Use of CohortDiagnostics for evaluating a phenotype of acute-on-chronic hepatic failure

Bürgler, Alexandra[1]; Burrows, Evanette[2,3]; Titaevski, Serge[4]; Swerdel, Joel[2,3]

1. Global Epidemiology, Johnson & Johnson Innovative Medicine, Gewerbestrasse 16, CH-4123, Allschwil, Switzerland
2. Observational Health Data Analytics, Global Epidemiology, Johnson & Johnson Innovative Medicine, Titusville, NJ, USA
3. Observational Health Data Sciences and Informatics (OHDSI), New York, NY, USA
4. Global Medical Organization, Johnson & Johnson Innovative Medicine, Hegenheimermattweg 167C, CH-4123, Allschwil, Switzerland

Background

Acute hepatic failure with prior chronic liver disease, also known as acute-on-chronic hepatic failure (ACHF), is a sudden, life-threatening deterioration of liver function and therefore an important safety event of interest in pharmacovigilance[1]. Thus, ACHF is a common safety event of interest when evaluating drug safety. We aimed to create a clinical algorithm (phenotype) for drug-related ACHF (ACHF phenotype), meant to contrast a previously developed phenotype of acute hepatic failure in a population without prior chronic liver disease (AHF phenotype, minimally adapted from Shoaibi, Rao[2]).

Methods

We followed the steps of 1) developing a clinical description of ACHF (using generative AI followed by medical review), 2) conducting a literature review of previously developed ACHF phenotypes to include relevant concepts, 3) developing the phenotype of ACHF with SNOMED terminology by adapting the pre-existing AHF phenotype and using open-source OHDSI tools ATLAS[3] and PHOEBE[4], and 4) reviewing the phenotype with the global medical safety team. The package CohortDiagnostics (version 3.2.5)[5] was used in R (version 4.3.2)[6] to evaluate both phenotypes (ACHF and AHF) regarding the capture of key clinical characteristics and to determine whether the two phenotypes were distinctive (overlap analysis) in seven observational databases standardized to the OMOP common data model version 5.4. The seven databases included six health claims databases and one electronic health record database from the US and Japan (Table 1).

Results

The index event of the AHF phenotype was defined as a condition occurrence of acute hepatic failure with 1) no chronic liver disease any time prior to index (AHF diagnosis), 2) no liver transplant any time prior to index, 3) no viral hepatitis or sequelae thereof, no alcoholic hepatitis or sequelae seven days before until seven days after index (a number of days chosen to account for variability of code entry) and 4) no hepatorenal syndrome

on day of diagnosis of acute hepatic failure. The AHF phenotype was adapted to create the ACHF phenotype by replacing criterion (1) with the inclusion of chronic liver disease any time prior to index. All other inclusion criteria (2-4), end date and cohort collapse strategy were kept the same as for the AHF phenotype.

Assessing both phenotypes (ACHF and AHF) in CohortDiagnostics, an overlap of 67.8% to 75.7% between the two cohorts was found across the seven databases (Figure 1). All patients in the AHF cohort were also included in the ACHF cohort, accounting for 67.8% to 75.7% of all captured patients. The remaining 24.3% to 32.2% of patients were only part of the ACHF cohort. Investigating this overlap, we found that it was caused by overlapping concepts in the inclusion criterion of prior chronic liver disease and the index event of acute hepatic failure, combined with the chosen time window for prior chronic liver disease. For example, the concept of "acute hepatic failure" was used as an index event but also as a proxy for prior chronic liver disease when occurring before index. Since the time window for the inclusion criterion of prior chronic liver disease allowed any time prior index until index date, everyone with an overlapping concept on the index date was wrongly assigned chronic liver disease. Adjustment of the time window of prior chronic liver disease to "until one day prior to index date" reduced the overlap to 7.9-14.7%.

After reduction of the overlap, the person count in the ACHF cohort decreased from 378,554 to 142,885 in the Health Verity Comprehensive Claims database and from 6,699 to 2,827 in the Japan Medical Data Center (JMDC) database. Comparing the ACHF and AHF cohort, the case count ranged from 2,837 to 142,885 (ACHF) and from 6,699 to 378,554 (AHF) in JMDC and Health Verity Comprehensive Claims databases, respectively. Across all databases, and consistent for both phenotypes, males had higher incidence than females and incidence was higher in older age. Typical symptoms before index presented more frequently in the ACHF cohort than in the AHF cohort (e.g., 82% versus 40% with pain and indication for inflammation in abdominal area).

A remaining overlap of 7.9–14.7% primarily reflects differences in the definition of the chronic liver disease concept set between the AHF and ACHF cohorts. Specifically, the ACHF definition employs a broader concept set, including 405 codes that cover conditions such as "Toxic Liver Disease," "Steatosis of the liver," and "Chronic passive congestion of the liver." In contrast, the AHF definition includes only seven codes, such as "Hepatic coma," "End-stage liver disease," and "Cirrhosis of the liver." These differences arise because concept set design is typically tailored to the specific clinical or research question, leading to variability in how conditions are defined and included.

