

# JMIR Bioinformatics and Biotechnology

Methods, devices, web-based platforms, open data and open software tools for big data analytics, understanding biological/medical data, and information retrieval in biology and medicine.  
Volume 6 (2025) ISSN 2563-3570 Editor in Chief: Ece D. Uzun, MS, PhD, FAMIA

## Contents

### Original Papers

Investigating Associations Between Prognostic Factors in Gliomas: Unsupervised Multiple Correspondence Analysis ( <a href="#">e65645</a> ) Maria Goes Job, Heidge Fukumasu, Tathiane Malta, Pedro Porfirio Xavier. . . . .	2
Decentralized Biobanking Apps for Patient Tracking of Biospecimen Research: Real-World Usability and Feasibility Study ( <a href="#">e70463</a> ) William Sanchez, Ananya Dewan, Eve Budd, M Eifler, Robert Miller, Jeffery Kahn, Mario Macis, Marielle Gross. . . . .	16
Designing a Finite Element Model to Determine the Different Fixation Positions of Tracheal Catheters in the Oral Cavity for Minimizing the Risk of Oral Mucosal Pressure Injury: Comparison Study ( <a href="#">e69298</a> ) Zhiwei Wang, Zhenghui Dong, Xiaoyan He, ZhenZhen Tao, Jinfang Qi, Yatian Zhang, Xian Ma. . . . .	39
Extracting Knowledge From Scientific Texts on Patient-Derived Cancer Models Using Large Language Models: Algorithm Development and Validation Study ( <a href="#">e70706</a> ) Jiarui Yao, Zinaida Perova, Tushar Mandloi, Elizabeth Lewis, Helen Parkinson, Guergana Savova. . . . .	48
A Hybrid Deep Learning–Based Feature Selection Approach for Supporting Early Detection of Long-Term Behavioral Outcomes in Survivors of Cancer: Cross-Sectional Study ( <a href="#">e65001</a> ) Tracy Huang, Chun-Kit Ngan, Yin Cheung, Madelyn Marcotte, Benjamin Cabrera. . . . .	63

# Investigating Associations Between Prognostic Factors in Gliomas: Unsupervised Multiple Correspondence Analysis

Maria Eduarda Goes Job<sup>1</sup>; Heidge Fukumasu<sup>1</sup>, PhD; Tathiane Maistro Malta<sup>2</sup>, PhD; Pedro Luiz Porfirio Xavier<sup>1</sup>, PhD

<sup>1</sup>Laboratory of Comparative and Translational Oncology, Department of Veterinary Medicine, School of Animal Science and Food Engineering, University of Sao Paulo, Avenida Duque de Caxias, 225, Jardim Elite, Pirassununga, Brazil

<sup>2</sup>Cancer Epigenomics Laboratory, Department of Clinical Analysis, Toxicology and Food Sciences, School of Pharmaceutical Sciences of Ribeirao Preto, University of Sao Paulo, Ribeirao Preto, Brazil

## Corresponding Author:

Pedro Luiz Porfirio Xavier, PhD

Laboratory of Comparative and Translational Oncology, Department of Veterinary Medicine, School of Animal Science and Food Engineering, University of Sao Paulo, Avenida Duque de Caxias, 225, Jardim Elite, Pirassununga, Brazil

## Abstract

**Background:** Multiple correspondence analysis (MCA) is an unsupervised data science methodology that aims to identify and represent associations between categorical variables. Gliomas are an aggressive type of cancer characterized by diverse molecular and clinical features that serve as key prognostic factors. Thus, advanced computational approaches are essential to enhance the analysis and interpretation of the associations between clinical and molecular features in gliomas.

**Objective:** This study aims to apply MCA to identify associations between glioma prognostic factors and also explore their associations with stemness phenotype.

**Methods:** Clinical and molecular data from 448 patients with brain tumors were obtained from the Cancer Genome Atlas. The DNA methylation stemness index, derived from DNA methylation patterns, was built using a one-class logistic regression. Associations between variables were evaluated using the  $\chi^2$  test with k degrees of freedom, followed by analysis of the adjusted standardized residuals (ASRs >1.96 indicate a significant association between variables). MCA was used to uncover associations between glioma prognostic factors and stemness.

**Results:** Our analysis revealed significant associations among molecular and clinical characteristics in gliomas. Additionally, we demonstrated the capability of MCA to identify associations between stemness and these prognostic factors. Our results exhibited a strong association between higher DNA methylation stemness index and features related to poorer prognosis such as glioblastoma cancer type (ASR: 8.507), grade 4 (ASR: 8.507), isocitrate dehydrogenase wild type (ASR:15.904), unmethylated MGMT (methylguanine methyltransferase) Promoter (ASR: 9.983), and telomerase reverse transcriptase expression (ASR: 3.351), demonstrating the utility of MCA as an analytical tool for elucidating potential prognostic factors.

**Conclusions:** MCA is a valuable tool for understanding the complex interdependence of prognostic markers in gliomas. MCA facilitates the exploration of large-scale datasets and enhances the identification of significant associations.

(JMIR Bioinform Biotech 2025;6:e65645) doi:[10.2196/65645](https://doi.org/10.2196/65645)

## KEYWORDS

brain tumors; bioinformatics; stemness; multiple correspondence analysis

## Introduction

Cancer is a dynamic and heterogeneous disease characterized by several hallmarks controlling and contributing to its development and progression [1]. Cancer research continually generates large scales of data encompassing clinical information, genomic and transcriptomic profiles, prognostic and diagnostic markers, and therapeutic targets [2]. Different approaches have been used to study and associate all these variables to manage this complexity, aiming to reduce the dimensionality and enhance data interpretation and decision-making process. Several features used to study and classify the different types of cancer are based on categorical variables. For instance, the

most widely used cancer staging system, TNM, is based on categorical variables, where “T” refers to the size of the primary tumor, “N” refers to the number of lymph nodes affected by cancer, and “M” refers to absence or presence of metastasis [3]. Thus, these biological and clinical variables interact, and their associations can be measured and diagnosed using statistical tests such as Fisher exact tests and  $\chi^2$  tests. However, these approaches could not provide a global and comprehensive picture of the associations between these variables, particularly in datasets with a large number of categorical variables. Therefore, using multivariate and visual analysis methods can significantly improve the analysis and interpretation of associations between clinical and molecular cancer phenotypes.

Brain tumors are a particularly aggressive type of cancer, mostly due to local tissue damage and highly invasive growth. Gliomas, which originate from neuroglial stem cells or progenitor cells, account for 30% of primary brain tumors and 80% of malignant brain tumors [4]. This heterogeneous disease is histologically classified based on anaplasia criteria and predominant cell types such as oligodendroglioma, astrocytoma, and glioblastoma (GBM) [5]. Nevertheless, as further investigation aimed to elucidate the neuropathological mechanisms of gliomas, new variables are considered for characterizing this cancer tumor, leading to reclassifications based on mutational profiles, clinical data, and epigenetic factors [6]. This scenario resulted in different prognosis predictions, diagnosis determination, and treatment responses, contributing to an increasingly complex and stratified understanding of gliomas.

Stemness is a key phenotype of cancer stem cells (CSCs), related to tumor initiation and progression, therapy resistance, and metastasis [7]. CSCs are referred to as a subpopulation of tumor cells able to self-renew and differentiate into distinct cell lineages, enabling those cells to adapt to different environmental situations [8]. Moreover, recent studies have demonstrated associations between stemness features and different histologic classifications or prognostic factors of gliomas [9–11]. Therefore, providing a comprehensive visualization of the associations between clinical features and stemness in brain tumors could be valuable for identifying and determining potential prognostic and therapeutic markers.

Multiple correspondence analysis (MCA) is an unsupervised data science methodology that aims to observe and represent associations between variables disposed in contingency tables, visualizing these associations in a 2D perceptual map. This approach allows for the simultaneous visualization of the relationship between 2 or more characteristics [12]. MCA shares general characteristics, and it is an extension of principal component analysis which is effective in reducing data dimensionality. Thus, MCA can significantly reduce the workload and simplify statistical analysis in healthy research [13]. The results of MCA are typically interpreted in a 2D map, where the relative positions of categories of each variable and their distribution along the dimensions are analyzed. Categories that cluster together and are closer are more likely to be associated, providing key insights into the relationship [14]. Despite its applicability, rigor, and success in other disciplines such as Geography, Epidemiology, and Human Physiology, MCA remains underused in Oncology research and few studies are applying [12,14–16].

