

Synthetic Creation of OMOP Oncology Patients (SCOOP)

Brianna Argall¹, Ghazaleh Niknam Shirvan^{1,2}, Tim Churches^{1,2}, Frank Lin^{1,3}, Jeremy L Warner⁴, Georgina Kennedy^{1,2}

1 School of Clinical Medicine, University of New South Wales, Sydney, Australia,

2 Ingham Institute for Applied Medical Research, Sydney, New South Wales, Australia,

3 Garvan Institute, Sydney, New South Wales, Australia,

4 Brown University and Rhode Island Hospital, Providence, RI, USA.

Background

Developing and testing analytical tools for the OMOP Common Data Model is greatly facilitated by access to realistic synthetic datasets that match the CDM's structure. The OHDSI community has invested in the oncology extension to the OMOP CDM, including the addition of `episode` and `episode_event` tables to provide structured context to capture disease progression as well as linking together complex multi-modal treatment events. However, no publicly available synthetic dataset currently populates these oncology tables with realistic treatment variation.

Existing synthetic resources such as Eunomia¹ and Synthea² provide sample OMOP CDM datasets that serve well for education and software testing across core CDM tables, but do not cover oncology extension features. Synthea does support generation of longitudinal synthetic patient records that includes cancer modules, however oncology treatment breadth is limited.

One approach to richer synthetic oncology data would be to extend existing synthetic patient generators with more detailed cancer modules. However, oncology treatment regimens follow precisely structured protocols with component-specific dosing, irregular scheduling, and cycle-level variation that differs between variants of nominally similar regimens. Synthea's lung cancer module illustrates the ceiling of this approach, comprising one fixed chemotherapy regimen per histological type, with no variation by stage, biomarker status, or line of therapy. Extending this by hand would require rebuilding what the HemOnckB already provides systematically across thousands of curated, evidence-backed, continually updating regimen variants. The HemOnc Alchemy interface has demonstrated the integration of broader HemOnckB data to OMOP CDM structures³. We therefore build upon this existing work to deliver SCOOP (Synthetic Creation of OMOP Oncology Patients)⁴ around HemOnckB as the treatment backbone.

Methods

We developed Synthetic Creation of OMOP Oncology Patients (SCOOP), an open-source Python package that generates detailed and realistically variable synthetic oncology patient cohorts with records that map natively to OMOP CDM tables, with a focus on population of the episode-related tables.

SCOOP organises patient generation into four sequential phases:

Sampling	Patient attributes are drawn from configurable statistical distributions. At the end of this stage characteristics of the patient at diagnosis are defined, including demographics, molecular markers, performance status, and disease stage at presentation. We also declare the calendar year of diagnosis to take advantage of the detailed HemOnckB historical information, allowing the generated data to include shifts in practice that can then be rendered in downstream systems, which is a very useful validation point for a number of use-cases and is not available in other
-----------------	--

	comparable synthetic datasets.
Simulation	A clinical decision graph routes each patient to a treatment endpoint based on their attributes and the applicable clinical guideline structure. These graphs are configured per disease and per clinical subgroup, encoding the decision factors (stage, biomarker status, histology, eligibility criteria), or reasonable proxies thereof, that typically guide treatment decisions.
Synthesis	Each patient is assigned a concrete collection of sigs drawn from the HemOnckB and consistent with their treatment endpoint. Dosing and cycle timing are derived directly from HemOnckB's curated protocol instructions, with realistic day-level variation applied to reflect scheduling that occurs in clinical practice.
Serialisation	Patient records are written to OMOP CDM tables using standard OMOP concept identifiers and conforming to oncology-specific CDM conventions.

Each disease is represented as a profile: a set of YAML configuration files that define the clinical decision structure and the mapping from treatment endpoints to HemOnckB-sourced sigs. The profile design deliberately separates these concerns. The decision graph answers the question "what type of treatment does this patient receive?" The protocol registry answers "how is that treatment type operationalized into a collection of sigs?" This means clinical review of decision logic and clinical review of regimen mapping can occur independently.

