

Synthetic data reliably reproduces brain tumor primary research data

Roy Khalaf

McGill University

William Davalan

McGill University

Amro H. Mohammad

McGill University

Roberto Jose Diaz (✉ roberto.diaz@mcgill.ca)

McGill University

Research Article

Keywords: Synthetic data, Neuro-oncology, Brain tumors

Posted Date: January 3rd, 2024

DOI: <https://doi.org/10.21203/rs.3.rs-3800842/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Purpose

Synthetic data has garnered heightened attention in contemporary research due to confidentiality barriers and its capacity to simulate variables challenging to obtain, notably in cases where premature death prevents adequate follow-up. Indeed, a significant challenge in clinical neuro-oncology research is the limited availability of data pertinent to rapid-onset conditions with relatively poor prognoses. This study aimed to evaluate the reliability and validity of synthetic data in the context of neuro-oncology research, comparing findings from two published studies with results from synthetic datasets.

Materials and Methods

Two published neuro-oncology studies focusing on prognostic factors were selected, and their methodologies were replicated using *MDCclone* Platform to generate five synthetic datasets for each. These datasets were assessed for inter-variability and compared against the original study results.

Results

Findings from synthetic data consistently matched outcomes from both original articles. Reported findings, demographic trends and survival outcomes showed significant similarity ($P < 0.05$) with synthetic datasets. Moreover, synthetic data produced consistent results across multiple datasets.

Conclusion

Integrating synthetic data into clinical research offers excellent potential for providing accurate predictive insights without compromising patient privacy. In neuro-oncology, where data fragmentation and patient follow-up pose significant challenges, the adoption of synthetic datasets can be transformative.

BACKGROUND AND SIGNIFICANCE

Recent advancements in computational algorithms and machine learning applications have brought synthetic data generation to the forefront of clinical research.[1] Within the healthcare sector, synthetic datasets may offer solutions to the challenges presented by confidentiality considerations.[2] However, there is currently a limited number of studies reporting on its validity and potential for extrapolation.[3] Furthermore, a lack of comparative analysis exists between statistical results derived from synthetic data and real data. A comprehensive validation process is required to establish the suitability of synthetic health data. This process should include meaningful clinical questions and encompass various data types and outcomes.

Synthetic data emulates the statistical characteristics and time-dependent properties of the original dataset, while omitting defining information to uphold patient privacy.^[1] It is generated through statistical modelling of the original data and enables analysis of simulations and predictive models. When utilizing synthetic data, users should prioritize a holistic approach to analysis, as it emphasizes probabilistic patterns rather than individual profiles.^[4] Indeed, many synthetic data tools impose a minimum cohort size requirement to ensure meaningful statistical analysis.

Synthetic datasets are particularly useful in clinical oncology research. In this field, inadequate follow-up time is a challenge due to multiple factors, such as premature death, limiting access to useful patient information. Another challenge to data completeness arises when laboratory results and documentation are distributed across different hospital systems. These considerations prevent proper analysis of factors with a potential association to patient survival. With synthetic data, researchers can access consolidated datasets that may be tailored to meet pre-determined sample size criteria.

In this study, we sought to determine the capacity of a synthetic dataset to validate the results of two published studies in the neuro-oncology literature. We also explored the limitations of the current synthetic data structure on data analysis and interpretation. Finally, we give guidance on future data structure design to enhance clinical oncology research.

METHODS

MDCclone Platform. This study focused on assessing the validity of synthetic data generated by the *MDCclone* Platform (Beer-Sheba, Israel). *MDCclone* is a platform allowing researchers and healthcare professionals to securely access, query, and analyze large-scale medical datasets. This project was accomplished through a collaboration with Sheba medical center, the largest tertiary hospital in Israel. External users are granted access to the synthetic iteration of this data lake, designed to ensure the preservation of patient confidentiality, while affording unrestricted entry to a diverse spectrum of patient profiles.

MDCclone platform enables users to query data longitudinally by defining a reference event (population of interest) and additional time-related events (TRE) to further include additional data elements. An example of a reference cohort is patients diagnosed with glioblastoma. This will be the foundation for that query. Every subsequent event defined is affiliated with the reference cohort. TRE can encompass detailed information on the use of specific medications or additional factors, such as laboratory test results or surgical interventions.

The user has the option to select precise demographic information about the population. Variables such as age, gender, and country of origin are included, with identifying data being censored. The cohort is defined and the features are selected based on *MDCclone* synthetic data statistical models ^[5] to transform real patient groups into synthetic derivatives.

