



INTRODUCTION

- In the alleged ‘age of big data’, publicly available large data sets are still scarce.
- Data sharing is often hindered by concerns of patient privacy, regulatory aspects, and proprietary interests.
- Unveiling new mechanisms of disease, developing molecularly targeted therapies and designing clinical machine learning algorithms all require access to large data sets.

AIM

- to generate synthetic patient data for acute myeloid leukemia trials
- to compare real and synthetic data regarding clinical, laboratory, cytogenetic, molecular genetics, and survival data
- to assess the fidelity and usability of synthetic data for exploratory analyses and clinical trial design

CONCLUSIONS

- Synthetic patient data mimics clinical, laboratory, cytogenetic and molecular genetic properties of real multicenter trial patients.
- Inter-variable dependencies and survival analysis is feasible in synthetic cohorts enabling investigations and synthetic trial controls.
- Synthetic patient cohorts overcome current hinderances in data sharing and hold the promise of synthetic control arms in future clinical trials.

MIMICKING CLINICAL TRIALS WITH SYNTHETIC ACUTE MYELOID LEUKEMIA PATIENTS USING GENERATIVE ARTIFICIAL INTELLIGENCE

J.-N. Eckardt^{1,2}, W. Hahn^{3,4}, C. Röllig¹, S. Stasik¹, U. Platzbecker⁵, C. Müller-Tidow⁶, H. Serve⁷, C.D. Baldus⁸, C. Schliemann⁹, K. Schäfer-Eckart¹⁰, M. Hanoun¹¹, M. Kaufmann¹², A. Burchert¹³, C. Thiede¹, J. Schetelig¹, M. Sedlmayr⁴, M. Bornhäuser^{1,14,15}, M. Wolfien^{3,4} and J.M. Middeke^{1,2}

1 Department of Internal Medicine I, University Hospital Carl Gustav Carus, Technical University Dresden, Dresden, Germany; 2 Else Kröner Fresenius Center for Digital Health, Technical University Dresden, Dresden, Germany; 3 Center for Scalable Data Analytics and Artificial Intelligence (ScaDS.AI) Dresden/Leipzig, Germany; 4 Institute for Medical Informatics and Biometry, Technical University Dresden, Dresden, Germany; 5 Medical Clinic and Policlinic I Hematology and Cell Therapy, University Hospital, Leipzig, Germany; 6 Department of Medicine V, University Hospital Heidelberg, Heidelberg, Germany; 7 Department of Medicine 2, Hematology and Oncology, Goethe University Frankfurt, Frankfurt, Germany; 8 Department of Hematology and Oncology, University Hospital Schleswig Holstein, Kiel, Germany; 9 Department of Medicine A, University Hospital Münster, Münster, Germany; 10 Department of Internal Medicine V, Paracelsus Medizinische Privatuniversität and University Hospital Nürnberg, Nürnberg, Germany; 11 Department of Hematology, University Hospital Essen, Essen, Germany; 12 Department of Hematology, Oncology and Palliative Care, Robert-Bosch-Hospital, Stuttgart, Germany; 13 Department of Hematology, Oncology and Immunology, Philipps-University-Marburg, Marburg, Germany; 14 German Consortium for Translational Cancer Research DKFZ, Heidelberg, Germany; 15 National Center for Tumor Diseases (NCT), Dresden, Germany

RESULTS

Synthetic clinical and genetic patient characteristics closely mimic those of real patients

- two equally sized (n=1606) synthetic data sets were generated
- compared to real data, synthetic data show minimal diversion regarding clinical, laboratory, molecular genetic, and cytogenetic variables
- In univariable analyses, effect sizes and effect directions are preserved for synthetic data effectively enabling exploratory analyses
- synthetic cohorts match real cohorts in outcome and survival analyses
- training sample patient privacy is safeguarded:
 - zero exact matches in synthetic cohorts
 - median number of variables per synthetic patient to be altered to match a real patient is nine