Treatment endpoints connect to HemOnckB via two paths. For systemic therapies, a component query retrieves HemOnckB regimen variants matching specified agent class and category requirements. For treatment structures that are not naturally represented as a single HemOnckB variant (e.g. surgery with adjuvant or neoadjuvant systemic therapy, multi-phase protocols) a bundle assembly path retrieves each component separately and assembles them into a clinically ordered sequence. Both paths populate downstream OMOP records using actual HemOnckB sig data, ensuring that drug exposure timing reflects protocol schedules rather than uniform approximations.

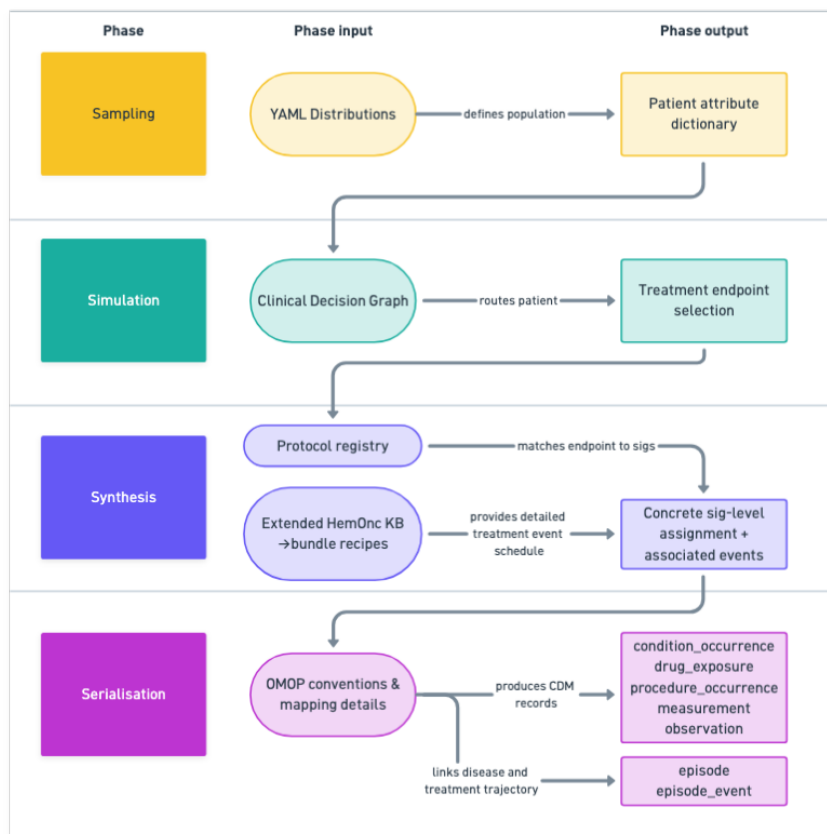
Results

SCOOP delivers a four-phase generation pipeline packaged as an open-source Python library, with disease profiles defined entirely through YAML configuration files. The package core handles graph execution, HemOnckB retrieval, and OMOP serialisation; the disease-specific clinical knowledge lives in the profile files and requires no changes to core code to add or modify.

A serialised patient record includes rows across `condition_occurrence`, `drug_exposure`, `procedure_occurrence`, `measurement`, `observation`, `episode`, and `episode_event`; the `episode` and `episode_event` tables are populated directly from the regimen structure, encoding treatment type, component sequence, and cycle-level linkage without post-hoc inference from drug exposure records alone.

The most developed disease profile is non-small cell lung cancer (NSCLC), modelled against 2026 ESMO living guidelines. It covers five clinical subgroups: stage I, resectable stages II–III, unresectable stage III, stage IV with actionable driver alterations, and stage IV without actionable drivers. Each subgroup has its own transformation graph and protocol registry, and the query-based variant path is fully executable end to end for systemic therapy endpoints. The NSCLC regimen registry contains 36 public entries

spanning all five subgroups across stage I (local), resectable stages II–III, unresectable stage III, stage IV with driver mutations, for stage IV with no driver mutations, plus a further 8 internal entries used in composite resolution. Of the 36 public entries, 19 use the component query path (direct HemOnckB variant retrieval by agent class and category), 12 use bundle assembly (chemoradiotherapy, surgery-plus-adjuvant, and sequential protocols), and 5 are composites that resolve to a union of component sub-entries.