The algorithm designed by *MDCclone* operates on a multivariate basis, generating variables using a covariance measure. Statistical models are developed by grouping similar patients and employing a

modified version of kernel density estimation to determine the multivariate probability density. [5] This approach ensures the preservation of univariate and multivariate relationships within data subsets and ensures patient anonymity, as the irreversible nature of the process prevents the possibility of re-identification.[5] The algorithm does not rely on predefined forms of underlying distributions, thus enabling the exploration of previously unrecognized relationships once the data is available. Linear and nonlinear relationships between the variables as well as the temporal arrangement of medical events are considered by the algorithm during the synthetic data generation. [3]

Each iteration of data synthesis derived from the same query produces a distinct cohort. However, it will retain analogous statistical characteristics. The synthetic dataset is then tabulated for user view.

Comparator Study Selection. Two published studies assessing prognostic factors relating to neuro-oncology were selected. The first study [6] focuses on serum albumin as an independent prognostic factor for post-operative survival in glioblastoma (GBM) patients. The second study[7] investigated serum inflammatory markers as potential prognostic indicators for survival following brain metastasis diagnosis. The two studies were reproduced on the *MDCOne* platform using the same inclusion criteria stipulated within each paper. Five synthetic datasets based on the initial query were reproduced for each article. These were compared to evaluate the stability and inter-variability of datasets generated from the same query.

Statistical Analysis. Categorical variables are displayed as frequencies and percentages, with the Chi-square test (χ^2) employed to assess statistical disparities between them. Continuous variables are presented as median $\pm$ standard deviation (SD) or ranges, and their statistical significance was determined using the Log-Rank test. Survival outcomes were evaluated using Kaplan-Meier survival analyses and T-test was used to determine statistical significance. Hazard ratios (HR) and odd ratios (OR) will be used to determine survival differences between subgroups of a population. The Kolmogorov-Smirnov test was used to examine variable distribution. Heterogeneity between synthetic datasets was evaluated using the ANOVA test. Statistical significance was defined as a P value less than 0.05.

RESULTS

Serum albumin and survival in GBM. Serum albumin, recognized as a prognostic marker in many acute and chronic illnesses, is associated with less favorable survival outcomes in multiple solid organ cancers. The study by Borg *et al.* [6] explored the predictive value of serum albumin for survival in the context of central nervous system tumors. Their study assessed the post-operative survival in 549 GBM patients. The patients were stratified into three distinct categories based on their pre-surgical serum albumin (ALB) concentrations: hypoalbuminemia (ALB < 30g/L), lower range of normal (ALB 30-40g/L), and upper range of normal (ALB > 40g/L). The study found a significant reduction in post-surgical survival rates in hypoalbuminemia patients (ALB < 30 g/L) compared to patients with normal albumin levels. Additionally, the survival duration for patients with lower normal serum albumin levels was substantially shorter than for those with upper-normal range albumin levels. The study hence concluded that the pre-surgical serum albumin level is a powerful predictor of survival in GBM patients.

Leveraging the Sheba data lake using the *MDC/one* platform, we identified 452 patients that fulfilled the inclusion criteria set by the study. The demographic characteristics of the synthetic population closely aligned with the real-world patient population, with a mean serum albumin of 34.7 ± 5.2 g/L in the original dataset and 35.6 ± 5.5 in the primary synthetic dataset (Table 1). The primary synthetic data points exhibited a normal Gaussian distribution consistent with the real dataset, as depicted in Fig. 1 (real data $n = 549$, $P = 0.06$; synthetic data $n = 452$, $P = 0.10$, Kolmogorov-Smirnov statistic $D = 0.06$). Furthermore, both the original and synthetic study cohorts demonstrated equivalent distribution among the three serum albumin categories (χ^2 $P = 0.99$).

Table 1
Patient characteristics in the original (Borg *et al.*) and synthetic datasets.

Characteristics	Real data (N = 1250)	Synthetic data 1 (N = 1320)	Synthetic data 2 (N = 1320)	Synthetic data 3 (N = 1320)	Synthetic data 4 (N = 1320)	Synthetic data 5 (N = 1320)	P values*
Age	60.1 ± 5.0	64.4 ± 11.3	64.4 ± 11.3	64.7 ± 11.1	64.4 ± 11.6	64.2 ± 11.6	0.98
Male:Female Ratio	1.58:1	1.79:1	1.75:1	1.81:1	1.78:1	1.69:1	0.98
Surgery	48.6%	36.1%	35.1%	36.9%	37.4%	37.8%	0.45
Biopsy	47.3%	63.9%	64.9%	63.1%	62.6%	62.2%	
Debulking	4.1%	0%	0%	0%	0%	0%	
Not ascertained							
Adjuvant therapy	36.6%	25.4%	28.1%	28.5%	27.7%	28.0%	1.00
None	48.0%	2.7%	3.3%	2.8%	3.0%	3.0%	
Radiotherapy	0%	30.8%	28.5%	29.0%	29.7%	29.0%	
Chemotherapy	15.3%	41.2%	40.1%	39.7%	39.6%	40.0%	
Chemoradiotherapy							
Number of Patients (Albumin < 30g/L)	15%	12%	12%	14%	13%	13%	0.91
Number of Patients (Albumin 30-40g/L)	68%	72%	70%	69%	70%	70%	0.95
Number of Patients (Albumin > 40g/L)	18%	16%	18%	17%	17%	17%	1.0
Median Albumin (g/L)	35 ± 5.2	36 ± 5.4	35 ± 5.5	36 ± 5.4	36 ± 5.5	36 ± 5.5	0.93
Mean Albumin (g/L)	34.7 ± 5.2	35.6 ± 5.4	35.4 ± 5.5	35.3 ± 5.4	35.4 ± 5.5	35.4 ± 5.5	0.93
Survival (months) (Median[range])	NA	13.0 [0.1–59.2]	12.8 [0.1–58.0]	13.0 [0.1–59.8]	13.0 [0.1–59.0]	12.4 [0.1–56.4]	0.90