By using MCA, we aimed to gain a deeper understanding of the interdependence between stemness and prognostic factors. Our findings revealed associations among molecular and clinical characteristics and prognostic factors, as previously described by the literature [17]. Additionally, we demonstrated the capability of MCA to identify associations between stemness and these prognostic factors. Our results exhibited a strong association between higher stemness index and features related to poorer prognosis, demonstrating the utility of MCA as an analytical tool for elucidating oncological heterogeneity and may also offer a valuable strategy for therapeutic decision-making. This study highlights MCA as a powerful tool

for overcoming the barrier of representing the heterogeneity and complexity of cancer variables, particularly in glioma.

## Methods

### Dataset of the Tumor Samples

Clinical and molecular information of a total of 448 patients with brain tumors was obtained from the Cancer Genome Atlas (TCGA). We tailored the dataset to contain only qualitative information, with 12 variables: cancer type, histology, grade, patient's vital status, IDH (isocitrate dehydrogenase) status, codeletion of chromosomes 1p and 19q arms, MGMT (methylguanine methyltransferase) gene methylation, telomerase reverse transcriptase (TERT) expression, gain of chromosome 19 and 20, chromosome 7 gain and chromosome 10 loss, ATRX (alpha thalassemia/mental retardation syndrome, X-linked) status, and GBM transcriptome subtypes. All categorical variables were selected based on their established role as prognostic factors for brain tumors.

### DNA Methylation Stemness Index

The DNA methylation stemness index (mDNAsi) based on DNA methylation was built using a one-class logistic regression [18] on the pluripotent stem cell samples (embryonic stem cell and induced pluripotent stem cell) from the Progenitor Cell Biology Consortium dataset [19,20]. The algorithm was built and validated as described in the original paper [21]. The mDNAsi was applied in 381 samples from the TCGA database. Malta's model presented a high correlation among other CSC signatures, providing significant insights into the biological and clinical features of pan-cancer. The workflow to generate the mDNAsi is available in the original paper [21].

### Multiple Correspondence Analysis

MCAs were conducted in the RStudio (version 4.3.1; Posit, PBC) environment using the packages FactoMineR (version 2.11; Institut Agro) [22] and cabooters (version 2.1.0; Cranfield University), for creating matrices for MCAs. Contingency tables for the categorical variables were generated, and associations between variables were assessed using a  $\chi^2$  test with k degrees of freedom. This was followed by the analysis of the adjusted standardized residuals (ASRs). The  $\chi^2$  test evaluates whether the observed associations between categorical variables are nonrandomly associated ( $P$  value  $< .05$ ). ASRs higher than 1.96 indicate a significant association between variables in the matrix. To perform MCA, the categorical variables should not be randomly associated. To create the perceptual map, inertia was determined as the total  $\chi^2$  divided by the number of samples, resulting in the number of associations in the dataset. MCA was performed based on the binary matrices and row and column profiles were determined to demonstrate the influence of each category of variables on the others. Matrices were defined based on the row and column profiles. Eigenvalues were then extracted to represent the number of dimensions that could be captured in the analysis. Finally, the x- and y-axis coordinates of the perceptual map were determined, allowing the category of the variables to be represented and established. In MCA, the spatial distance between categories of different variables reflects their associations. Categories with high coordinates that are close in

space are directly associated, while categories presenting high coordinates but opposing coordinates are inversely associated.

### Statistical Analysis

Fisher exact tests and  $\chi^2$  tests were performed using RStudio 4.3.1 environment and GraphPad Prism (version 10.3.0; Dotmatics, USA).

### Ethical Considerations

The results published in this paper are in whole based upon data generated by the TCGA Research Network [23]. TCGA Ethics and Policies was originally published by the National Cancer Institute [24].

## Results

### MCA Can Identify Associations Between Different Variables of Gliomas and Patient Vital Status

To determine the suitability of glioma variables for MCA, we first evaluated whether categorical glioma variables were randomly or nonrandomly associated. This involved creating individual contingency tables for each pair of glioma variables (Multimedia Appendices 1-13). Then, we applied  $\chi^2$  tests for each contingency table to confirm nonrandom associations ( $P$  value  $< .05$ ). We also confirmed the associations between categorical variables and patients' vital status using the Fisher exact test ( $P$  value  $< .05$ ) (Multimedia Appendix 14). Based on the  $\chi^2$  test, the results indicated that only 2 categorical variables, gender and DAXX expression, were randomly associated, suggesting no significant association patterns between these

variables and the others. Consequently, gender and DAXX expression were excluded from further analysis.

In the subsequent analysis, we observed and measured the strength of associations between the patient vital status (0-alive; 1-dead) and different factors including cancer type, histology, grade, IDH status, 1p19q codeletion, MGMT promoter methylation, gain of chromosome (Chr) 7 and loss of Chr10 (7+/10-), co-gain of Chr19 and Chr20 (19+/20+), TERT expression, ATRX status, and transcriptome subtype, aiming to determine whether MCA could identify associations between prognostic factors for this disease. We used ASRs to assess these associations, considering a category of each variable to be associated with either alive or dead vital status when the ASR values were higher than 1.96. Patients' vital status classified as dead were associated with poorer prognostics factors such as GBMs, grade 4, IDH wild type, non-codeleted 1p19q, unmethylated MGMT promoter, gain of Chr7 and loss of Chr10, expression of TERT, ATRX wild type, and classical (CL) and mesenchymal (ME) transcriptome subtypes (Table 1). In contrast, patients classified as alive were linked to favorable prognostic variables, including oligoastrocytomas and oligodendrogliomas, grade 2, IDH mutant, codeleted 1p19q, methylated MGMT promoter, absence of combined Chr7+/Chr10- (chromosome 7 gain and 10 loss), lack of TERT expression, ATRX mutant, and the proneural (PN) and neural (NE) transcriptome subtypes (Table 1). Histological classification, grade, IDH status, and Chr7+/Chr10- were the most strongly associated features with patient vital status. These associations were further illustrated in a heatmap (Figure 1A-D).

**Table.** Table exhibiting the values of the adjusted standardized residuals. Categories of variables with values higher than 1.96 are considered associated. We could observe a strong association between poorer prognostic factors and dead vital status. In contrast, better prognostic factors were associated with alive vital status.

Glioma variables	Patient vital status		Categories associated with
	Alive	Dead	
Glioblastoma	— <sup>a</sup>	8.127	Dead
Oligoastrocytoma	2.64	—	Alive
Oligodendroglioma	3.309	—	Alive
Astrocytoma	1.756	—	Not associated
Grade 2	6.809	—	Alive
Grade 3	0.155	—	Not associated
Grade 4	—	8.127	Dead
IDH <sup>b</sup> wild type	—	8.804	Dead
IDH mutant	8.804	—	Alive
1p/19q codeletion	5.265	—	Alive
1p/19q non-codeletion	—	5.265	Dead
Methylated MGMT <sup>c</sup> promoter	5.26	—	Alive
Unmethylated MGMT promoter	—	5.26	Dead
No combined Chr7+/Chr10 <sup>-d</sup>	5.756	—	Alive
Chr7+/Chr10 <sup>-</sup>	—	5.756	Dead
Not expressed TERT <sup>e</sup>	3.078	—	Alive
Expressed TERT	—	3.078	Dead
ATRX <sup>f</sup> mutant	2.311	—	Alive
ATRX wild type	—	2.311	Dead
Proneural subtype	4.122	—	Alive
Neural subtype	3.593	—	Alive
Mesenchymal subtype	—	4.635	Dead
Classical subtype	—	4.852	Dead

<sup>a</sup>Not applicable.

<sup>b</sup>IDH: isocitrate dehydrogenase.

