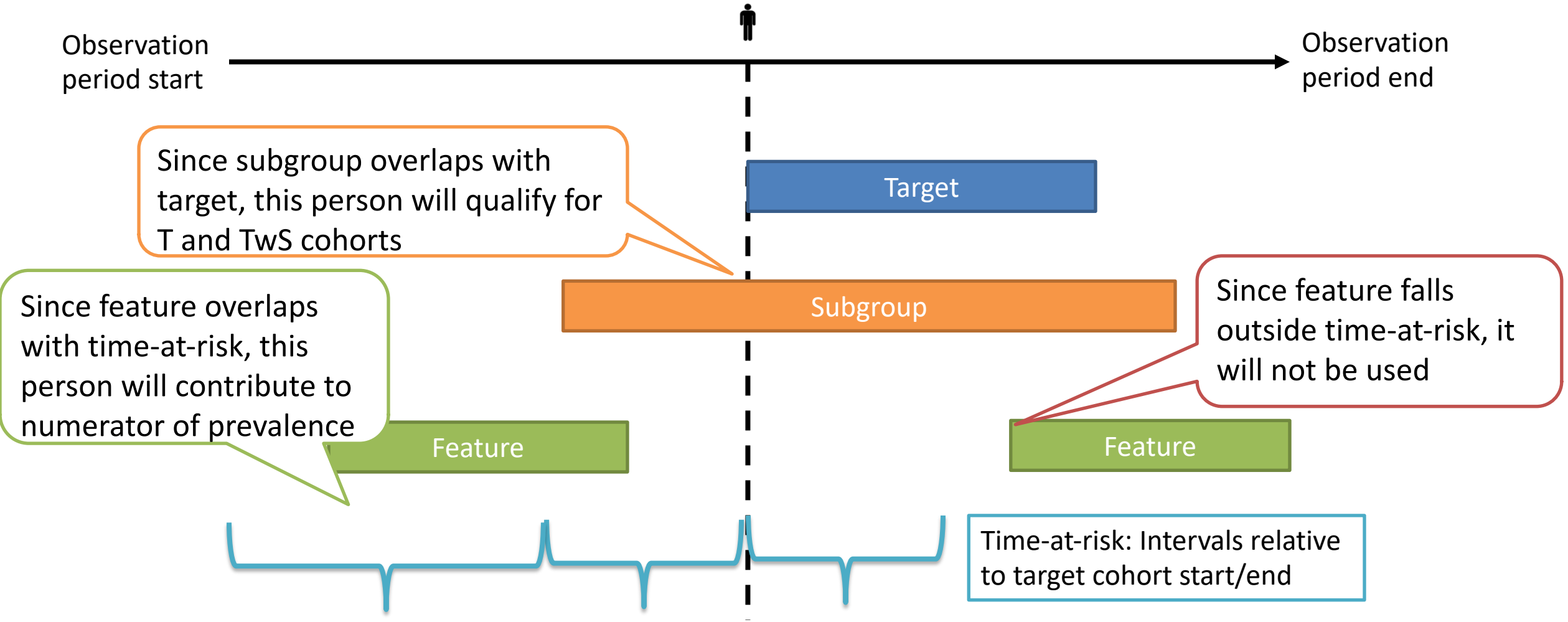
Characterizing a population: subgroups



Characterizing a population: features



Characterizing time in a characterization



Large scale characterization framework

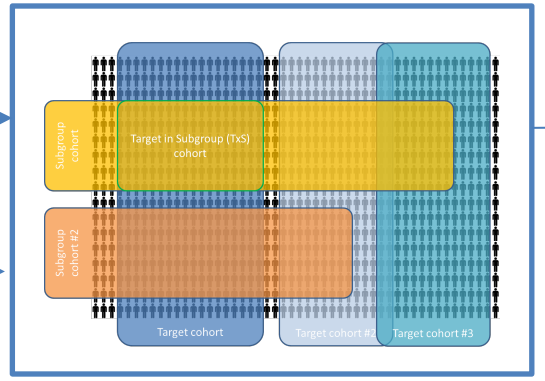
- Define a characterization study in terms of:
  - **Target cohorts (T)**: those to characterize
  - **Subgroup cohorts (S)**: those to use as subgroups of the target cohort(s)
  - **Feature cohorts (F):** cohorts used to construct features (outcomes) for characterization
  - **Time at risk windows**: Define windows of time to characterize all features (F) and concepts