*P-values indicate the probability that the value of the specific variable is not significantly different between synthetic datasets.

With regard to survival, our analysis of the primary synthetic data demonstrated a substantial correlation between a GBM patient's serum albumin levels and post-surgical survival ($P < 0.0001$) (Fig. 2).

Patients in the hypoalbuminemia group exhibited a significantly shorter post-operative survival period, with a median duration of 7.0 months (HR 2.27 [95% CI 1.55–3.31], $P < 0.01$). Moreover, the patients with hypoalbuminemia experienced increased peri-operative (≤ 1 year post-operatively) mortality (61.1% vs. 26.1%; OR 4.18 [95% CI 2.20–7.99], $P < 0.01$). In contrast, the groups with lower and upper normal albumin levels demonstrated significantly longer survival times of 12.9 and 16.2 months, respectively (Fig. 2). It must be noted that the median survival times reported in the original article vary slightly from the results derived from synthetic data. This variance can be attributed to recent advances in evidence-based surgical and oncological practices implemented in multidisciplinary healthcare settings, which have significantly enhanced short-term survival outcomes for neurosurgical patients. Notably, the prevailing treatment strategy for GBM patients currently involves the combined use of adjuvant chemotherapy and radiotherapy. This approach is exemplified in the synthetic data, in which 41.3% of the patients have undergone a multimodal treatment regimen, whereas only 15.3% received multimodal treatment in the original study cohort ($\chi^2 P < 0.01$). Contemporary research indicates that combined chemoradiotherapy results in a median survival duration of 14.6 months (range: 13.2–16.8 months), [8] which aligns with the median survival of 15.4 months [95% CI 3.1–23.8] observed in the primary synthetic dataset for patients with the same multimodal treatment. Therefore, the median survival times observed from the synthetic data cohort are consistent with the current literature, when taking into consideration treatment discrepancies across cohorts.

Across all five datasets, overall survival times and demographic characteristics demonstrated noteworthy consistency, with very little inter-variability (Table 1). Each of the five synthetic datasets yielded a range of 452–468 viable patients, all exhibiting a median albumin level between 35–36g/L ($P = 0.93$). The characteristics variables across the five synthetic datasets showed no significant difference. Furthermore, median overall survival (OS) times were uniform for all five datasets (Table 1, $P = 0.90$).

Systemic inflammation scores correlate with survival prognosis in patients with newly diagnosed brain metastases. The Starzer *et al.* [7] study focused on using inflammatory markers as prognostic factors for survival in patients with brain metastasis. A group of 1250 patients was selected from the electronic medical record of the Medical University of Vienna between 1990 and 2019. Data collection included the source of brain metastases, brain metastases correlation with the primary tumor, and the presence of extra-cranial disease. Other factors included treatment methods such as radiotherapy, surgery, and chemotherapy. The authors collected data on systemic inflammatory markers, which included the neutrophil-to-lymphocyte ratio (NLR), platelet-to-lymphocyte ratio (PLR), monocyte-to-lymphocyte ratio (MLR), and the C-reactive protein-to-albumin ratio. These markers were measured at the time of diagnosis of the brain metastasis (± 14 days).

Results showed a notable correlation between the progression of extra-cranial disease and elevated systemic inflammatory markers, particularly the platelet-to-lymphocyte ratio (PLR) and C-reactive protein-to-albumin ratio (CRP/Alb), which were found to be significantly higher in patients with advancing extra-cranial disease. On the other hand, lower serum inflammatory markers were associated with longer overall survival.

The synthetic population contained 1320 patients compared to 1250 patients in the original study. Similar categorical features were present in the synthetic dataset such as cancer entity and first-line treatment to the brain metastasis (Table 2). Systematic inflammatory markers were also measured from the synthetic dataset, showing ranges comparable to the real dataset (Table 3).