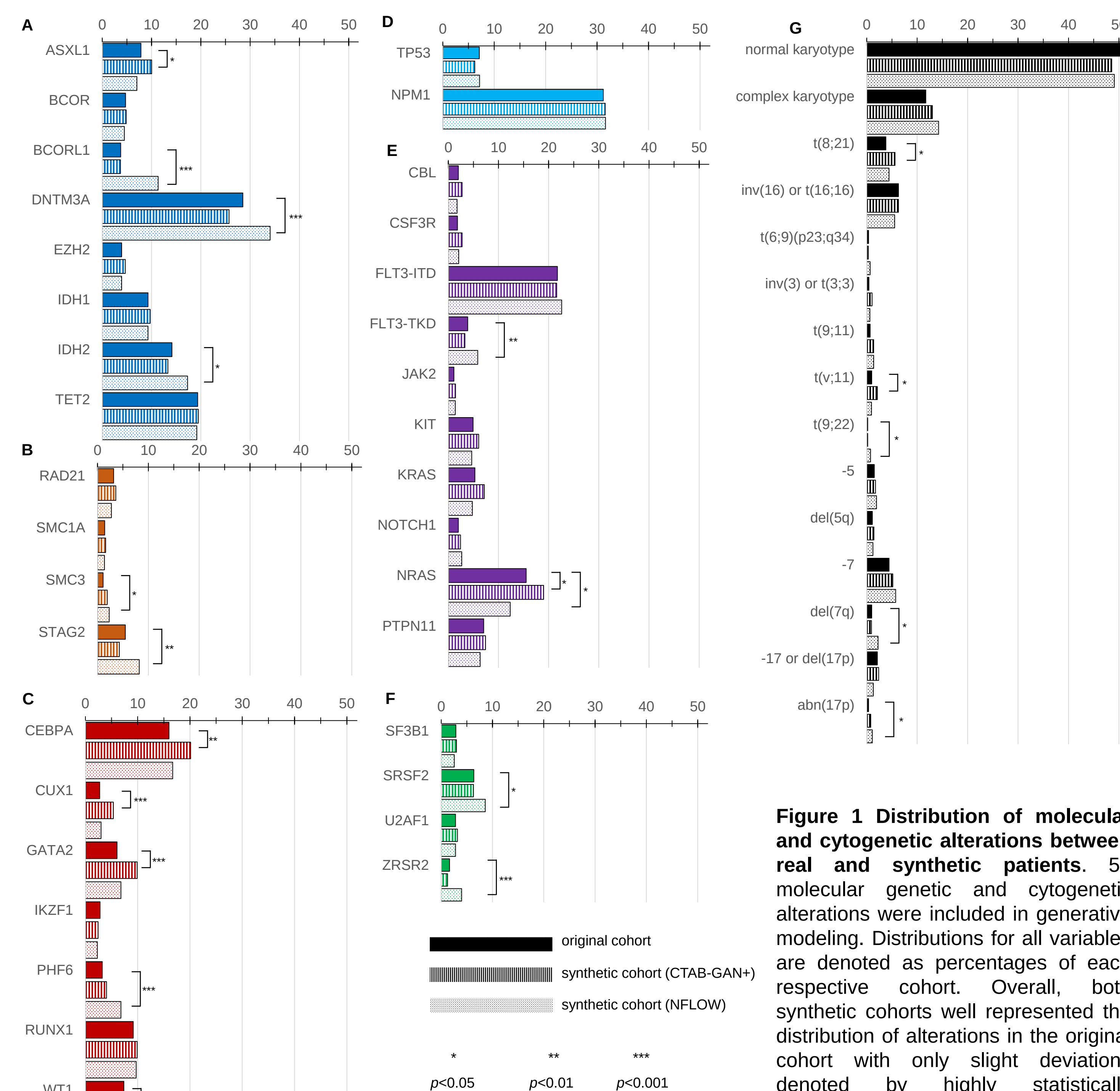


Figure 1 Distribution of molecular and cytogenetic alterations between real and synthetic patients. 50 molecular genetic and cytogenetic alterations were included in generative modeling. Distributions for all variables are denoted as percentages of each respective cohort. Overall, both synthetic cohorts well represented the distribution of alterations in the original cohort with only slight deviations denoted by highly statistically significant ($p<0.001$) differences in *BCORL1*, *DNMT3A*, *PHF6*, and *ZRSR2* for NFlow, as well as *CUX1* and *GATA2* for CTAB-GAN+.