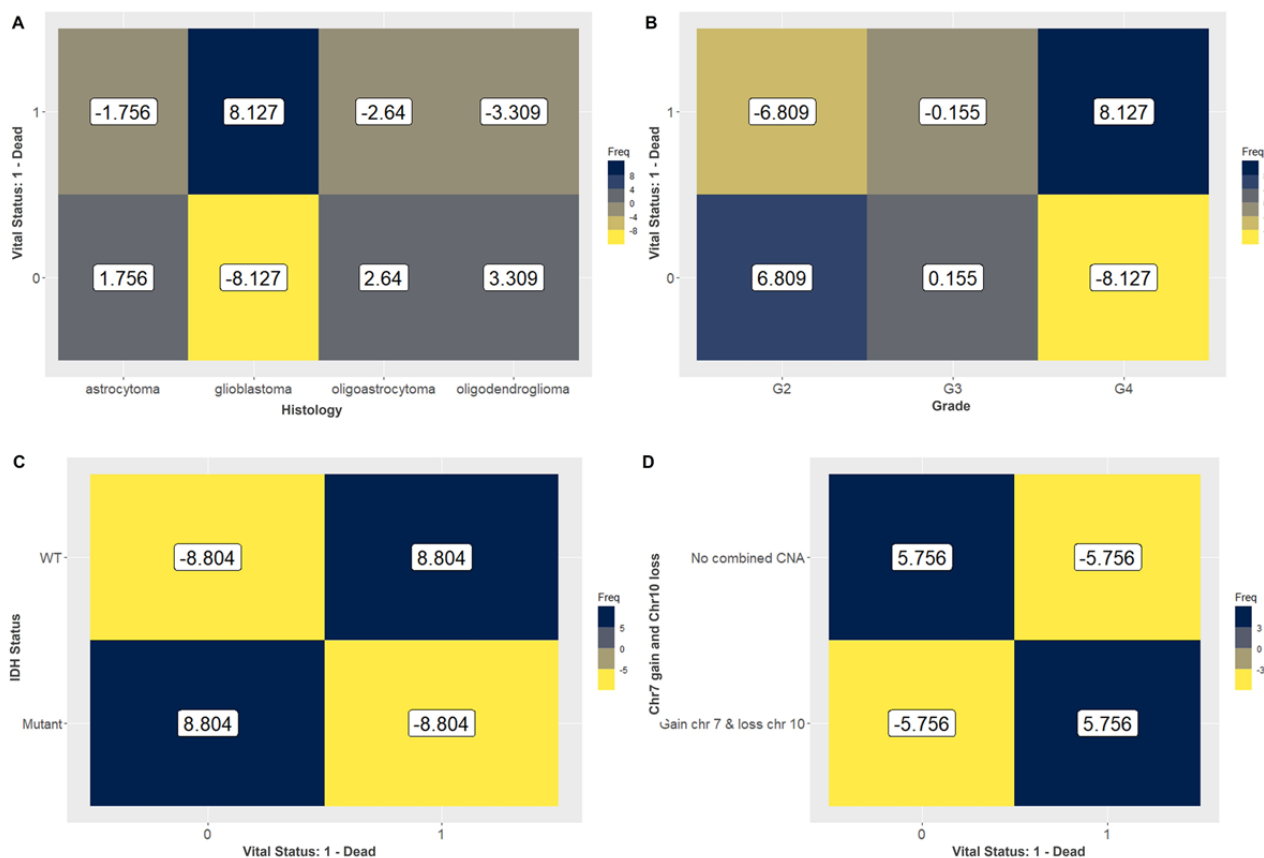
<sup>c</sup>MGMT: methylguanine methyltransferase.

<sup>d</sup>Chr7+/Chr10<sup>-</sup>: chromosome 7 gain and 10 loss.

<sup>e</sup>TERT: telomerase reverse transcriptase.

<sup>f</sup>ATRX: alpha thalassemia/mental retardation syndrome, X-linked.

**Figure 1.** Heatmap exhibiting the values of the adjusted standardized residuals. Categories of variables with values higher than 1.96 are associated. We could observe a strong association of (A) glioblastoma (8.127), (B) grade 4 (8.127), (C) IDH wild type (8.804), and (D) Chr7+/Chr10- (5.756) with dead vital status. Favorable prognostic factors including (A) oligoastrocytoma and oligodendroglioma, (B) grade 2, (C) IDH mutant, and (D) no combined copy number alterations were associated with alive vital status. Chr7+/Chr10-: chromosome 7 gain and 10 loss; IDH: isocitrate dehydrogenase.



Using MCA, we observed that dimension 1 (x-axis) accounted for 33.71% of the variance, while dimension 2 (y-axis) accounted for 14.08%. The inertia (sum of the variances) for these 2 dimensions was 47.79%. The variance of the overall dimensions (17 dimensions) for the combinations of the variables is illustrated in [Multimedia Appendix 15](#). The main idea was to present the percentage of explained variance for each dimension and not the influence of individual variables. The total inertia (sum of the variances) was 1.41.

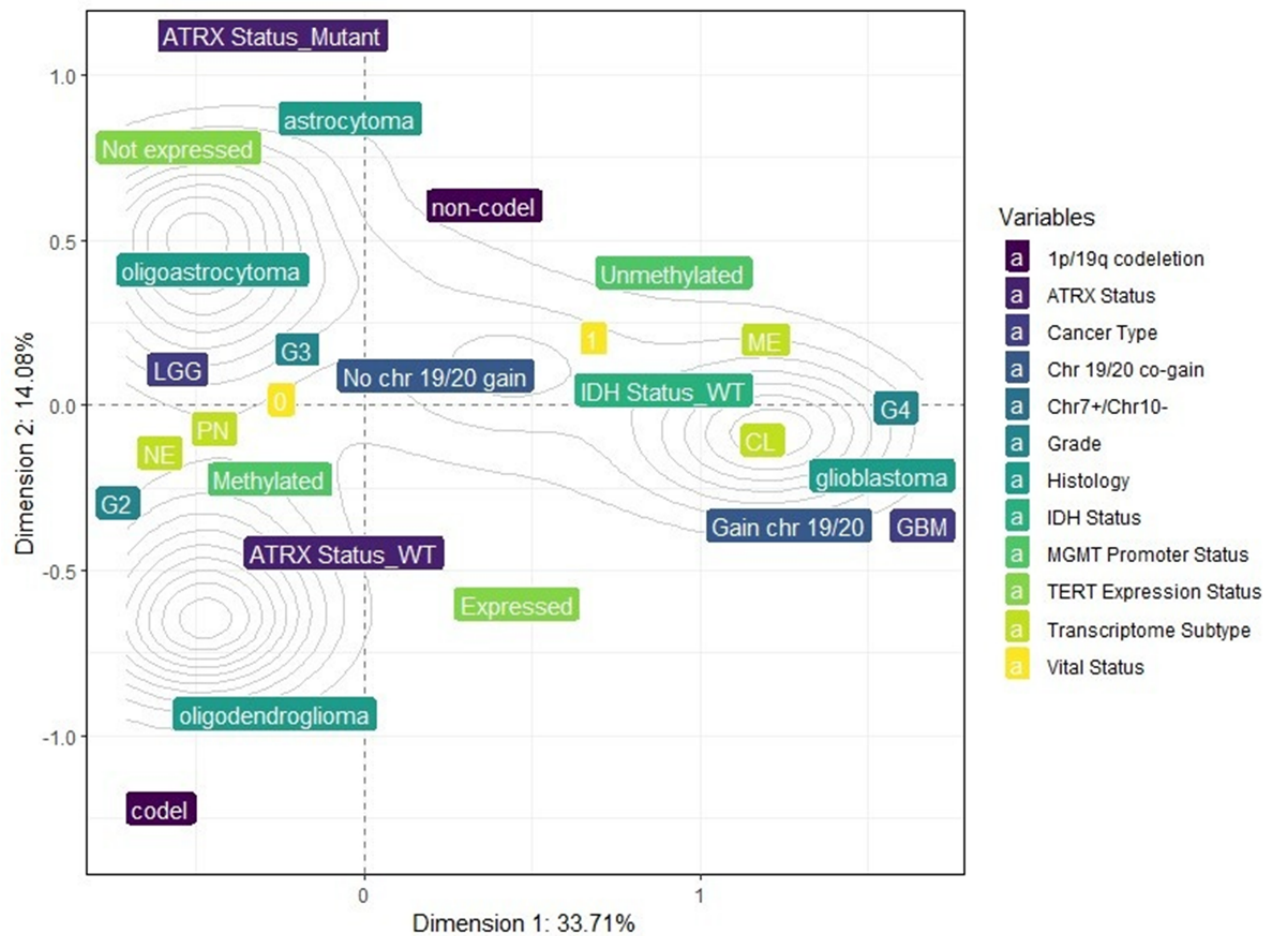
The results obtained from the MCA were visualized in a 2D perceptual map ([Figure 2](#)), highlighting the associations between the categories of each variable. The coordinates of each category are detailed in [Table 2](#). The perceptual map reveals that categories such as GBM, unmethylated MGMT promoter, IDH wild type, Chr7 gain and Chr10 loss, grade 4, GBM ATRX wild type, TERT expression, non-codel 1p.19q, and CL and ME

transcriptome subtypes are closely associated with dead vital status, appearing along the positive x-axis (dimension 1). Conversely, categories like oligoastrocytomas and oligodendrogliomas, grade 2, IDH mutant, codel 1p19q, methylated MGMT promoter, no combined copy number alterations, no expression of TERT, ATRX mutant, and PN and NE transcriptome subtypes are closely associated with alive vital status, appearing along the negative x-axis (dimension 1) ([Figure 2](#)).