Table 2
Patient characteristics in the original (Starzer *et al.*) and synthetic datasets.

Characteristics	Real data (N = 1250)	Synthetic data 1 (N = 1320)	Synthetic data 2 (N = 1320)	Synthetic data 3 (N = 1320)	Synthetic data 4 (N = 1320)	Synthetic data 5 (N = 1320)	P values*
Age at diagnosis of BM (Median[range])	62 [23–91]	67.2 [25.6–95.4]	67.01 [25.6–95.5]	66.8 [26.0–95.8]	66.6 [25.4–95.4]	67.3 [26.7–95.7]	1.0
Sex (Male/Female)	662 / 588	724 / 596	668 / 596	670/594	668/596	669/596	0.89
Cancer entity	994	499	499	499	498	498	0.99
Lung cancer	86	94	94	94	94	94	
Breast cancer	106	146	147	147	147	147	
Melanoma	7	28	28	28	28	28	
Renal cell carcinoma	11	60	60	60	60	60	
Colorectal cancer	36	493	494	494	495	495	
Others							
First-line treatment to BM	87	83	83	82	83	83	1.0
Surgery	973	676	676	676	675	675	
Radiotherapy							
Adjuvant chemotherapy	346	914	914	913	915	914	1.0
Yes	904	406	406	407	405	406	
No							

*P-values indicate the probability that the value of the specific variable is not significantly different between synthetic datasets.

Table 3
Systematic inflammatory markers in the original (Starzer *et al.*) and synthetic datasets.

	Real data	Synthetic data 1	Synthetic data 2	Synthetic data 3	Synthetic data 4	Synthetic data 5	P values*
	Median	Median	Median	Median	Median	Median	
Survival (months)	N/A	4.2 [1.00–72.46]	4.21 [1.01–76.46]	4.20 [1.02–73.43]	4.20 [0.80–72.78]	4.15 [0.87–73.33]	1.0
Leucocytes (g/L)	8.69 [0.89–48.8]	8.61 [1.16–53.63]	8.59 [1.17–53.41]	8.59 [1.21–53.62]	8.61 [1.18–53.47]	8.61 [1.17–53.35]	1.0
Neutrophils (g/L)	6.15 [0.24–66.0]	6.44 [0.64–44.26]	6.47 [0.63–45.22]	6.59 [0.63–45.07]	6.51 [0.6–44.36]	6.47 [0.62–43.39]	1.0
Lymphocytes (g/L)	1.30 [0.1–26.5]	1.05 [0.14–11.35]	1.04 [0.14–11.12]	1.06 [0.14–11.25]	1.06 [0.14–11.43]	1.06 [0.14–10.9]	1.0
Platelets (g/L)	272.0 [2.0–894.0]	242.0 [13–769]	239.0 [13–769]	239.0 [13–756]	238.0 [13–769]	239.0 [1.17–53.35]	1.0
Monocytes (g/L)	0.70 [0.01–2.42]	0.58 [0.02–2.09]	0.59 [0.02–2.11]	0.57 [0.02–2.09]	0.58 [0.02–2.1]	0.58 [0.02–2.11]	1.0
CRP (mg/L)	1.22 [0–43.7]	2.23 [0.2–39.06]	2.16 [0–39.49]	2.09 [0–39.52]	2.22 [0–39.55]	2.12 [0–39.53]	1.0
PLR	207.57 [2.0–894.0]	228.80 [7.43–2471.43]	230.09 [4.74–3657.14]	228.44 [4.59–2506.25]	227.83 [3.30–2162.50]	226.63 [5.44–3077.78]	0.99
LLR	6.39 [0.12–997.5]	7.91 [1.0–133.06]	7.86 [1.09–131.57]	7.79 [1.16–158.60]	7.87 [1.47–87.92]	7.79 [1.14–111.0]	1.0
MLR	0.52 [0.01–3.06]	0.53 [1.02–5.33]	0.53 [0.02–5.22]	0.53 [0.02–4.69]	0.53 [0.02–5.43]	0.52 [0.02–3.89]	1.0

	Real data	Synthetic data 1	Synthetic data 2	Synthetic data 3	Synthetic data 4	Synthetic data 5	P values*
	Median	Median	Median	Median	Median	Median	
NLR	4.76 [0.07-98.0]	6.01 [0.30-143.35]	6.09 [0.37-131.86]	6.03 [0.40-139.37]	6.01 [0.39-86.96]	6.05 [0.44-104.0]	1.0
CRP/Alb	2.41 [0.02-121.64]	6.26 [0.04-205.56]	6.03 [0.04-190.26]	5.94 [0.04-187.61]	6.08 [0.04-168.33]	5.99 [0.04-197.51]	1.0

*P-values indicate the probability that the value of the specific variable is not significantly different between synthetic datasets.

