Cohort Diagnostics

Gowtham Rao, Azza Shoaibi, Jamie Gilbert, Martijn Schuemie

Background:

Cohort Diagnostics (1), an OHDSI HADES package (2), has gone through substantial development in the past year and its use is considered a recommended best practice step prior to performing an OHDSI network study. The software enables iterative decision making by enabling the comparison of one or more cohort definition design choices for similar clinical ideas. Users can then infer from the variations introduced by said choices on sensitivity, specificity, and consistency over a network of data sources. Cohort Diagnostics enables decisions such as the feasibility to develop cohort definitions for a clinical idea, improvement of definitions by comparing diagnostic performance to each other and a-priori expectations. Also, it may be used to generate descriptive-based evidence.

We illustrate how we used Cohort Diagnostics to compare 2 alternate definitions for the clinical idea of anaphylaxis - one attempting to capture any type of anaphylaxis (C3), and other not including anaphylaxis due environmental causes such as insect bites, food exposure (C1) anaphylaxis - to empirically decide of which definition is appropriate for the clinical idea of drug/vaccine induced anaphylaxis.

Methods:

Given the two cohort definitions (C1 vs C3), cohort diagnostics generated an output in the form of a pre-specified standard results data model. This output, a non-person level de-identified result set, obtained from multiple independent data sources over a network is combined into one. The output is reviewed using the Diagnostics Explorer viewer application (part of Cohort Diagnostics) and published at data.ohdsi.org (3) as part of a short tutorial.

Briefly using Cohort Diagnostics, inferences are drawn at the data source level, cohort definition level and concept-id level:

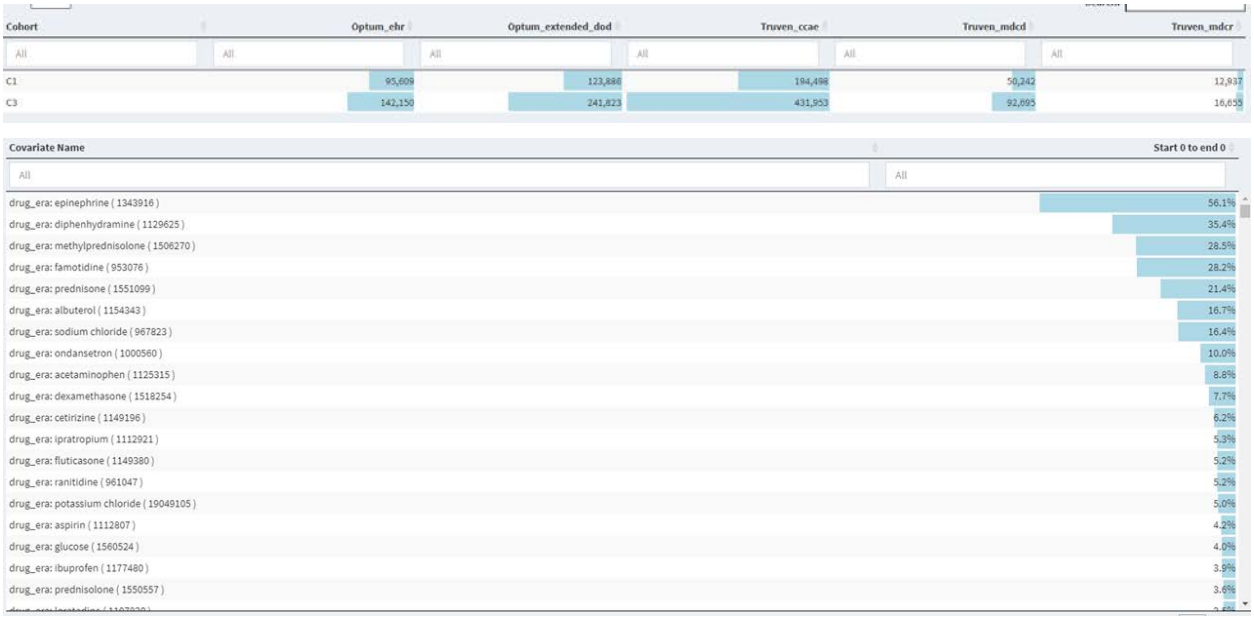
- Data source level diagnostics may be used to infer about the data source heterogeneity. If data source heterogeneity is observed, then researchers should attempt to explain the observed heterogeneity by understanding the source data and make determination if that understanding might introduce limitations on the interpretation of the results of the proposed study.
- Concept id level diagnostics: such as *orphan concepts* are any concepts that appear to have substring similarity with the concepts in the cohorts being diagnosed but are not present in the cohort definition but seem to capture the clinical idea behind the cohort definition and have sufficient counts in one or more data sources. If yes, was this code missed in the original cohort definition i.e., should the cohort definition include these concepts?
- Cohort level diagnostics: These are main diagnostics in cohort diagnostics and help us to infer about the cohorts by reviewing and comparing descriptive characteristics of the instantiated cohorts across and within data sources.