Scaling cohort construction

Input:
Cohort specifications
(JSON)

Target cohorts (T)

Subgroups cohorts (S)



Output:
Instantiated cohorts
(COHORT table in CDM)

- Targets (T)
- Targets with Subgroup (TwS)
- Targets without Subgroup (TwoS)

Illustrative Example:
Input:

- 3 Targets:
  - Diabetes, hypertension, depression
- 4 Subgroups:
  - Female, Black, Young, Old

Output:

- 3 T + 3*4 TwS + 3*4 TwoS = 27 cohorts
  1. Persons with diabetes
  2. Persons with diabetes AND Female
  3. Persons with diabetes AND NOT Female
  4. Persons with diabetes AND Black
  5. Persons with diabetes AND NOT Black
  6. …

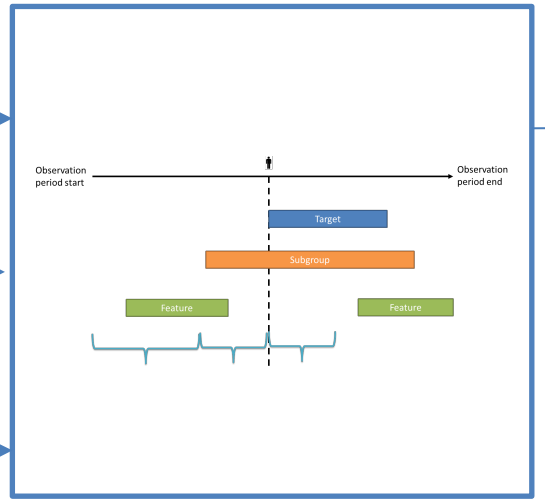
Scaling feature construction

Input:

Instantiated cohorts (T, TwS, TwoS)

Feature specifications (F)
- Cohorts
- Concept-based events

Time-at-risk windows



Output:

Summary statistics:
- Prevalence = Persons in cohort with feature during time-at-risk / Persons in cohort

Illustrative Example:

Input:
- 27 Cohorts:
  - Diabetes, hypertension, depression in Female/Black/Young/Old subgroups
- 5 Features:
  - Hospitalization, AMI, Death, Surgery, Drug initiation
- 2 Time-at-risk:
  - -365d to -1d; +1d to 365d

Output:
- 27 T * 5 F * 2 TAR = 270 statistics
  1. Persons with diabetes AND hospitalization IN -365d to -1d
  2. Persons with diabetes AND hospitalization IN +1d to +365d
  3. Persons with diabetes AND AMI IN -365d to -1d
  4. Persons with diabetes AND AMI IN +1d to +365d
  5. Persons with diabetes AND Death IN -365d to -1d
  6. Persons with diabetes AND Death IN +1d to +365d
  7. …

Large-scale characterization

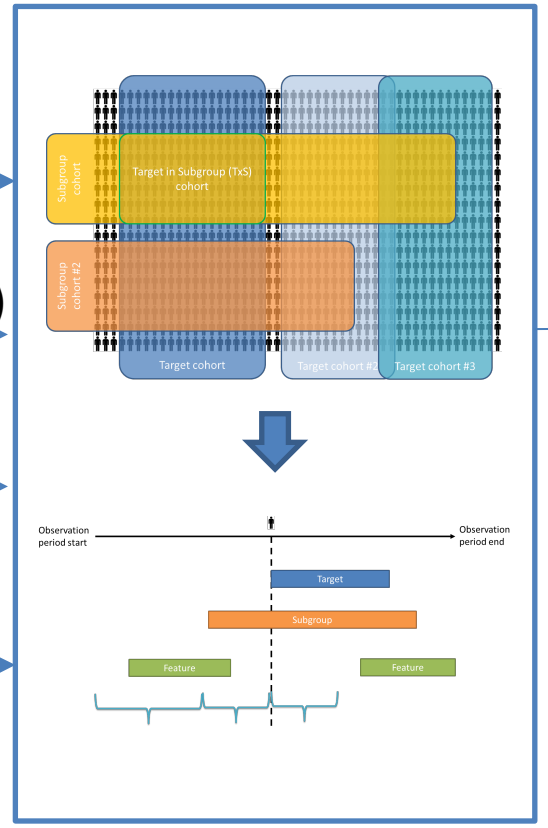
Input:

Target specifications (T)

Subgroup specifications (S)

Feature specifications (F)
- Cohorts
- Concept-based events

Time-at-risk windows

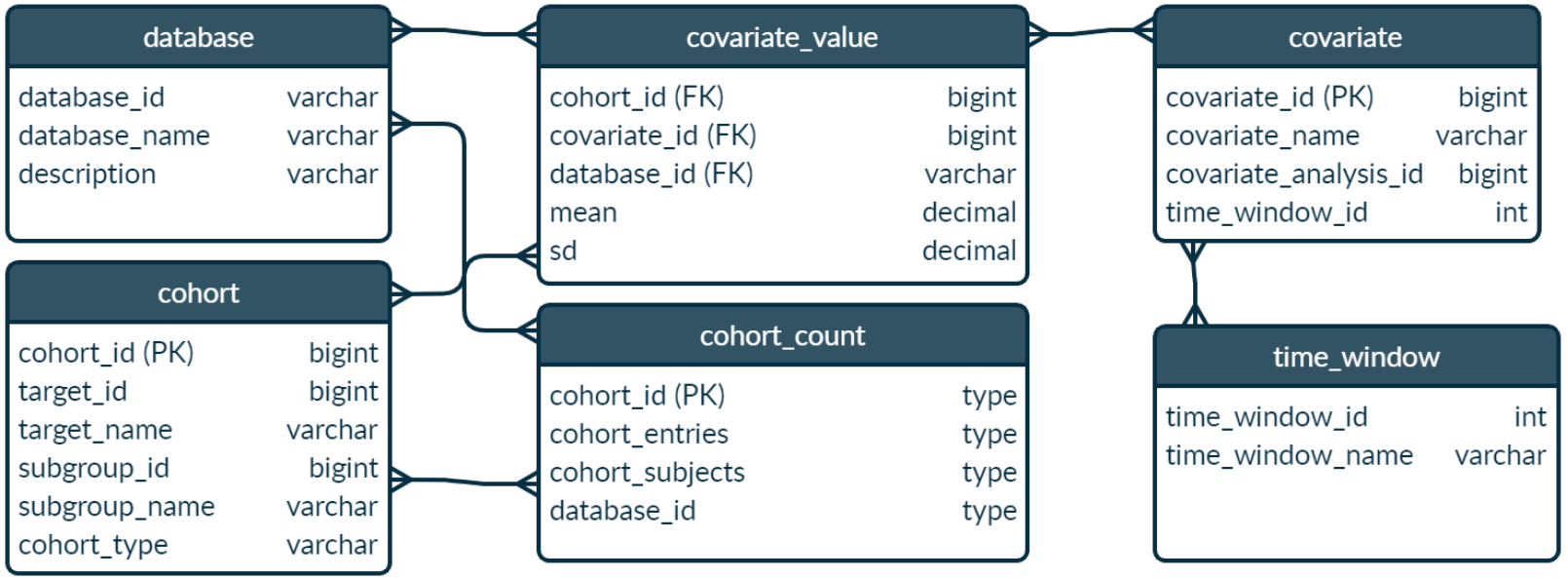


Output:

Summary statistics:
- Prevalence = Persons in cohort with feature during time-at-risk / Persons in cohort