In the primary synthetic dataset, lower NLR (cut-off: <5.0 g/L) was associated with longer overall survival (OS), 5.1 months compared to 3.7 months in patients with a higher NLR ($P < 0.01$; mean of differences 2.2 [95% CI 1.0-3.4]). Furthermore, patients with lower LLR (cutoff: <5.7 g/L; 5.1 vs 3.8 months with $P < 0.01$; mean of differences 2.5 [95% CI 1.1–3.8]), lower PLR (cutoff: <3.5 g/L; 4.8 vs 3.5 months with $P < 0.04$; mean of differences 1.3 [95% CI 0.07–2.7]), lower MLR (cutoff: <0.5 g/L; 4.7 vs 3.7 months with $P < 0.01$; mean of differences 1.6 [95% CI 0.6–2.7]), and lower CLR/Alb (cutoff: <10.0 g/L; 5.2 months vs 3.4 months with $P < 0.01$; mean of differences 3.5 [95% CI 2.5–4.5]) presented with a more favorable survival prognosis (Fig. 3). The original article similarly depicted a comparable pattern, albeit with subtle variations noted among the monthly survival rates.

An advantage of synthetic data, in this case, is the ability to choose within different time frames and generate data in a timely fashion. All systemic inflammatory markers were taken within 14 days of brain metastasis diagnosis. However, if it was indicated otherwise, it would have been possible to rapidly generate another set of synthetic data within another time-frame.

Negligible variability can be observed between the five datasets generated for this study. All five synthetic datasets exhibit nearly identical median overall survival times, attesting to the reliability of the *MDCclone* software in generating various cohorts from a singular query (Table 3, $P = 0.99$). Furthermore, the different inflammatory markers were consistent across all five datasets showing no significant difference (Table 3).

DISCUSSION

In this study, we compared the findings of two published studies against results obtained from synthetic data. Our primary aim was to assess the reliability and validity of synthetic data in neuro-oncology research. More precisely, we examined prognostic factors for neuro-oncological patients, recognizing the potential of synthetic data to offer valuable insights. Our study also investigated the inter-variability of five synthetic datasets emanating from the same initial query. We used two clinical studies that differed in terms of parameters examined to comprehensively evaluate the reliability of synthetic data.

The overall findings from the synthetic data exhibited remarkable consistency and significant similarities to the results reported in both articles. Every reported result and cohort demographic showed remarkable similarity to those derived from our synthetic datasets. Consequently, this validation study provided compelling evidence that results obtained from synthetic data predict real data results and cohort demographics. This study demonstrated the ability of synthetic data generated from the *MDCclone* platform to reliably reproduce statistical characteristics of real patient data.

MDCclone platform enables in-depth analysis of large cohorts, with noted exceptions. Certain clinical elements such as specific tumor genotypes (e.g., isocitrate dehydrogenase 1/2 gene mutation status or O-6-methylguanine DNA methyltransferase gene promoter methylation) could not be extracted using *MDCclone*. This limitation stems from the restricted access to the information within a patient's "free-text". A patient's "free-text", commonly called non-structured data, encompasses notably all information present in medical reports or imaging results [3]. Non-structured data is not currently reflected in the synthetic data generated by the *MDCclone* platform. To overcome this limitation, exploring innovative data synthesis approaches, such as integrated image analysis or natural language processing, will become essential to extract statistical characteristics from non-structured data. Some studies^[3] have reported that higher accuracy is achieved in synthetic data when a larger number of patients relative to the number of variables are used.

Another limitation encountered corresponds to the liaison between each time-independent event. Interdependence of categories precludes determination of specific values. For example, to calculate NLR for an individual patient, neutrophil and lymphocyte values are needed, but data completeness is not assured. Unfortunately, this resulted in a decrease in the synthetic cohort size.

Synthetic data has several potential applications in the healthcare sector, ranging from hypothesis testing and epidemiological predictions to educational training. Several platforms presently offer diverse machine learning approaches to generate synthetic data.^[1] However, there is limited research on the validity of synthetic data across these diverse applications, and even fewer studies have reached a consensus or established a standard protocol for validating synthetic data [9]. This emphasizes the importance of further exploration in the realm of synthetic data validation, with the potential of developing a shared framework. Indeed, synthetic data generation platforms must prioritize establishing in-depth validation metrics. These metrics should assess individual data attributes and their correlations, ensuring synthetic data meet the standards of real datasets. Strengthening this process requires close collaboration between developers of synthetic databases and oncologists to ensure accurate encapsulation of nuances in the field and evolving standards. To validate results and guarantee wider applicability, these platforms should also collaborate with external institutions to expand their initial real-patient database.