Conclusions

CohortDiagnostics is an important tool for understanding possible errors of cohort definitions. This study highlights that when creating distinct phenotypes with

overlapping concept sets, special attention should be paid to selecting the appropriate time windows, and carefully observing the cohort overlap in CohortDiagnostics to detect any potential misclassification.

Table 1: databases

Databases	Years	Country	Data type	Details	Number of captured persons in ACHF phenotype / total persons
Merative® MarketScan Commercial Claims and Encounters (CCAE)	01/2000 – 01/2025	United States	Health insurance claims	Commercial claims for the working-age population	32,457 / 176,034,685
Health Verity Comprehensive Claims	06/2009 – 03/2025	United States	Health insurance claims	De-identified medical, pharmacy, and death data from closed claims and open claims	142,885 / 228,757,497
Merative® MarketScan Multi-State Medicaid Database (MDCD)	01/2006 – 06/2024	United States	Health insurance claims	Claims data for Medicaid populations (low-income)	31,463 / 36,726,487
Optum® Electronic Health Record dataset (Optum EHR)	01/2007 – 12/2024	United States	Electronic health records (EHR)	Inpatient and outpatient EHR from healthcare provider organizations	65,892 / 124,583,412
Optum's Clinformatics® Data Mart – Socioeconomic status (Optum SES)	05/2000 – 12/2024	United States	Health insurance claims	Claims data for commercial (<65 years) and Medicare Advantage health plan populations (≥65 years), including race, ethnicity, SES	71,023 / 102,881,984
Merative® MarketScan Medicare Supplemental (MDCR)	01/2000 – 01/2025	United States	Health insurance claims	Claims for retirees (aged ≥65 years) with Medicare supplemental coverage through employer-sponsored plans	15,441 / 11,394,087
Japan Medical Data Center (JMDC)	01/2005 – 09/2024	Japan	Health insurance claims	Claims data from 250 Health Insurance Associations covering workers aged < 75 years old and their dependents	2,827 / 22,314,340

Figure 1: cohort overlap

Original overlap

Target database (T): AHF
Control database (C): ACHF

Database Name	T Only	C Only	Both	Total Subjects
Health Verity Comprehensive Claims - Closed Claims Enrollment	0.0%	24.3%	75.7%	378,554
Optumâ€™s ClinformaticsÂ® Extended Data Mart â€“ Socio-Economic Status (SES)	0.0%	27.1%	72.9%	181,461
Optum EHR	0.0%	28.9%	71.1%	173,673
Merative MarketScanÂ® Commercial Claims and Encounters Database	0.0%	29.3%	70.7%	75,715
Merative MarketScanÂ® Multi-State Medicaid Database	0.0%	27.5%	72.5%	75,058
Merative MarketScanÂ® Medicare Supplemental and Coordination of Benefits Database	0.0%	32.2%	67.8%	38,659
Japan Medical Data Center (JMDC)	0.0%	29.6%	70.4%	6,699

Overlap resolved

Target database (T): AHF
Cohort database (C): ACHF

Database Name	T Only	C Only	Both	Total Subjects
Health Verity Comprehensive Claims - Closed Claims Enrollment	62.0%	23.8%	14.2%	376,047
Optumâ€™s ClinformaticsÂ® Extended Data Mart â€“ Socio-Economic Status (SES)	60.8%	27.0%	12.2%	181,292
Optum EHR	62.1%	28.9%	9.0%	173,673
Merative MarketScanÂ® Commercial Claims and Encounters Database	57.1%	29.2%	13.8%	75,579
Merative MarketScanÂ® Multi-State Medicaid Database	58.0%	27.3%	14.7%	74,827
Merative MarketScanÂ® Medicare Supplemental and Coordination of Benefits Database	60.0%	32.1%	7.9%	38,569
Japan Medical Data Center (JMDC)	57.8%	29.6%	12.6%	6,699

References

1. Sarin SK, Choudhury A. Acute-on-chronic liver failure: terminology, mechanisms and management. *Nature reviews Gastroenterology & Hepatology* 2016; **13**(3): 131-49. https://doi.org/10.1038/nrgastro.2015.219.
2. Shoaibi A, Rao G, Swerdel JN, et al. Development and evaluation of phenotype algorithms for the European Medicines Agency designated medical events in a

network of real-world data. In *Abstracts of ICPE 2023*, the 39th International Conference on Pharmacoepidemiology and Therapeutic Risk Management (ICPE), Halifax, Canada, 25–27 August 2023. *Pharmacoepidemiology and Drug Safety* 2023; **32**(S1): 3-612. https://doi.org/10.1002/pds.5687.

3. ATLAS: open source software tool for researchers to conduct scientific analyses on standardized observational data. https://github.com/OHDSI/Atlas.
4. Ostropolets A, Ryan P, Hripcsak G. Phenotyping in distributed data networks: selecting the right codes for the right patients. *AMIA Annual Symposium Proceedings* 2022: 826-35. PMC10148372.
5. Rao GA, Shoaibi A, Makadia R, et al. CohortDiagnostics: Phenotype evaluation across a network of observational data sources using population-level characterization. *PLoS One* 2025; **20**(1): e0310634. https://doi.org/10.1371/journal.pone.0310634.
6. R Core Team. R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2022. https://www.R-project.org/.