These findings highlight the utility and capacity of MCA in reducing data dimensionality and demonstrate that, in gliomas, variables interact cohesively. MCA allows us to further visualize these interactions on a global perceptual map, organizing the characteristics into distinct clusters that correspond to different prognostic profiles.



**Figure 2.** Multiple correspondence analysis (MCA) 2D perceptual map demonstrating the association between the categories of each categorical variable. Categories that are closely clustered are strongly associated with each other. Categories such as glioblastoma, unmethylated MGMT promoter, IDH wild type, chromosome 7 gain and 10 loss (Chr7+/Chr10-), grade 4, glioblastoma ATRX wild type, TERT expression, non-codel 1p.19q, CL and ME transcriptome subtypes are closely associated with dead vital status (1), appearing along the positive x-axis (dimension 1). ATRX: alpha thalassemia/mental retardation syndrome, X-linked; CL: classical; GBM: glioblastoma; IDH: isocitrate dehydrogenase; ME: mesenchymal; MGMT: methylguanine methyltransferase; NE: neural; PN: proneural; TERT: telomerase reverse transcriptase.



**Table .** Coordinates of each category compounding the perceptual map.

Category	Dimension 1 (x-axis)	Dimension 2 (y-axis)
GBM <sup>a</sup>	1.6650830	−0.0896760
Low-grade glioma	−0.4723301	0.0254382
Astrocytoma	−0.2672355	0.9527631
Glioblastoma	1.6650830	−0.0896760
Oligoastrocytoma	−0.5334711	0.3276318
Oligodendroglioma	−0.6011671	−0.9346433
Grade 2	−0.6611308	−0.1971919
Grade 3	−0.2970898	0.2320783
Grade 4	1.6650830	−0.0896760
0-Alive	−0.3185609	−0.0551369
1-Dead	0.7544862	0.1305874
IDH <sup>b</sup> mutant	−0.6734117	−0.0548104
IDH wild type	1.1888626	0.0967641
1p/19q code1	−0.6877365	−13.034.766
1p/19q non-code1	0.2750946	0.5213906
Methylated	−0.3429710	−0.1087842
Unmethylated	1.0048449	0.3187185
Chr7+/Chr10 <sup>−c</sup>	1.4087248	−0.0210234
No combined Chr7+/Chr10 <sup>−</sup>	−0.4205758	0.0062766
Chr 19/20 co-gain	1.4900007	−0.1295089
No Chr 19/20 co-gain	−0.0843397	0.0073307
Expressed TERT <sup>d</sup>	0.3715020	−0.6845760
Not expressed TERT	−0.4690682	0.8643636
ATRX <sup>e</sup> mutant	−0.6448249	1.0773395
ATRX wild type	0.2693572	−0.4500279
Classical	1.2675815	−0.0217510
Mesenchymal	1.0920361	0.2687642
Neural	−0.5475482	−0.0650952
Proneural	−0.5971662	−0.0604168

<sup>a</sup>GBM: glioblastoma.  
<sup>b</sup>IDH: isocitrate dehydrogenase.  
<sup>c</sup>Chr7+/Chr10<sup>−</sup>: chromosome 7 gain and 10 loss.  
<sup>d</sup>TERT: telomerase reverse transcriptase.  
<sup>e</sup>ATRX: Alpha Thalassemia/Mental Retardation Syndrome X-linked.

**MCA Can Associate an Epigenetic Stemness Index (mDNAsi) as a Prognostic Factor in Gliomas**

After demonstrating that MCA effectively reduces dimensionality and identifies associations between prognostic factors and clinical data in the glioma database, we proceeded to explore whether MCA could also associate these variables with stemness phenotype. For this analysis, we updated our database by including mDNAsi as a new variable, categorized into low, intermediate, and high levels of stemness. These

categories were based on the DNA methylation index related to tumor pathology and clinical outcomes, as previously studied by [21].

First, we evaluated whether the categorical glioma variables were randomly or nonrandomly associated with mDNAsi by creating individual contingency tables for each pair of glioma variables and applying  $\chi^2$  tests (Multimedia Appendix 16). We also confirmed the associations between categorical variables using the Fisher exact test ( $P$  value <.05) ( Multimedia



Appendix 17). All the variables were found to be suitable for MCA. Then, using ASR values to evaluate the strength of these associations, our results indicated strong associations between high mDNAsi levels and poor prognostic and clinical factors. Higher mDNAsi levels were associated with GBM, IDH wild-type, absence of 1p19q co-deletion, unmethylated MGMT promoter, TERT expression, grade 3 and 4, patient's vital status as dead, Chr7+/Chr10-, chromosomes 19/20 co-gain, ATRX

wildtype and ME and CL transcriptome subtypes (Table 3). Conversely, intermediate and lower levels of mDNAsi were associated with characteristics related to favorable prognosis, including oligodendroglioma, IDH mutant, 1p19q co-deletion, methylation of MGMT promoter, absence of TERT expression, grade 2, patient's vital status as alive, no combined copy number alteration, absence of chromosomes 19/20 co-gain, ATRX mutant, and PN and NE transcriptome subtypes (Table 3).

**Table .** Table exhibiting the values of the adjusted standardized residuals. Categories of variables with values higher than 1.96 are considered associated. We could observe a strong association between poorer prognostic factors and a higher stemness index (DNA methylation stemness index [mDNAsi]). In contrast, better prognostic factors were associated with lower stemness index.

Glioma Variables	mDNAsi			Categories associated with
	Low	Intermediate	High	
Glioblastoma	— <sup>a</sup>	—	8.507	High
Oligoastrocytoma	—	—	—	Not associated
Oligodendroglioma	3.949	—	—	Low
Astrocytoma	—	—	2.832	High
G2	3.279	4.057	—	Low and intermediate
G3	—	—	2.392	High
G4	—	—	8.507	High
IDH <sup>b</sup> wild type	—	—	15.904	High
IDH mutant	8.743	7.057	—	Low and intermediate
1p/19q codeletion	5.772	2.102	—	Low and intermediate
1p/19q non-codeletion	—	—	7.964	High
Methylated MGMT <sup>c</sup> promoter	5.944	3.961	—	Low and intermediate
Unmethylated MGMT promoter	—	—	9.983	High
No combined Chr7+/Chr10- <sup>d</sup>	6.436	5.927	—	Low and intermediate
Chr7+/Chr10-	—	—	12.433	High
Not expressed TERT <sup>e</sup>	—	3.216	—	Intermediate
Expressed TERT	—	—	3.351	High
ATRX <sup>f</sup> mutant	—	3.505	—	Intermediate
ATRX wild type	—	—	4.949	High
Proneural subtype	8.476	—	—	Low
Neural subtype	—	4.218	—	Intermediate
Mesenchymal subtype	—	—	4.771	High
Classical subtype	—	—	10.981	High

<sup>a</sup>Not applicable.

<sup>b</sup>IDH: isocitrate dehydrogenase.

<sup>c</sup>MGMT: methylguanine methyltransferase.

<sup>d</sup>Chr7+/Chr10-: chromosome 7 gain and 10 loss.

<sup>e</sup>TERT: telomerase reverse transcriptase.

<sup>f</sup>ATRX: Alpha Thalassemia/Mental Retardation Syndrome X-linked.