Several key strategies must be explored to enhance the implementation of synthetic data in medicine. Incorporating molecular profiles of brain pathologies, such as genetic pathological variants, epigenetic markers, and protein expression, would allow for better evaluation of the natural history of specific molecular pathologies. For instance, O6-methylguanine DNA methyltransferase (MGMT) gene promoter methylation is a crucial prognostic marker for glioblastoma [8,10]. Additionally, refining the integration of synthetic datasets based on radiological imaging, such as MRI, carries substantial significance. There is

also a potential for innovation in generating longitudinal synthetic data by employing methods to mimic disease progression and treatment effects over time.

CONCLUSION

Synthetic data offers a significant potential to improve clinical data analysis and drive research, serving as a valid predictive tool for identifying clinical trends while preserving patient confidentiality. In particular, data synthesis platforms are key to advancing medical research by providing faster insights and facilitating improved data sharing among researchers. With the landscape of medical research continuing to evolve, embracing synthetic data as a powerful and ethical alternative can revolutionize the way cohort studies are conducted in oncology, ultimately leading to more robust and generalizable research findings.

Declarations

The authors have no personal, financial, or institutional interest in any drugs, materials, or devices described in this article.

Funding and Disclosures: Access to MDClone provided by RI-MUHC (Research Institute of the McGill University Health Centre) and the CLIP (Clinical Innovation Platform).

Acknowledgements: We would like to acknowledge Santiago Marquez Fosser facilitating access to the *MDClone* platform and Lena Zack for their assistance during its utilization.

Author Contribution

R.K and W.D wrote the main manuscript text and prepared figures 1-3. A.M reviewed the manuscript, provided guidance for statistical analysis. R.D is the supervisor who gave the main directives as to what analysis should be done. R.D reviewed the manuscript and provided guidance for statistical analysis.

References

1. Gonzales A, Guruswamy G, Smith SR (2023) Synthetic data in health care: A narrative review. *PLOS Digit Health* 2:e0000082
2. Mohan, Sa (2017) A Review Of Synthetic Data Generation Methods For Privacy Preserving Data Publishing. *J Sci Technol Res* ;95–101
3. Reiner, Benaim et al (2020) Analyzing Medical Research Results Based on Synthetic Data and Their Relation to Real Data Results: Systematic Comparison From Five Observational Studies. *JMIR Med Inform*
4. Moniz L, Buczak AL, Hung L et al (2009) Construction and validation of synthetic electronic medical records. *Online J Public Health Inform*

5. Foraker RE, Yu SC, Gupta A et al (2020) Spot the difference: comparing results of analyses from real patient data and synthetic derivatives. *JAMIA Open* 3:557–566
6. Borg N, Guilfoyle MR, Greenberg DC et al (2011) Serum albumin and survival in glioblastoma multiforme. *J Neurooncol* 105:77–81
7. Starzer AM, Steindl A, Mair MJ et al (2022) Correction: Systemic inflammation scores correlate with survival prognosis in patients with newly diagnosed brain metastases. *Br J Cancer* 126:968
8. Stupp R, Mason WP, van den Bent MJ et al (2005) Radiotherapy plus concomitant and adjuvant temozolomide for glioblastoma. *N Engl J Med* 352:987–996
9. Mc Lahlan et al (2018) The ATEN framework for creating the realistic synthetic electronic health record. *Proceedings of the 11th International Joint Conference on Biomedical Engineering Systems and Technologies* ;5:182–191
10. Hegi ME, Diserens AC, Godard S et al (2004) Clinical trial substantiates the predictive value of O-6-methylguanine-DNA methyltransferase promoter methylation in glioblastoma patients treated with temozolomide. *Clin Cancer Res* 10:1871–1874