Results:

We first reviewed the two cohorts C3 (with environmental exposure) and C1 (without environmental exposure) and determined if there were sufficient counts across all data sources (> 1,000). As expected, C3 was found to have more counts compared to C1 across many data sources, but the relative difference appeared to be consistent across the data sources we studied.

To improve confidence in the cohort, we looked at the drugs used on the index date, and found a relatively high proportion of persons to have received drugs commonly used in the treatment of anaphylaxis. When C3 was compared to C1, we found the proportion to be higher in C1 compared to C3 suggesting that C3 is a more specific cohort for the phenotype anaphylaxis.

Figure 1: Cohort counts (A) and temporal characterization (B)

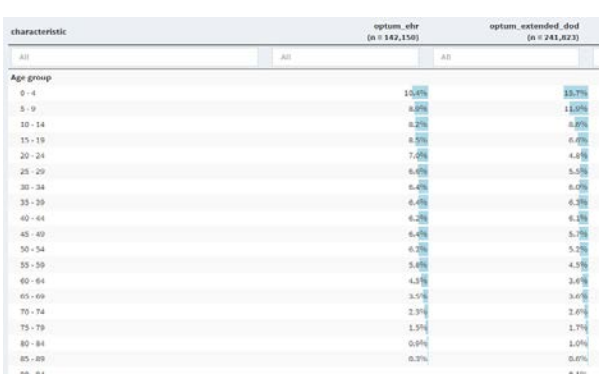


Cohort	Optum_ehr	Optum_extended_dod	Truven_ccae	Truven_mdcd	Truven_mdcr
C1	95,609	123,886	194,498	50,242	12,937
C3	142,150	241,823	431,953	92,695	16,655

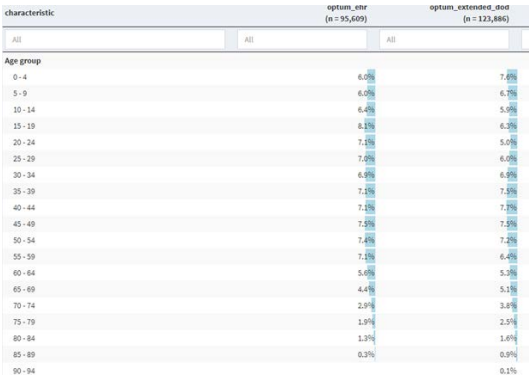
Covariate Name	Start 0 to end 0
drug_era: epinephrine (1343916)	56.1%
drug_era: diphenhydramine (1129625)	35.4%
drug_era: methylprednisolone (1506270)	28.5%
drug_era: famotidine (953076)	28.2%
drug_era: prednisone (1551099)	21.4%
drug_era: albuterol (1154343)	16.7%
drug_era: sodium chloride (967823)	16.4%
drug_era: ondansetron (1000560)	10.0%
drug_era: acetaminophen (1125315)	8.8%
drug_era: dexamethasone (1518254)	7.7%
drug_era: cetirizine (1149196)	6.2%
drug_era: ipratropium (1112921)	5.3%
drug_era: fluticasone (1149380)	5.2%
drug_era: ranitidine (961047)	5.2%
drug_era: potassium chloride (19049105)	5.0%
drug_era: aspirin (1112807)	4.2%
drug_era: glucose (1560524)	4.0%
drug_era: ibuprofen (1177480)	3.9%
drug_era: prednisolone (1550557)	3.6%

We looked at the age distribution of the two cohorts. In the C3 cohort we found that it was weighted more towards the younger population, compared to the C1 cohort which had a more uniform distribution. The uniform distribution was more in line with our prior expectation for the drug/vaccine induced anaphylaxis, as we believe that environmental exposure induced anaphylaxis is more likley in children.