Large scale characterization - design

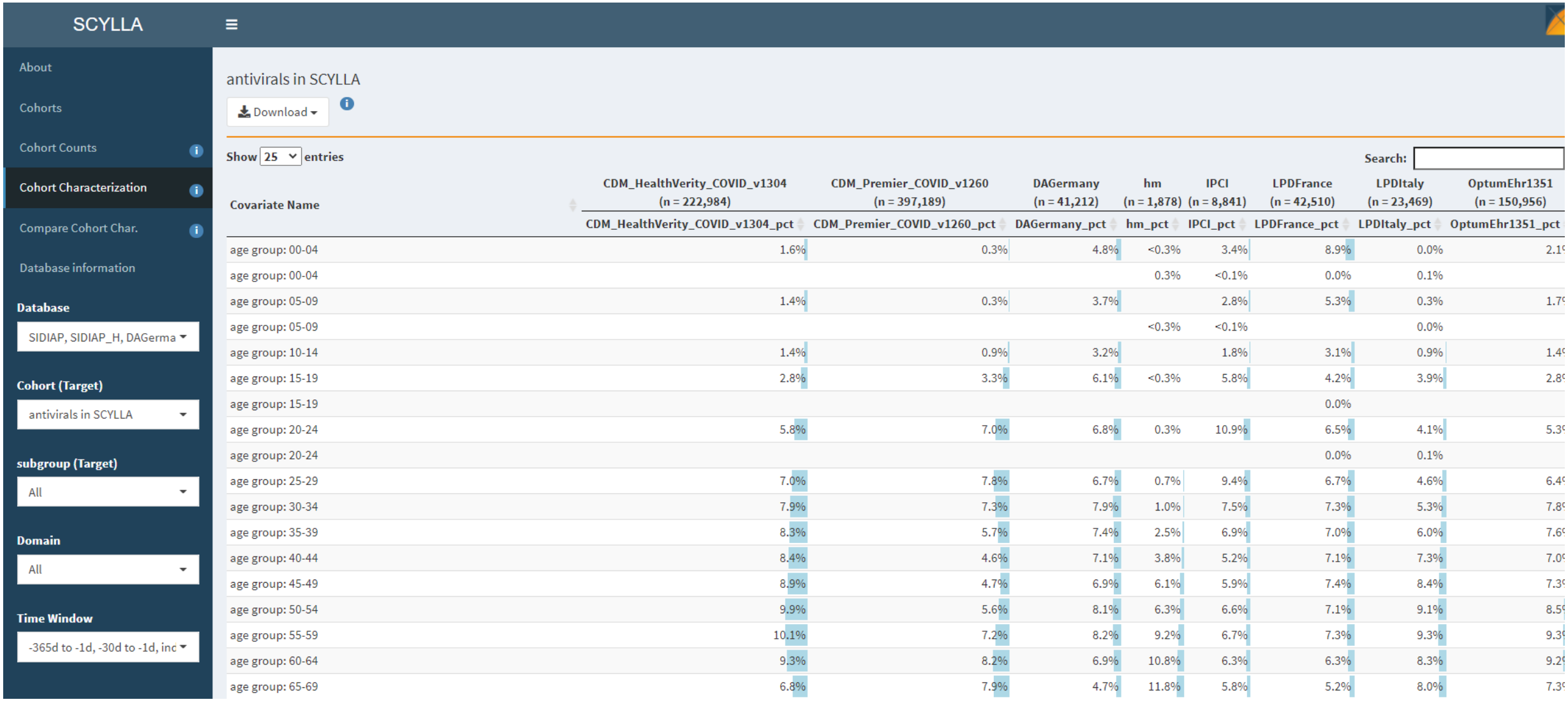
- Results are assembled for each database and stored in a simple data model:



Implementation

- Implemented in a series of R study packages
  - Run cohort diagnostics for all study cohorts
  - Apply large scale characterization for all T/S/F combinations for all time windows
  - Compute all demographics and concept-based features for all time windows (FeatureExtraction)
- Results reviewed via R Shiny application locally
- Results provided to study coordinators for inclusion into network results.