Figures

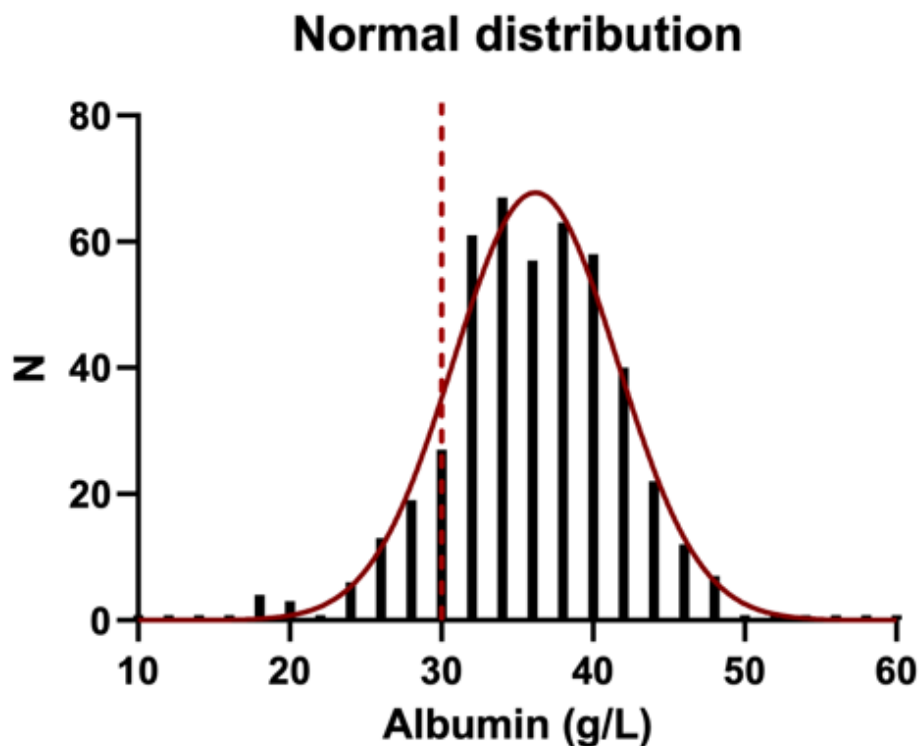


Figure 1

Distribution of serum albumin levels across the primary synthetic dataset cohort (n=452, Kolmogorov-Smirnov P=0.10, Kolmogorov-Smirnov statistic D=0.06). The red curve represents the Gaussian distribution for these values, while the vertical red-dotted line denotes the threshold for the hypoalbuminemic group (Albumin <30g/L)

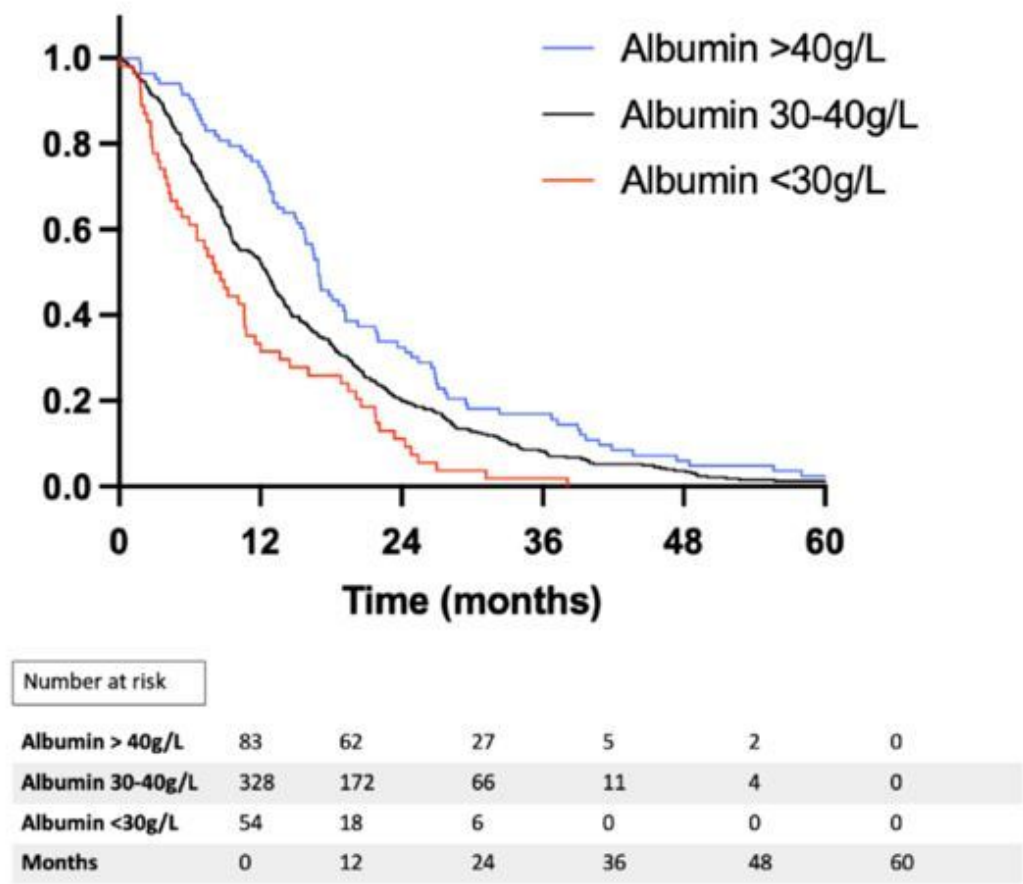


Figure 2

Kaplan-Meier survival curve of the primary synthetic dataset cohort subdivided into three categories: Hypoalbuminemic (red-lined, Albumin <30g/L, median= 7.0 months), lower normal serum albumin (black-lined, Albumin 30-40g/L, median = 12.9 months) and upper normal serum albumin (blue-lined, Albumin >40g/L, median= 16.2 months) (Log-rank P<0.0001)

