Development and validation of generative-model-derived synthetic cohorts for outcome prediction and comparative effectiveness research in chronic lymphocytic leukemia (CLL)

Authors:

Talha Munir[1], Paolo Sportoletti[2], Leyla Mohseninejad[3], Sheng Xu[4], Igor Chebuniaev[5], Samuel Aballéa[5], Rami Zouari[5], Noemi Kreif[6], Mondher Toumi[7], Lianne Barnieh[8], Marcus Lefebure[8], Rhys Williams[9], Remus Vezan[9], Eric Chen[10], Iwona Hus[11], Anne-Sophie Michallet[12], Anna Maria Frustaci[13], Loïc Ysebaert[14], Mazyar Shadman[15]

Affiliations:

[1]LEEDS TEACHING HOSPITALS NHS TRUST Leeds, United Kingdom; [2]University of Perugia and Santa Maria della Misericordia Hospital Perugia, Italy; [3]BeOne Medicines Netherlands B.V. Schiphol, Netherlands; [4] BeOne Medicines Shanghai, China; [5]InovIntell, Krakow, Poland; [6]University of Washington School of Pharmacy, Seattle, WA, United States; [7]Aix-Marseille University, Marseille, France; [8]BeOne Medicines UK, Ltd London, United Kingdom; [9]BeOne Medicines, Ltd San Carlos, CA, United States; [10]BeOne Medicines, Ltd, Cambridge, MA, United States; [11]National Medical Institute of the Ministry of the Interior and Administration, Warsaw, Poland; [12]Centre Léon Bérard Lyon, France; [13]ASST Grande Ospedale Metropolitano Niguarda, Milan, Italy; [14]IUCT Oncopole Toulouse, France; [15] Fred Hutchinson Cancer Center Seattle, WA, United States

Review Category: 803. Emerging Tools, Techniques, and Artificial Intelligence in Hematology

Keywords: Diseases, Lymphoid Malignancies, Lymphoid Leukemias, CLL, Artificial Intelligence(AI), Generative AI, Machine Learning

Sub-category: Health Services, Quality Improvement, or Outcomes Research

Character count: 4,381/4,500

Introduction

Artificial intelligence (AI) generated synthetic patient cohorts have the potential to transform evidence generation by creating realistic synthetic patient cohorts that may be used to support a range of research applications. Realizing this potential requires demonstrating that synthetic data can accurately reproduce observed outcomes across diverse patient populations and generate reliable comparative-effectiveness estimates.

We developed a generative machine-learning model using individual patient data (IPD) from phase 3 zanubrutinib trials to simulate complete patient trajectories in chronic lymphocytic leukemia (CLL) and evaluated its ability to reproduce outcomes and support comparative effectiveness across patient subgroups.

Methods

A latent ordinary-differential-equation generative model was trained using IPD from SEQUOIA (treatment-naive CLL; NCT03336333) and ALPINE (relapsed/refractory CLL; NCT03734016). In addition to modelling survival outcomes, the model learned longitudinal disease trajectories, including the evolution of laboratory measurements over time and their relationship with progression and death conditional on baseline characteristics, enabling simulation of complete patient trajectories for specified baseline profiles.

Validation was conducted in two steps. First, synthetic and observed Kaplan–Meier (KM) curves were compared across clinically relevant patient subgroups. Analyses included subgroups used during fine-tuning (del(17p)/TP53, IGHV, complex karyotype, prior treatment and β2-microglobulin status) and subgroups defined by characteristics not used for model fine-tuning (age, Binet stage, bulky disease, and gender), with baseline profiles generated separately from the model. The gap between synthetic and observed KM curves was expressed in terms of median absolute difference (AD).

Second, the utility of the synthetic cohort for comparative analyses was evaluated from a comparison of synthetic data generated from SEQUOIA for comparison with AMPLIFY (NCT03836261). Resulting treatment-effect estimates were compared with previously published estimates obtained using matching-adjusted indirect comparison (MAIC) (Shadman et al, 2026, Blood Advances) and network meta-analysis (NMA) (Shadman et al, poster presented at SOHO 2025). Investigator-assessed progression-free survival (PFS) was the primary endpoint; OS was also explored. Uncertainty was quantified using bootstrapping.

Results

Across subgroups, model-generated KM curves closely reproduced observed outcomes. The median AD in PFS was 0.016, and approximately 92% of each synthetic 95% confidence interval (CI) was contained within the bounds of the corresponding observed CI for both PFS and OS. Agreement was maintained in subgroups not used for model fine-tuning (median PFS AD 0.014; approximately 94% CI containment), with larger deviations limited to smaller subgroups and sparsely populated tails of survival curves.

For the comparative analysis, population adjustment aligned the synthetic bendamustine-rituximab (BR) arm with the AMPLIFY chemoimmunotherapy control. At

36 months, the model-generated BR arm closely reproduced the AMPLIFY control (66.9% vs 65.9% PFS). The comparison of synthetic generated zanubrutinib versus AV yielded a PFS hazard ratio (HR) of 0.34 (95% CI 0.21–0.55). This estimate was consistent with a previously published NMA (HR 0.46; 95% CI 0.28–0.76) and a MAIC (HR 0.45; 95% CI 0.23–0.88 without age adjustment, and HR 0.26; 95% CI 0.13–0.54 with age adjustment), while producing narrower CIs with the synthetic cohort (CI width:0.34) vs. the MAIC (range 0.41–0.65) and the NMA (0.48).

Conclusions

An AI-enabled framework generated CLL cohorts that reproduced survival outcomes not only in the subgroups used for tuning but also in subgroups defined along dimensions outside of fine-tuning, using synthetic patient profiles not seen in training. Treatment-effect estimates generated from synthetic data were consistent with previously published NMA and MAIC analyses while offering greater precision. These findings support the use of synthetic patient cohorts for indirect treatment comparisons and external control arms and highlight broader future applications including subgroup outcome prediction, virtual cohort generation, evaluation of hypothetical clinical scenarios, and modelling of treatment sequences.