Adding a new disease profile requires four YAML artifacts: a profile manifest, one or more transformation graphs encoding clinical decision logic, a protocol registry defining the canonical treatment endpoint vocabulary, and resolver configurations connecting each endpoint to HemOnckB.

Conclusion

SCOOP addresses a specific gap in the OMOP CDM synthetic data ecosystem: the absence of realistic oncology data across the episode-related tables introduced by the oncology extension. By grounding synthetic treatment records directly in HemOnckB, SCOOP produces data with genuine regimen fidelity — correct agents, schedules, and dosing — rather than simplified hand-coded approximations, and does so for a treatment space that continues to evolve as HemOnckB is updated.

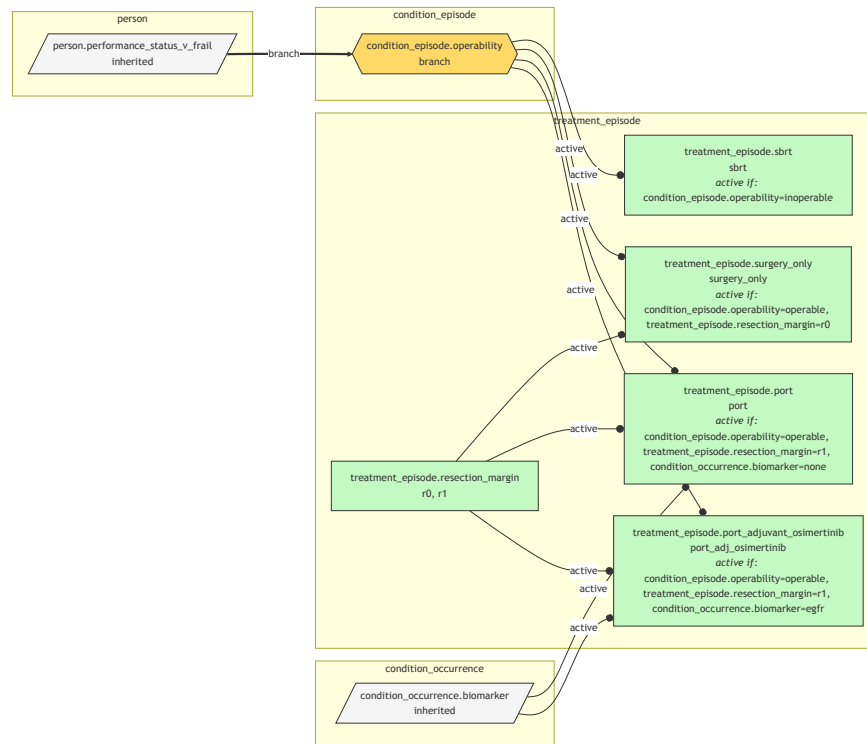
The package is designed to be extensible: adding a new disease profile requires only configuration files describing the clinical decision structure and connecting treatment endpoints to HemOnckB; no changes to the core package are required.

References

1. DeFalco F, Schuemie M, Sena A, Adulyanukosol N, Liu S, Black A. Eunomia: Standard Dataset Manager for Observational Medical Outcomes Partnership Common Data Model Sample Datasets, <https://ohdsi.github.io/Eunomia/>
2. Walonoski J, Kramer M, Nichols J, et al. Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record. *J Am Med Inform Assoc.* 2018;25(3):230–238. doi:10.1093/jamia/ocx079
3. Kennedy G, Zack T, Cerelia Valerie I, Gurley M, Warner JL. Interface for description and analyses of systemic oncology protocols. OHDSI Global Symposium; October 2024;
4. Argall, B., Niknam Shirvan G., Churches, T., Kennedy, G., SCOOP: Synthetic Creation of OMOP Oncology Patients <https://github.com/AustralianCancerDataNetwork/SCOOP/>

NSCLC:

Stage I



Stage II-III Resectable