Characterization viewer



antivirals in SCYLLA

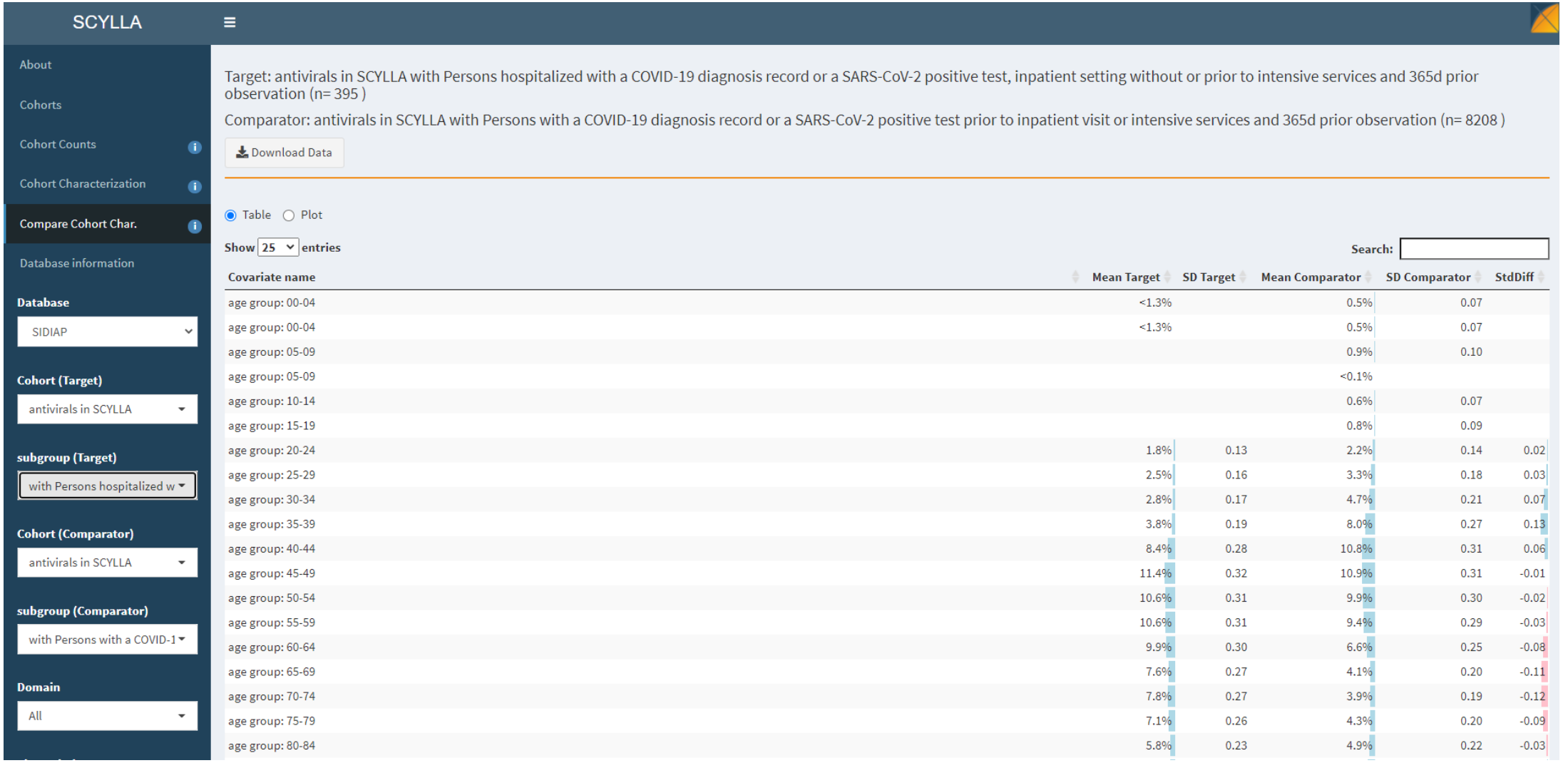
Download

Show 25 entries

Search:

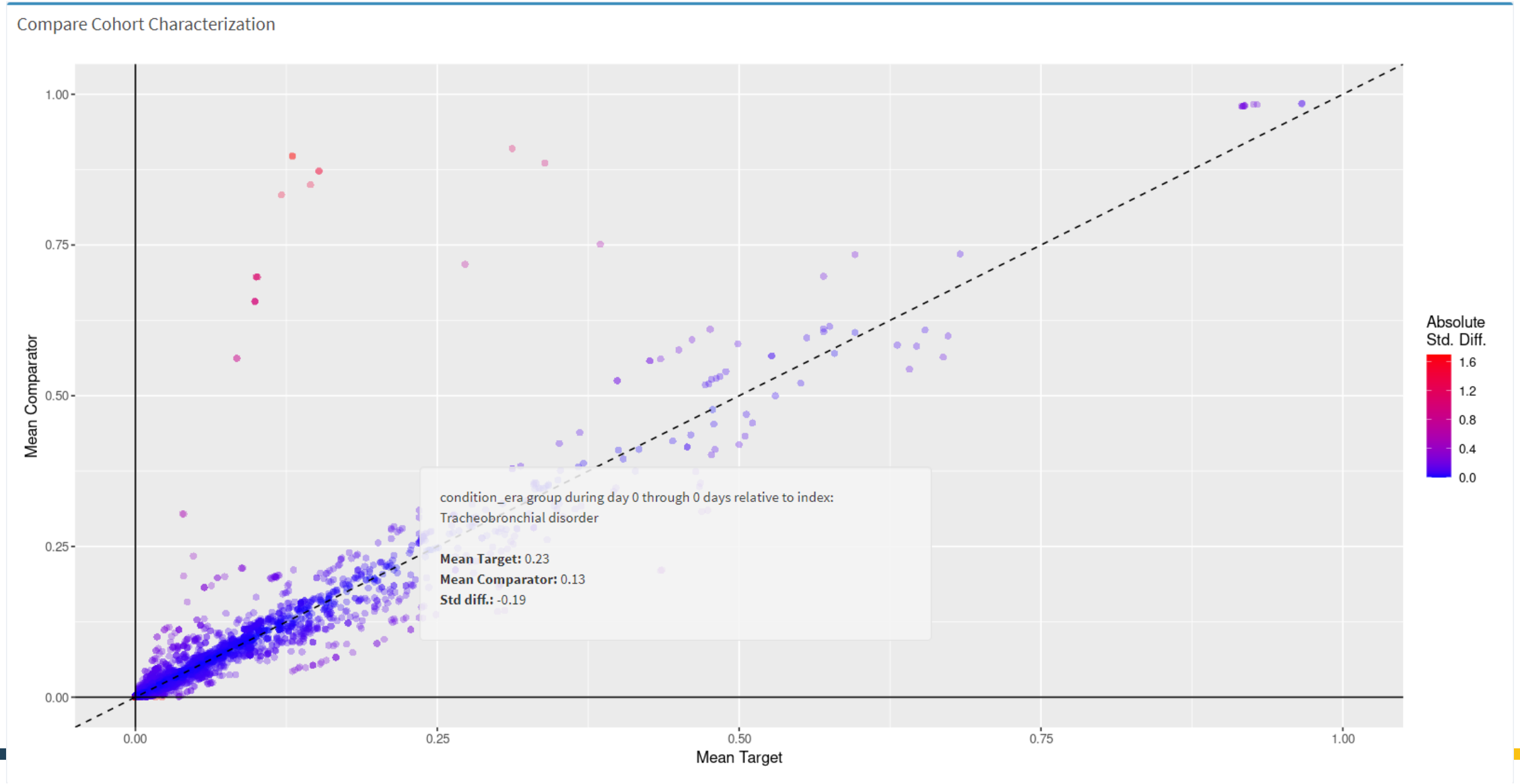
Covariate Name	CDM_HealthVerity_COVID_v1304 (n = 222,984)	CDM_Premier_COVID_v1260 (n = 397,189)	DAGermany (n = 41,212)	hm (n = 1,878)	IPCI (n = 8,841)	LPDFrance (n = 42,510)	LPDItaly (n = 23,469)	OptumEhr1351 (n = 150,956)
	CDM_HealthVerity_COVID_v1304_pct	CDM_Premier_COVID_v1260_pct	DAGermany_pct	hm_pct	IPCI_pct	LPDFrance_pct	LPDItaly_pct	OptumEhr1351_pct
age group: 00-04	1.6%	0.3%	4.8%	<0.3%	3.4%	8.9%	0.0%	2.1%
age group: 00-04				0.3%	<0.1%	0.0%	0.1%	
age group: 05-09	1.4%	0.3%	3.7%		2.8%	5.3%	0.3%	1.7%
age group: 05-09				<0.3%	<0.1%		0.0%	
age group: 10-14	1.4%	0.9%	3.2%		1.8%	3.1%	0.9%	1.4%
age group: 15-19	2.8%	3.3%	6.1%	<0.3%	5.8%	4.2%	3.9%	2.8%
age group: 15-19						0.0%		
age group: 20-24	5.8%	7.0%	6.8%	0.3%	10.9%	6.5%	4.1%	5.3%
age group: 20-24						0.0%	0.1%	
age group: 25-29	7.0%	7.8%	6.7%	0.7%	9.4%	6.7%	4.6%	6.4%
age group: 30-34	7.9%	7.3%	7.9%	1.0%	7.5%	7.3%	5.3%	7.8%
age group: 35-39	8.3%	5.7%	7.4%	2.5%	6.9%	7.0%	6.0%	7.6%
age group: 40-44	8.4%	4.6%	7.1%	3.8%	5.2%	7.1%	7.3%	7.0%
age group: 45-49	8.9%	4.7%	6.9%	6.1%	5.9%	7.4%	8.4%	7.3%
age group: 50-54	9.9%	5.6%	8.1%	6.3%	6.6%	7.1%	9.1%	8.5%
age group: 55-59	10.1%	7.2%	8.2%	9.2%	6.7%	7.3%	9.3%	9.3%
age group: 60-64	9.3%	8.2%	6.9%	10.8%	6.3%	6.3%	8.3%	9.2%
age group: 65-69	6.8%	7.9%	4.7%	11.8%	5.8%	5.2%	8.0%	7.3%