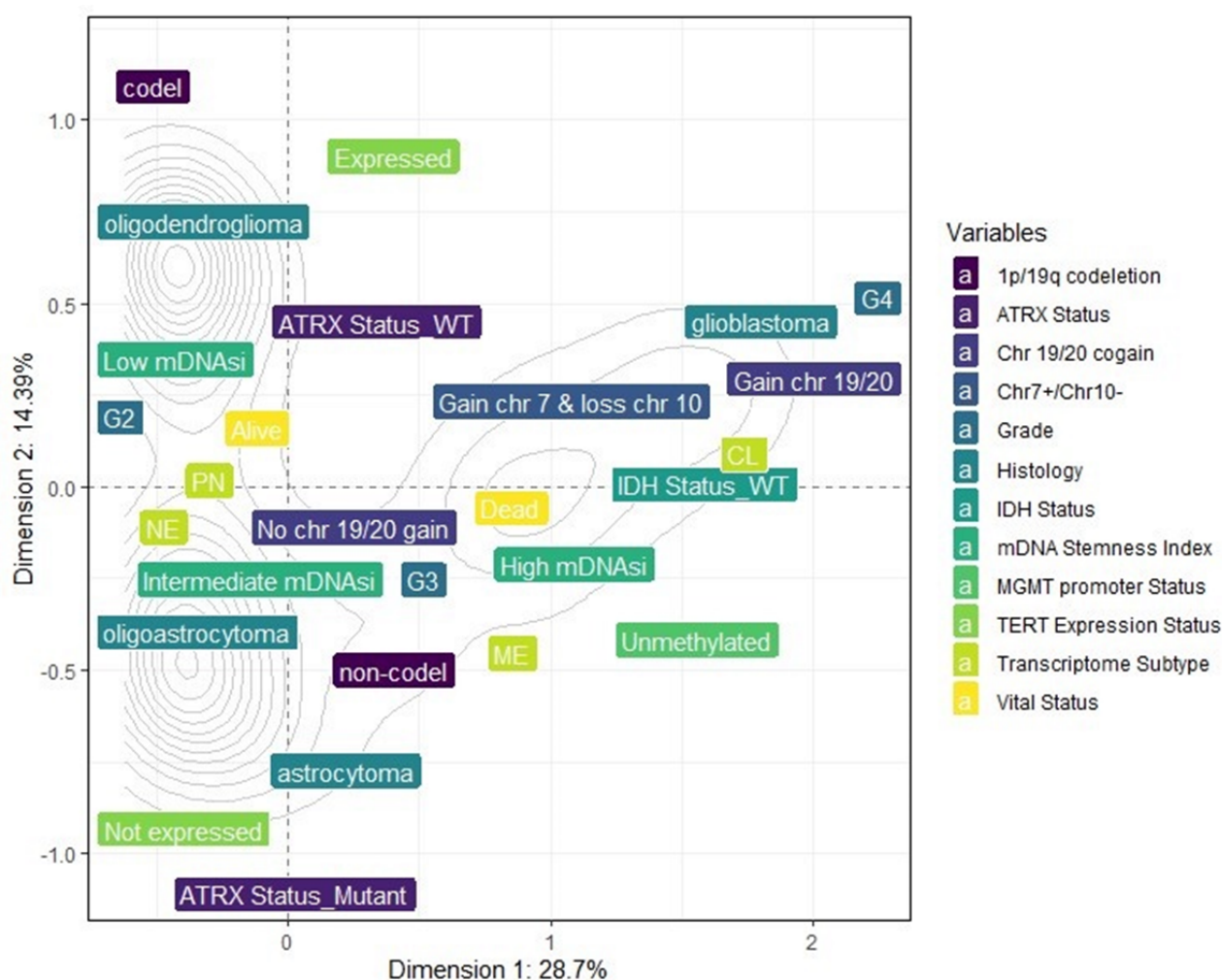
Using MCA, dimension 1 (x-axis) accounted for 28.7% of the variance, while dimension 2 (y-axis) accounted for 14.39%.

The inertia (sum of the variances) for these 2 dimensions was 43.09%. The variance of the overall dimensions (18 dimensions)

for the combinations of the variables is illustrated in [Multimedia Appendix 18](#). The total inertia (sum of the variances) was 1.5. The 2D perceptual map exhibited the associations between the categories of each variable ([Figure 3](#)). The perceptual map reveals categories such as GBM, unmethylated MGMT promoter, IDH wild type, Chr7 gain and Chr10 loss, grade 4, GBM ATRX wild type, TERT expression, non-codel 1p.19q, and CL and ME transcriptome subtypes are closely associated

with high mDNAsi, appearing along the positive x-axis (dimension 1). Conversely, categories like oligoastrocytomas and oligodendrogliomas, grade 2, IDH mutant, codel 1p19q, methylated MGMT promoter, no combined copy number alterations, no expression of TERT, ATRX mutant, and PN and NE transcriptome subtypes are closely associated with alive vital status, appearing along the negative x-axis (dimension 1) ([Figure 3](#)).

**Figure 3.** Multiple correspondence analysis (MCA) 2D perceptual map demonstrating the association between the categories of each categorical variable. Categories that are closely clustered are strongly associated with each other. Categories such as glioblastoma, unmethylated MGMT promoter, IDH wild type, chromosome 7 gain and 10 loss (Chr7+/Chr10-), grade 4, glioblastoma ATRX wild type, TERT expression, non-codel 1p.19q, and CL and ME transcriptome subtypes are closely associated with high mDNAsi, appearing along the positive x-axis (dimension 1). ATRX: alpha thalassemia/mental retardation syndrome, X-linked; CL: classical; IDH: isocitrate dehydrogenase; mDNAsi: DNA methylation stemness index; ME: mesenchymal; MGMT: methylguanine methyltransferase; NE: neural; PN: proneural; TERT: telomerase reverse transcriptase.



## Discussion

### Principal Findings

Multiple efforts have been made to explore the diversity of oncologic diseases, with significant contributions from genetics, cell and tissue biology, as well as computational and experimental technologies, providing a wealth of information on cancer manifestations. In the field of glioma research, emerging approaches have sought to clarify tumor pathology and grading through the introduction of novel types and subtypes, as well as by identifying molecular markers and genetic mutations that contribute to predicting diagnosis and

prognosis. However, it also results in an accumulation of extensive datasets, presenting challenges in interpretation and visualization regarding the associations between prognostic factors. In this study, we used MCA, an unsupervised data science approach, to establish statistical associations between different qualitative variables of gliomas. This method was able to reduce data dimensionality and represent it on a 2D perceptual map, revealing associations between various established glioma prognostic factors, including histological classification, IDH status, MGMT promoter methylation, and transcriptome subtypes. Furthermore, we associated these clinical and prognostic variables with an epigenetic-based stemness index

(mDNAsi), demonstrating that higher stemness levels were associated with poorer prognostic factors, providing a useful tool to associate prognostic markers in brain tumors.

### Comparison to Prior Studies

Several clinical and molecular factors are considered in predicting the prognosis and survival of brain tumors, more specifically for gliomas. Beyond histological classification and tumor grade, genetic and molecular biomarkers have been incorporated as potential prognostic indicators. Thus, we first evaluated the ability of MCA to associate these consolidated prognostic variables with the patient's vital status. Our findings demonstrate that MCA effectively clusters poor prognostic factors with dead vital status. All these prognostic factors are well consolidated and associated with malignancy of gliomas. IDH mutation represents one of the main prognostic markers for gliomas [25]. It has been identified that one of the mechanisms given by this favorable outcome is the impaired production of nicotinamide adenine dinucleotide phosphate in Krebs cycle caused by IDH1 enzyme mutation that can sensitize tumor cells to chemotherapy and explain the favorable prognosis of patients with IDH mutation [25]. Likewise, co-deletion of 1p19q chromosome arms, especially when combined with other biomarkers such as IDH mutation and TERT expression, has been used as a predictive biomarker and recent studies investigated biological mechanisms to be significantly linked to genes involved in cell division, angiogenesis, and DNA repair responses [26]. Thus, we demonstrated that MCA was able to capture and associate key glioma hallmarks with patients' vital status, which was applied to different clinical variables.

Subsequently, we applied MCA to explore the association between high stemness levels (mDNAsi) and characteristics related to poor prognosis. Stemness has been considered an important phenotype in glioma malignancy and is potentially associated with CL genetic alterations, such as the gain of chromosome 7. Chromosome 7 harbors some key genes related to stemness, including Epidermal Growth Factor Receptor (EGFR), Mesenchymal-Epithelial Transition Factor (MET), and Homeobox A gene (HOXA). A study of 86 GBMs reported that EGFR amplification occurs with higher probability in samples that have a gain of chromosome 7 (82.1%) compared with those without it (66.7%) [27]. In addition, EGFR amplification is more prevalent in IDH-wildtype diffuse gliomas (66.0%) and GBM (85.5%) [28], which are also associated with poorer prognostic factors, consistent with our findings. High mDNAsi has been previously linked to EGFR mutations [21]. The HOXA and MET loci, also located on chromosome 7, have been implicated in stemness-related pathways. Notably, studies have demonstrated interactions between chromosome 7 gain and the expression of a stem cell-related HOX signature in GBMs [29]. Analysis of the MET gene at 7q31.2 revealed that gain occurs in 47% of primary and 44% of secondary GBMs, suggesting that this genetic alteration contributes to the pathogenesis of both GBM subtypes [30].

Overall, relatively few studies have used MCA to explore associations with cancer phenotypes. Previous studies have applied MCA to different approaches, such as analyzing prognosis low rectal cancer surgery [31], investigating the association between some types of cancer in rural or urban areas [15], examining the association between Traditional Chinese Medicine Syndrome and histopathology of colorectal cancer [32], assessing clinically relevant demographic variables across multiple gastrointestinal cancers [33], and the relationship between types of diagnostic classification in breast cancer [34]. Our study also highlights the utility of MCA in investigating associations within the context of brain tumors. MCA enables the investigation of the pattern among many categorical factors in gliomas, providing a powerful computational approach to identify and test prognostic variables. It was possible to visually and quantitatively represent the associations, which facilitates the identification of distinct patient clusters based on shared prognostic characteristics. Our findings were consistent with previous literature and emphasized stemness as an important phenotype for gliomas.