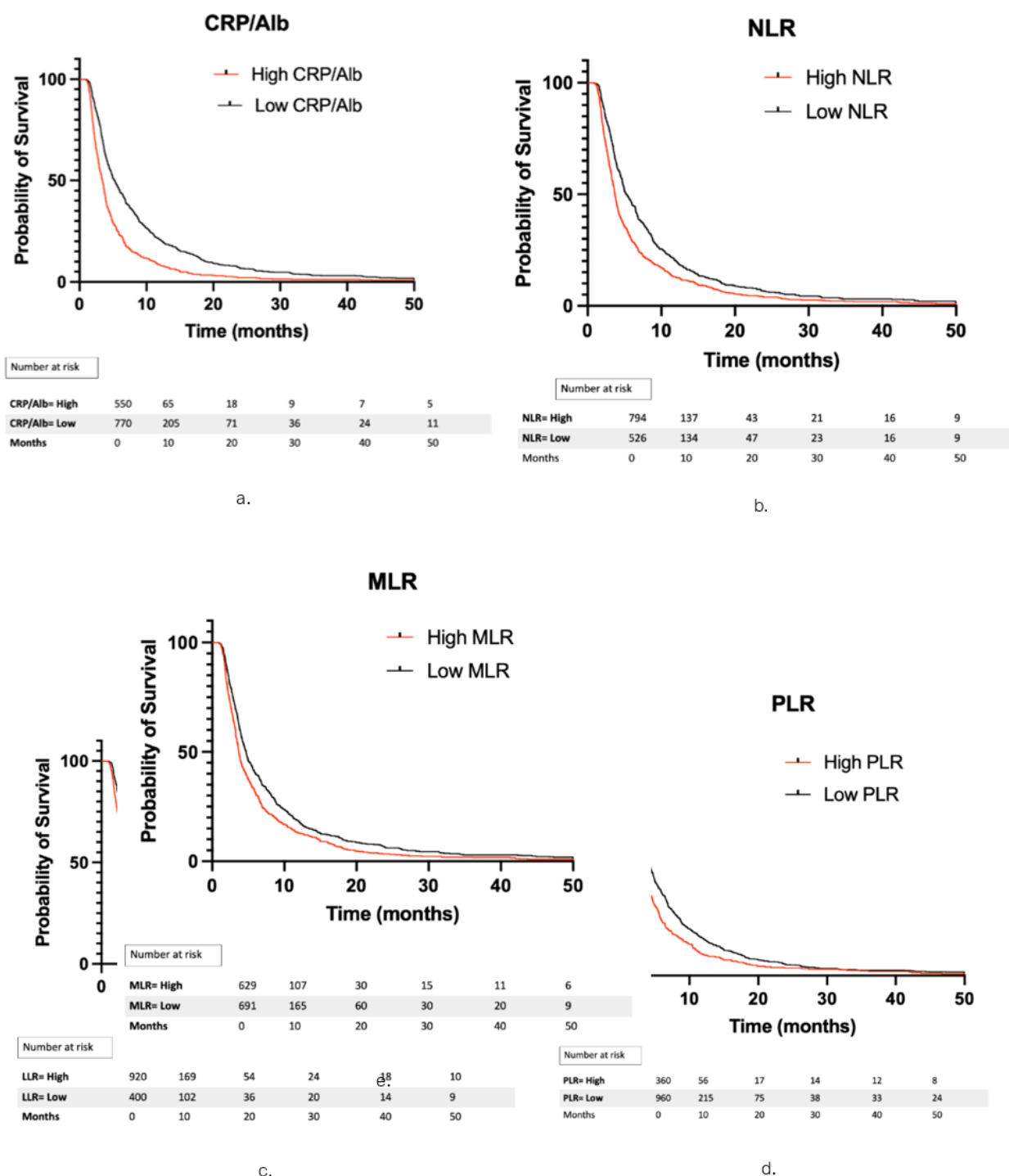


Figure 3

Kaplan-Meier survival curves of the primary synthetic dataset of patients with brain metastasis depending on level of serum inflammatory markers. A) Compares high CRP/Alb with low serum CRP/Alb (median survival 3.4 vs 5.2 months, mean of differences 3.5 [95% CI 2.5-4.5], $P < 0.0001$). B) Compares high NLR with low NLR (median survival 3.7 vs 5.1 months, mean of differences 2.2 [95% CI 1.0-3.4], $P < 0.0004$). C) Compares high LLR with low LLR (median survival 3.8 vs. 5.1 months, mean of differences 2.5 [95% CI 1.1-3.8], $P < 0.0003$). D) Compares high PLR with low PLR (median survival 3.6 vs 4.7 months, mean of

differences 1.3 [95% CI 0.07-2.7], $P<0.04$). E) Compares high MLR with low MLR (median survival 3.7 vs 4.7 months, mean of differences 1.6 [95% CI 0.6-2.7], $P<0.002$)