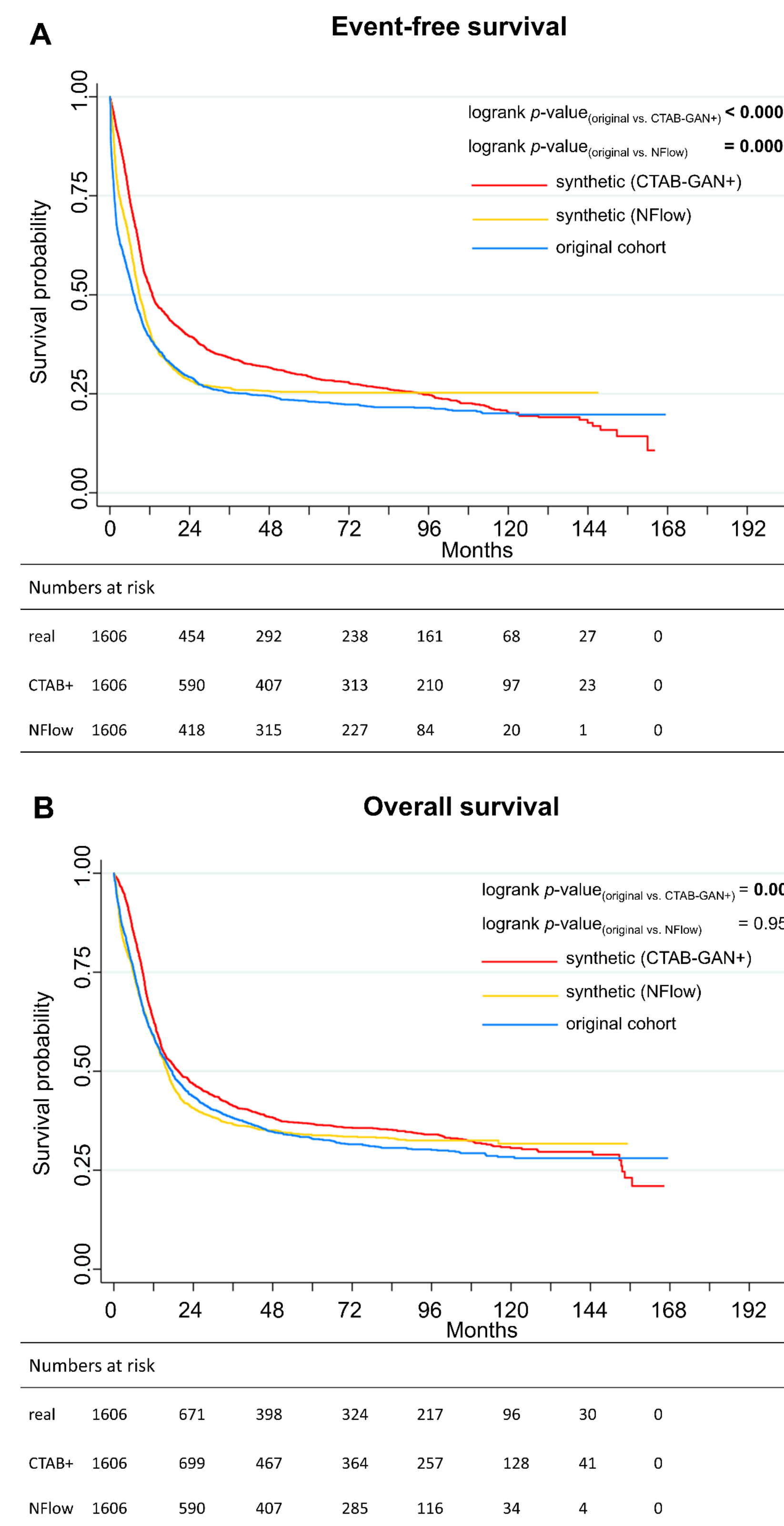


Figure 2 Comparison of survival curves between original and synthetic cohorts.

Event-free survival deviated significantly from the original cohort for both synthetic cohorts (A). For the NFlow-generated cohort, there was no significant deviation from the original distribution for overall survival, while the CTAB-GAN+-generated cohort again differed significantly (B). Interestingly, while the survival curve for CTAB-GAN+ displays a plausible curve up until ten years of follow-up, the curve shows no stabilization of survival rates in the end as the original cohort does. Contrastingly, the survival curve for NFlow shows an overall plausible course, however, NFlow tends to overtly censor patients after two years of follow-up.

METHODS

- 1606 AML patients from four previous multicenter clinical trials of the German Study Alliance Leukemia (SAL) were included
 - eligibility: non-APL AML, age ≥ 18 years, intensive induction therapy
 - training variables: clinical (n=4), laboratory (n=3), molecular genetics (n=35), cytogenetics (n=15), outcome (n=5)
- Two fundamentally different generative artificial intelligence technologies were used:
 - the generative adversarial network CTAB-GAN+
 - normalizing flows (NFlow)
- Hyperparameter tuning via Optuna
- no imputation of missing data
 - missing data is included in training and generated via AI in synthetic cohorts
- Patient privacy conservation was measured using Hamming distances.
 - the average minimum distance effectively denotes the number of variables that one would need to alter for a synthetic patient to match a real patient

ACKNOWLEDGEMENTS

We thank all contributing physicians, laboratories, and nurses associated with the German Study Alliance Leukemia and especially participating patients for their valuable contributions.

CONTACT INFORMATION

email: jan-niklas.eckardt@uniklinikum-dresden.de
phone: +49 (0) 351 458 11542
web: www.ai-in-cancer.org