### Limitations

Our study has inherent limitations. First, as a retrospective analysis of TCGA data, it is subject to selection bias. Second, we associated all the prognostic variables with patients' vital status, which may not be the most optimal variable for determining prognosis. For the future, we intend to improve our model validating its applicability in other prospective datasets. Third, the absence of therapy data is another limitation of this study. Finally, an intrinsic limitation of MCA is that retaining only 2 or 3 dimensions may not sufficiently capture all the significant features in the data. In our analysis, the percentage of explained inertia was approximately 40%. While there is not an accepted threshold for adequately explained inertia, common guidelines recommend retaining dimensions that represent over 70% of the inertia [35]. However, explained inertia in the range of 40% - 60% is often considered informative, and the interpretability and relevance of the patterns revealed by the dimensions are frequently more important than the exact percentage of inertia explained, especially in a complex heterogeneous disease such as brain tumors [36].

### Conclusion and Future Perspectives

In conclusion, our findings suggest that MCA is a valuable tool for understanding the interdependence between prognostic markers in gliomas. MCA facilitates the exploration of a large-scale dataset and enhances the identification of associations. Considering the advances in computational oncology and the emergence of new oncological features, such as stemness phenotype, incorporating MCA into cancer research as an approach to exploring the complex heterogeneity of the oncologic field becomes a powerful tool for simplifying data management. It contributes to researchers statistically identifying associations between variables within extensive databases and improves the visual representation, leading to a deeper understanding of cancer findings.

## Acknowledgments

This study has been supported by grants from the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) and from the Sao Paulo Research Foundation (FAPESP), Brazil (2018/00583-0, 2022/06305-7, 2022/09378-5, 2023/05099-7, 2023/07358-0).

## Data Availability

The datasets generated or analyzed during this study are available at National Institutes of Health Genomic Data Commons (GDC) [37]. The workflow to generate the DNA methylation stemness index (mDNAsi) can be accessed at GitHub [38].

## Authors' Contributions

MEGJ conducted the study, contributing to the acquisition of data, data analysis and interpretation, production of tables and figures, and wrote the first version of the manuscript. HF contributed to the interpretation and discussion of data and corrected the final version of the manuscript. TMM contributed to the acquisition, interpretation, and discussion of data, and corrected the final version of the manuscript. PLPX contributed to the concept and design of the study, data analysis and interpretation, funding, and corrected the final version of the manuscript.

## Conflicts of Interest

None declared.

### Multimedia Appendix 1

Individual contingency tables for cancer type.

[[XLSX File, 29 KB](#) - [bioinform\\_v6i1e65645\\_app1.xlsx](#)]

### Multimedia Appendix 2

Individual contingency tables for histology.

[[XLSX File, 32 KB](#) - [bioinform\\_v6i1e65645\\_app2.xlsx](#)]

### Multimedia Appendix 3

Individual contingency tables for grade.

[[XLSX File, 27 KB](#) - [bioinform\\_v6i1e65645\\_app3.xlsx](#)]

### Multimedia Appendix 4

Individual contingency tables for gender.

[[XLSX File, 24 KB](#) - [bioinform\\_v6i1e65645\\_app4.xlsx](#)]

### Multimedia Appendix 5

Individual contingency tables for vital status.

[[XLSX File, 23 KB](#) - [bioinform\\_v6i1e65645\\_app5.xlsx](#)]

### Multimedia Appendix 6

Individual contingency tables for IDH (isocitrate dehydrogenase) status.

[[XLSX File, 21 KB](#) - [bioinform\\_v6i1e65645\\_app6.xlsx](#)]

### Multimedia Appendix 7

Individual contingency tables for X1p.19q.codeletion.

[[XLSX File, 20 KB](#) - [bioinform\\_v6i1e65645\\_app7.xlsx](#)]

### Multimedia Appendix 8

Individual contingency tables for MGMT (methylguanine methyltransferase) promoter.

[[XLSX File, 18 KB](#) - [bioinform\\_v6i1e65645\\_app8.xlsx](#)]

### Multimedia Appendix 9

Individual contingency tables for Chr 7 gain and Chr 10 loss.

[[XLSX File, 17 KB](#) - [bioinform\\_v6i1e65645\\_app9.xlsx](#)]

## Multimedia Appendix 10

Individual contingency tables for Chr 19/20 co-gain.

[[XLSX File, 16 KB](#) - [bioinform\\_v6ile65645\\_app10.xlsx](#) ]

## Multimedia Appendix 11

Individual contingency tables for TERT (telomerase reverse transcriptase) expression status.

[[XLSX File, 13 KB](#) - [bioinform\\_v6ile65645\\_app11.xlsx](#) ]

## Multimedia Appendix 12

Individual contingency tables for ATRX (Alpha Thalassemia/Mental Retardation Syndrome X-linked alpha thalassemia/mental retardation syndrome, X-linked) status.

[[XLSX File, 11 KB](#) - [bioinform\\_v6ile65645\\_app12.xlsx](#) ]

## Multimedia Appendix 13

Individual contingency tables for DAXX status.

[[XLSX File, 10 KB](#) - [bioinform\\_v6ile65645\\_app13.xlsx](#) ]

## Multimedia Appendix 14

Fisher exact test and  $\chi^2$  test for vital status  $\times$  glioma prognostic factors.

[[XLSX File, 24 KB](#) - [bioinform\\_v6ile65645\\_app14.xlsx](#) ]

## Multimedia Appendix 15

Percentage of explained variances of the overall (17) dimensions.

[[PNG File, 161 KB](#) - [bioinform\\_v6ile65645\\_app15.png](#) ]

## Multimedia Appendix 16

Individual contingency table for mDNAsI.

[[XLSX File, 30 KB](#) - [bioinform\\_v6ile65645\\_app16.xlsx](#) ]

## Multimedia Appendix 17

Fisher exact test and  $\chi^2$  test for mDNAsI (DNA methylation stemness index)  $\times$  glioma prognostic factors.

[[XLSX File, 22 KB](#) - [bioinform\\_v6ile65645\\_app17.xlsx](#) ]

## Multimedia Appendix 18

Percentage of explained variances of the overall (18) dimensions.

[[PNG File, 9 KB](#) - [bioinform\\_v6ile65645\\_app18.png](#) ]

## References

1. Hanahan D. Hallmarks of cancer: new dimensions. *Cancer Discov* 2022 Jan;12(1):31-46. [doi: [10.1158/2159-8290.CD-21-1059](#)] [Medline: [35022204](#)]
2. Dagogo-Jack I, Shaw AT. Tumour heterogeneity and resistance to cancer therapies. *Nat Rev Clin Oncol* 2018 Feb;15(2):81-94. [doi: [10.1038/nrclinonc.2017.166](#)] [Medline: [29115304](#)]
3. Brierley J, O'Sullivan B, Asamura H, et al. Global consultation on cancer staging: promoting consistent understanding and use. *Nat Rev Clin Oncol* 2019 Dec;16(12):763-771. [doi: [10.1038/s41571-019-0253-x](#)] [Medline: [31388125](#)]
4. Weller M, Wick W, Aldape K, et al. Glioma. *Nat Rev Dis Primers* 2015 Jul 16;1:15017. [doi: [10.1038/nrdp.2015.17](#)] [Medline: [27188790](#)]
5. Louis DN, Ohgaki H, Wiestler OD, et al. The 2007 WHO classification of tumours of the central nervous system. *Acta Neuropathol* 2007 Aug;114(2):97-109. [doi: [10.1007/s00401-007-0243-4](#)] [Medline: [17618441](#)]
6. Louis DN, Perry A, Wesseling P, et al. The 2021 WHO classification of tumors of the central nervous system: a summary. *Neuro Oncol* 2021 Aug 2;23(8):1231-1251. [doi: [10.1093/neuonc/noab106](#)] [Medline: [34185076](#)]
7. Ayob AZ, Ramasamy TS. Cancer stem cells as key drivers of tumour progression. *J Biomed Sci* 2018 Mar 6;25(1):20. [doi: [10.1186/s12929-018-0426-4](#)] [Medline: [29506506](#)]
8. Battle E, Clevers H. Cancer stem cells revisited. *Nat Med* 2017 Oct 6;23(10):1124-1134. [doi: [10.1038/nm.4409](#)] [Medline: [28985214](#)]
9. Wang Q, Hu B, Hu X, et al. Tumor evolution of glioma-intrinsic gene expression subtypes associates with immunological changes in the microenvironment. *Cancer Cell* 2017 Jul 10;32(1):42-56. [doi: [10.1016/j.ccell.2017.06.003](#)] [Medline: [28697342](#)]