Figure 2: Age distributions by database



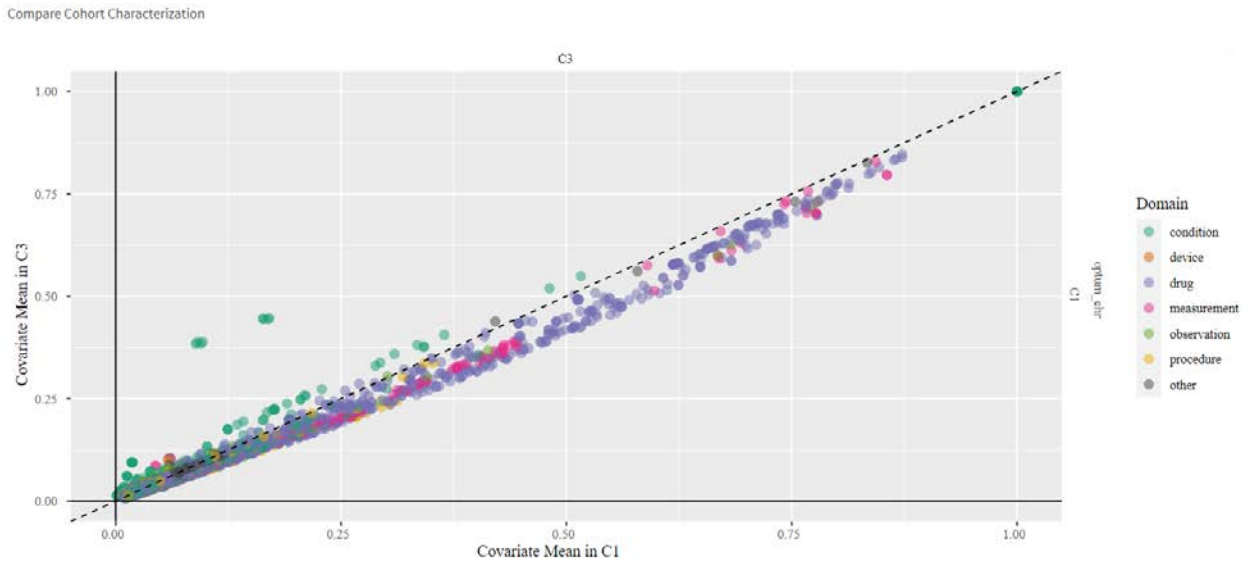
characteristic	optum_ehr (n = 142,150)	optum_extended_dod (n = 241,823)
Age group		
0 - 4	10.4%	13.7%
5 - 9	8.0%	11.0%
10 - 14	8.2%	8.8%
15 - 19	8.5%	8.4%
20 - 24	7.0%	6.8%
25 - 29	6.6%	5.5%
30 - 34	6.4%	6.0%
35 - 39	6.4%	6.3%
40 - 44	6.2%	6.1%
45 - 49	6.4%	5.7%
50 - 54	6.2%	5.2%
55 - 59	5.4%	4.5%
60 - 64	4.3%	3.6%
65 - 69	3.5%	3.6%
70 - 74	2.3%	2.6%
75 - 79	1.5%	1.7%
80 - 84	0.9%	1.0%
85 - 89	0.3%	0.6%



characteristic	optum_ehr (n = 95,609)	optum_extended_dod (n = 123,886)
Age group		
0 - 4	6.0%	7.6%
5 - 9	6.0%	6.7%
10 - 14	6.4%	5.9%
15 - 19	8.1%	6.3%
20 - 24	7.1%	5.0%
25 - 29	7.0%	6.0%
30 - 34	6.9%	6.9%
35 - 39	7.1%	7.5%
40 - 44	7.1%	7.7%
45 - 49	7.5%	7.5%
50 - 54	7.4%	7.2%
55 - 59	7.1%	6.4%
60 - 64	5.6%	5.3%
65 - 69	4.4%	5.1%
70 - 74	2.9%	3.8%
75 - 79	1.8%	2.5%
80 - 84	1.3%	1.6%
85 - 89	0.3%	0.9%
90 - 94		0.1%

When we compared the baseline characteristics of the two cohorts, we found them to be remarkably similar for most of the covariates (see scatter plot below of covariate means) except for the external exposures (green color outliers). This observation indicates that the two definitions are identifying people with comparable baseline characteristics. This improved our confidence in the generalizability of the definition.

Figure 3: Comparison of baseline characteristics between cohorts



We then reviewed the temporal characteristics of the cohorts in the period immediately prior to and not including index date (-30days to -1 days). We found that some proportions of persons were exposed to drugs commonly used to treat anaphylaxis, before they had the diagnosis, e.g. epinephrine was 8.9%. This although lower than observed proportion of > 50% epinephrine use on index date, suggests that some people in the cohort may have index date misclassification. In specific, the true index date may be prior that that defined by the definition. This index event misclassification in an acute outcome like anaphylaxis may lead to significant misclassification bias in time to event studies. Researchers may need to develop methods to reduce such biases.

Conclusion

We were able to improve our confidence in our anaphylaxis cohort definition using cohort diagnostics. The tool gathers and explores empirical evidence that is key metrics to understand different element of phenotype implementation on various data sources. The included descriptive diagnostics on levels of data, codes and cohorts can be used as a proxy for evaluating sensitivity/specificity errors tradeoffs. These

empirical insights can help researchers identify the impact of choices they make for any given phenotype. We encourage the use of Cohort Diagnostics in OHDSI network studies as a best practice.

Reference/citations

1. Observational Health Data Sciences and Informatics. (2021). *Cohort Diagnostics*. Cohort Diagnostics: An R package for performing various cohort diagnostics. Retrieved June 15, 2021, from https://ohdsi.github.io/CohortDiagnostics/

2. Observational Health Data Sciences and Informatics. (2021). *HADES*. Health Analytics Data-to-Evidence Suite (HADES): A collection of R packages for performing analytics against the Common Data Model. Retrieved June 15, 2021, from https://ohdsi.github.io/Hades/

3. Observational Health Data Sciences and Informatics. (2021). *Ohdsi 10 Minute Tutorial CohortDiagnostics Anaphylaxis,* Ohdsi 10 Minute Tutorial CohortDiagnostics Anaphylaxis. Retrieved June 15, 2021, https://data.ohdsi.org/Ohdsi10MinuteTutorialCohortDiagnosticsAnaphylaxis/