Cohort Comparison Viewer – Tabular View



Covariate name	Mean Target	SD Target	Mean Comparator	SD Comparator	StdDiff
age group: 00-04	<1.3%		0.5%	0.07	
age group: 00-04	<1.3%		0.5%	0.07	
age group: 05-09			0.9%	0.10	
age group: 05-09			<0.1%		
age group: 10-14			0.6%	0.07	
age group: 15-19			0.8%	0.09	
age group: 20-24	1.8%	0.13	2.2%	0.14	0.02
age group: 25-29	2.5%	0.16	3.3%	0.18	0.03
age group: 30-34	2.8%	0.17	4.7%	0.21	0.07
age group: 35-39	3.8%	0.19	8.0%	0.27	0.13
age group: 40-44	8.4%	0.28	10.8%	0.31	0.06
age group: 45-49	11.4%	0.32	10.9%	0.31	-0.01
age group: 50-54	10.6%	0.31	9.9%	0.30	-0.02
age group: 55-59	10.6%	0.31	9.4%	0.29	-0.03
age group: 60-64	9.9%	0.30	6.6%	0.25	-0.08
age group: 65-69	7.6%	0.27	4.1%	0.20	-0.11
age group: 70-74	7.8%	0.27	3.9%	0.19	-0.12
age group: 75-79	7.1%	0.26	4.3%	0.20	-0.09
age group: 80-84	5.8%	0.23	4.9%	0.22	-0.03

Cohort Comparison Viewer – Plot



Large-scale characterization in action: CHARYBDIS

Objective: Large scale characterization of COVID-19 disease natural history

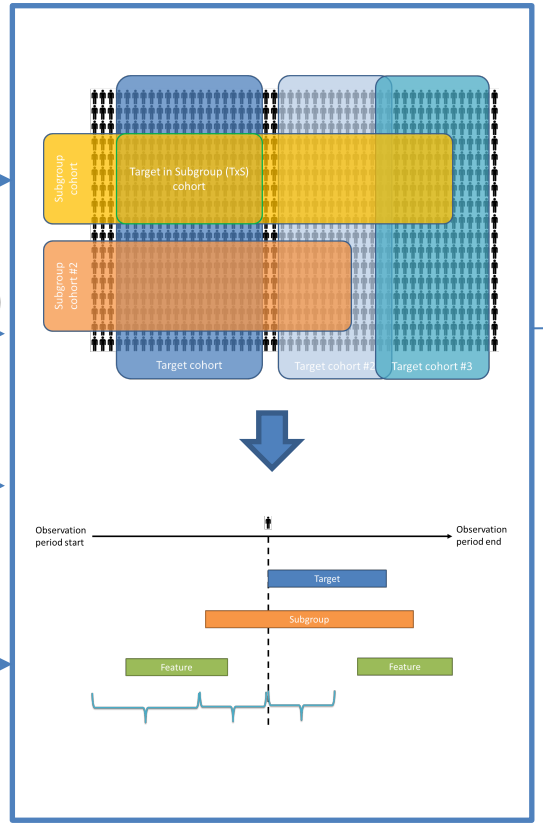
Input:

22 Target specifications (T)

58 Subgroup specifications (S)

64 Feature specifications (F)
- Cohorts
- Concept-based events

4 Time-at-risk windows



Output:

326,656

Summary statistics:
- Prevalence = Persons in cohort with feature during time-at-risk / Persons in cohort

.... x **22** Contributing Databases

= **>286 million** results and counting

Large-scale characterization – all the feels

Credit: Kristin Kostka

Large-scale characterization in action: SCYLLA

Objective: Large scale characterization for COVID-19 therapeutic evaluation

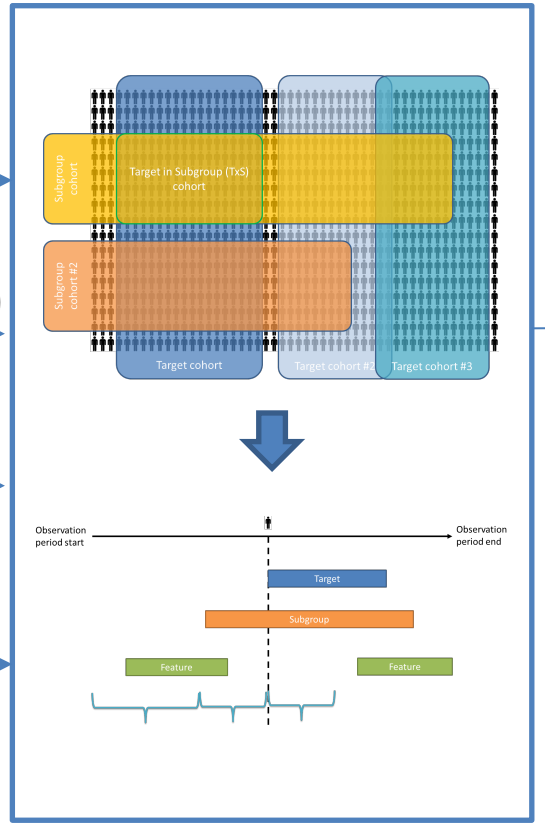
Input:

89 Target specifications (T)

6 Subgroup specifications (S)

64 Feature specifications (F)
- Cohorts
- Concept-based events

4 Time-at-risk windows



Output:

136,704

Summary statistics:
- Prevalence = Persons in cohort with feature during time-at-risk / Persons in cohort

.... x **13** Contributing Databases

= **>27 million** results and counting

Large-scale characterization in action: HERA

Objective: Large scale characterization of gender and racial disparities

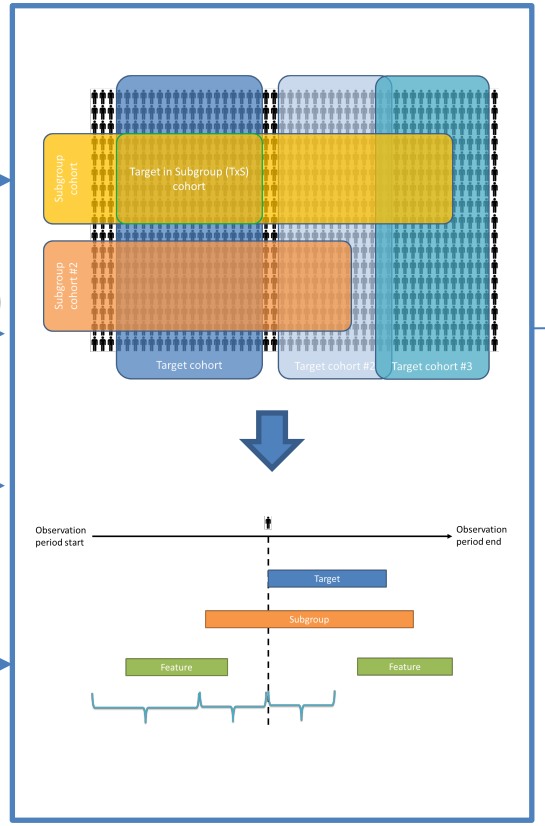
Input:

132 Target specifications (T)

8 Subgroup specifications (S)

124 Feature specifications (F)
- Cohorts
- Concept-based events

3 Time-at-risk windows



Output:

392,832

Summary statistics:
- Prevalence = Persons in cohort with feature during time-at-risk / Persons in cohort

.... x **5** Contributing Databases

= **>107 million** results and counting