10. Ortensi B, Setti M, Osti D, Pelicci G. Cancer stem cell contribution to glioblastoma invasiveness. *Stem Cell Res Ther* 2013 Feb 28;4(1):18. [doi: [10.1186/scrt166](https://doi.org/10.1186/scrt166)] [Medline: [23510696](https://pubmed.ncbi.nlm.nih.gov/23510696/)]
11. Tan J, Zhu H, Tang G, et al. Molecular subtypes based on the stemness index predict prognosis in glioma patients. *Front Genet* 2021;12:616507. [doi: [10.3389/fgene.2021.616507](https://doi.org/10.3389/fgene.2021.616507)] [Medline: [33732284](https://pubmed.ncbi.nlm.nih.gov/33732284/)]
12. Sourial N, Wolfson C, Zhu B, et al. Correspondence analysis is a useful tool to uncover the relationships among categorical variables. *J Clin Epidemiol* 2010 Jun;63(6):638-646. [doi: [10.1016/j.jclinepi.2009.08.008](https://doi.org/10.1016/j.jclinepi.2009.08.008)] [Medline: [19896800](https://pubmed.ncbi.nlm.nih.gov/19896800/)]
13. Li BH, Sun ZQ, Dong SF. Correspondence analysis and its application in oncology. *Commun Stat Theory Methods* 2010 Mar 19;39(7):1229-1236. [doi: [10.1080/03610920902871446](https://doi.org/10.1080/03610920902871446)]
14. Costa PS, Santos NC, Cunha P, Cotter J, Sousa N. The use of multiple correspondence analysis to explore associations between categories of qualitative variables in healthy ageing. *J Aging Res* 2013;2013:302163. [doi: [10.1155/2013/302163](https://doi.org/10.1155/2013/302163)] [Medline: [24222852](https://pubmed.ncbi.nlm.nih.gov/24222852/)]
15. Florensa D, Godoy P, Mateo J, et al. The use of multiple correspondence analysis to explore associations between categories of qualitative variables and cancer incidence. *IEEE J Biomed Health Inform* 2021 Sep;25(9):3659-3667. [doi: [10.1109/JBHI.2021.3073605](https://doi.org/10.1109/JBHI.2021.3073605)] [Medline: [33857006](https://pubmed.ncbi.nlm.nih.gov/33857006/)]
16. van Horn A, Weitz CA, Olszowy KM, et al. Using multiple correspondence analysis to identify behaviour patterns associated with overweight and obesity in Vanuatu adults. *Public Health Nutr* 2019 Jun;22(9):1533-1544. [doi: [10.1017/S1368980019000302](https://doi.org/10.1017/S1368980019000302)] [Medline: [30846019](https://pubmed.ncbi.nlm.nih.gov/30846019/)]
17. Śledzińska P, Bebyn MG, Furtak J, Kowalewski J, Lewandowska MA. Prognostic and predictive biomarkers in gliomas. *Int J Mol Sci* 2021 Sep 26;22(19):10373. [doi: [10.3390/ijms221910373](https://doi.org/10.3390/ijms221910373)] [Medline: [34638714](https://pubmed.ncbi.nlm.nih.gov/34638714/)]
18. Sokolov A, Paull EO, Stuart JM. ONE-class detection of cell states in tumor subtypes. Presented at: Proceedings of the Pacific Symposium; Jan 4-8, 2016; Kohala Coast, Hawaii, USA. [doi: [10.1142/9789814749411\\_0037](https://doi.org/10.1142/9789814749411_0037)]
19. Salomonis N, Dexheimer PJ, Omberg L, et al. Integrated genomic analysis of diverse induced pluripotent stem cells from the progenitor cell biology consortium. *Stem Cell Rep* 2016 Jul 12;7(1):110-125. [doi: [10.1016/j.stemcr.2016.05.006](https://doi.org/10.1016/j.stemcr.2016.05.006)] [Medline: [27293150](https://pubmed.ncbi.nlm.nih.gov/27293150/)]
20. Daily K, Ho Sui SJ, Schriml LM, et al. Molecular, phenotypic, and sample-associated data to describe pluripotent stem cell lines and derivatives. *Sci Data* 2017 Mar 28;4:170030. [doi: [10.1038/sdata.2017.30](https://doi.org/10.1038/sdata.2017.30)] [Medline: [28350385](https://pubmed.ncbi.nlm.nih.gov/28350385/)]
21. Malta TM, Sokolov A, Gentles AJ, et al. Machine learning identifies stemness features associated with oncogenic dedifferentiation. *Cell* 2018 Apr 5;173(2):338-354. [doi: [10.1016/j.cell.2018.03.034](https://doi.org/10.1016/j.cell.2018.03.034)] [Medline: [29625051](https://pubmed.ncbi.nlm.nih.gov/29625051/)]
22. Lê S, Josse J, Husson F. FactoMineR: an R package for multivariate analysis. *J Stat Softw* 2008 Mar;25(1):1-18. [doi: [10.18637/jss.v025.i01](https://doi.org/10.18637/jss.v025.i01)] [Medline: [27348562](https://pubmed.ncbi.nlm.nih.gov/27348562/)]
23. The Cancer Genome Atlas program (TCGA). Center for Cancer Genomics. URL: <https://www.cancer.gov/tcga> [accessed 2025-03-06]
24. The Cancer Genome Atlas program. National Cancer Institute. URL: <https://www.cancer.gov/ccg/research/structural-genomics/tcga/history/policies/tcga-human-subjects-data-policies.pdf> [accessed 2025-03-06]
25. Bleeker FE, Atai NA, Lamba S, et al. The prognostic IDH1( R132 ) mutation is associated with reduced NADP+-dependent IDH activity in glioblastoma. *Acta Neuropathol* 2010 Apr;119(4):487-494. [doi: [10.1007/s00401-010-0645-6](https://doi.org/10.1007/s00401-010-0645-6)] [Medline: [20127344](https://pubmed.ncbi.nlm.nih.gov/20127344/)]
26. Chai RC, Zhang KN, Chang YZ, et al. Systematically characterize the clinical and biological significances of 1p19q genes in 1p/19q non-codeletion glioma. *Carcinogenesis* 2019 Oct 16;40(10):1229-1239. [doi: [10.1093/carcin/bgz102](https://doi.org/10.1093/carcin/bgz102)] [Medline: [31157866](https://pubmed.ncbi.nlm.nih.gov/31157866/)]
27. McNulty SN, Cottrell CE, Vigh-Conrad KA, et al. Beyond sequence variation: assessment of copy number variation in adult glioblastoma through targeted tumor somatic profiling. *Hum Pathol* 2019 Apr;86:170-181. [doi: [10.1016/j.humpath.2018.12.004](https://doi.org/10.1016/j.humpath.2018.12.004)] [Medline: [30594748](https://pubmed.ncbi.nlm.nih.gov/30594748/)]
28. Wang H, Zhang X, Liu J, et al. Clinical roles of EGFR amplification in diffuse gliomas: a real-world study using the 2021 WHO classification of CNS tumors. *Front Neurosci* 2024;18:1308627. [doi: [10.3389/fnins.2024.1308627](https://doi.org/10.3389/fnins.2024.1308627)] [Medline: [38595969](https://pubmed.ncbi.nlm.nih.gov/38595969/)]
29. Kurscheid S, Bady P, Sciuscio D, et al. Chromosome 7 gain and DNA hypermethylation at the HOXA10 locus are associated with expression of a stem cell related HOX-signature in glioblastoma. *Genome Biol* 2015 Jan 27;16(1):16. [doi: [10.1186/s13059-015-0583-7](https://doi.org/10.1186/s13059-015-0583-7)] [Medline: [25622821](https://pubmed.ncbi.nlm.nih.gov/25622821/)]
30. Pierscianek D, Kim YH, Motomura K, et al. MET gain in diffuse astrocytomas is associated with poorer outcome. *Brain Pathol* 2013 Jan;23(1):13-18. [doi: [10.1111/j.1750-3639.2012.00609.x](https://doi.org/10.1111/j.1750-3639.2012.00609.x)] [Medline: [22672415](https://pubmed.ncbi.nlm.nih.gov/22672415/)]
31. Mancini R, Pattaro G, Diodoro MG, et al. Tumor regression grade after neoadjuvant chemoradiation and surgery for low rectal cancer evaluated by multiple correspondence analysis: ten years as minimum follow-up. *Clin Colorectal Cancer* 2018 Mar;17(1):e13-e19. [doi: [10.1016/j.clcc.2017.06.004](https://doi.org/10.1016/j.clcc.2017.06.004)] [Medline: [28865674](https://pubmed.ncbi.nlm.nih.gov/28865674/)]
32. Wu T, Zhang S, Guo S, et al. Correspondence analysis between traditional Chinese medicine (TCM) syndrome differentiation and histopathology in colorectal cancer. *Eur J Integr Med* 2015 Aug;7(4):342-347. [doi: [10.1016/j.eujim.2015.07.003](https://doi.org/10.1016/j.eujim.2015.07.003)]
33. Kramer RJ, Rhodin KE, Therien A, et al. Unsupervised clustering using multiple correspondence analysis reveals clinically-relevant demographic variables across multiple gastrointestinal cancers. *Surgical Oncology Insight* 2024 Mar;1(1):100009. [doi: [10.1016/j.soi.2024.100009](https://doi.org/10.1016/j.soi.2024.100009)]



34. Nadjib Bustan M, Arif Tiro M, Annas S. Correspondence analysis of breast cancer diagnosis classification. J Phys Conf Ser 2019 Jun 1;1244(1):012030. [doi: [10.1088/1742-6596/1244/1/012030](https://doi.org/10.1088/1742-6596/1244/1/012030)]
35. Higgs NT. Practical and innovative uses of correspondence analysis. R Stat Soc Ser D (The Statistician) 1991;40(2):183. [doi: [10.2307/2348490](https://doi.org/10.2307/2348490)]
36. Husson F, Le S, Pagès J. Exploratory Multivariate Analysis by Example Using R: CRC Press; 2011.
37. Machine learning identifies stemness features associated with oncogenic dedifferentiation. National Cancer Institute. URL: <https://gdc.cancer.gov/about-data/publications/PanCanStemness-2018> [accessed 2025-03-06]
38. PanCanStem: reproducing mrnasi from PMID: 29625051. GitHub. URL: <https://github.com/ArtemSokolov/PanCanStem> [accessed 2025-03-06]

## Abbreviations

**ASR:** adjusted standardized residual  
**ATRX:** alpha thalassemia/mental retardation syndrome, X-linked  
**Chr:** chromosome  
**Chr7+/Chr10-:** chromosome 7 gain and 10 loss  
**CL:** classical  
**CSC:** cancer stem cell  
**GMB:** glioblastoma  
**IDH:** isocitrate dehydrogenase  
**MCA:** multiple correspondence analysis  
**mDNAsi:** DNA methylation stemness index  
**ME:** mesenchymal  
**MGMT:** methylguanine methyltransferase  
**NE:** neural  
**PR:** proneural  
**TCGA:** the Cancer Genome Atlas  
**TERT:** telomerase reverse transcriptase

*Edited by E Uzun; submitted 21.08.24; peer-reviewed by C Tang, J Lai; revised version received 22.11.24; accepted 04.02.25; published 12.03.25.*

### *Please cite as:*

Goes Job ME, Fukumasu H, Malta TM, Porfirio Xavier PL

*Investigating Associations Between Prognostic Factors in Gliomas: Unsupervised Multiple Correspondence Analysis*

*JMIR Bioinform Biotech* 2025;6:e65645

URL: <https://bioinform.jmir.org/2025/1/e65645>

doi: [10.2196/65645](https://doi.org/10.2196/65645)

© Maria Eduarda Goes Job, Heidge Fukumasu, Tathiane Maistro Malta, Pedro Luiz Porfirio Xavier. Originally published in JMIR Bioinformatics and Biotechnology (<https://bioinform.jmir.org/>), 12.3.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Bioinformatics and Biotechnology, is properly cited. The complete bibliographic information, a link to the original publication on <https://bioinform.jmir.org/>, as well as this copyright and license information must be included.

Original Paper

# Decentralized Biobanking Apps for Patient Tracking of Biospecimen Research: Real-World Usability and Feasibility Study

William Sanchez<sup>1</sup>, BA; Ananya Dewan<sup>2</sup>, BA; Eve Budd<sup>3</sup>; M Eifler<sup>1</sup>, BA, MFA; Robert C Miller<sup>4,5</sup>, MD; Jeffery Kahn<sup>6</sup>, PhD; Mario Macis<sup>7</sup>, PhD; Marielle Gross<sup>1,6</sup>, MD

<sup>1</sup>de-bi, co., Greencastle, PA, United States

<sup>2</sup>Johns Hopkins School of Medicine, Johns Hopkins University, Baltimore, MD, United States

<sup>3</sup>Harpur College of Arts and Sciences, State University of New York, Binghamton, NY, United States

<sup>4</sup>Faculty of Medicine, Mayo Clinic, Rochester, MN, United States

<sup>5</sup>School of Medicine, Indiana University Hospital, Indianapolis, IN, United States

<sup>6</sup>Johns Hopkins Berman Institute of Bioethics, Johns Hopkins University, Baltimore, MD, United States

<sup>7</sup>Carey School of Business, Johns Hopkins University, Baltimore, MD, United States

**Corresponding Author:**

Marielle Gross, MD

Johns Hopkins Berman Institute of Bioethics

Johns Hopkins University

1809 Ashland Ave.

Baltimore, PA, 17225

United States

Phone: 1 8135416103

Email: [mariellesophiagross@gmail.com](mailto:mariellesophiagross@gmail.com)

## Abstract

**Background:** Biobank privacy policies strip patient identifiers from donated specimens, undermining transparency, utility, and value for patients, scientists, and society. We are advancing decentralized biobanking apps that reconnect patients with biospecimens and facilitate engagement through a privacy-preserving nonfungible token (NFT) digital twin framework. The decentralized biobanking platform was first piloted for breast cancer biobank members.

**Objective:** This study aimed to demonstrate the technical feasibility of (1) patient-friendly biobanking apps, (2) integration with institutional biobanks, and (3) establishing the foundation of an NFT digital twin framework for decentralized biobanking.

**Methods:** We designed, developed, and deployed a decentralized biobanking mobile app for a feasibility pilot from 2021 to 2023 in the setting of a breast cancer biobank at a National Cancer Institute comprehensive cancer center. The Flutter app was integrated with the biobank's laboratory information management systems via an institutional review board-approved mechanism leveraging authorized, secure devices and anonymous ID codes and complemented with a nontransferable ERC-721 NFT representing the *soul-bound* connection between an individual and their specimens. Biowallet NFTs were held within a custodial wallet, whereas the user experiences simulated token-gated access to personalized feedback about collection and use of individual and collective deidentified specimens. Quantified app user journeys and NFT deployment data demonstrate technical feasibility complemented with design workshop feedback.

**Results:** The decentralized biobanking app incorporated key features: "biobank" (learn about biobanking), "biowallet" (track personal biospecimens), "labs" (follow research), and "profile" (share data and preferences). In total, 405 pilot participants downloaded the app, including 361 (89.1%) biobank members. A total of 4 central user journeys were captured. First, all app users were oriented to the ≥60,000-biospecimen collection, and 37.8% (153/405) completed research profiles, collectively enhancing annotations for 760 unused specimens. NFTs were minted for 94.6% (140/148) of app users with specimens at an average cost of US \$4.51 (SD US \$2.54; range US \$1.84-\$11.23) per token, projected to US \$17,769.40 (SD US \$159.52; range US \$7265.62-\$44,229.27) for the biobank population. In total, 89.3% (125/140) of the users successfully claimed NFTs during the pilot, thereby tracking 1812 personal specimens, including 202 (11.2%) distributed under 42 unique research protocols. Participants embraced the opportunity for direct feedback, community engagement, and potential health benefits, although user onboarding requires further refinement.

**Conclusions:** Decentralized biobanking apps demonstrate technical feasibility for empowering patients to track donated biospecimens via integration with institutional biobank infrastructure. Our pilot reveals potential to accelerate biomedical research through patient engagement; however, further development is needed to optimize the accessibility, efficiency, and scalability of platform design and blockchain elements, as well as a robust incentive and governance structure for decentralized biobanking.

(*JMIR Bioinform Biotech* 2025;6:e70463) doi:[10.2196/70463](https://doi.org/10.2196/70463)

## KEYWORDS

patient empowerment; biobanking; biospecimens; transparency; community engagement; nonfungible tokens; NFTs; blockchain technology; decentralized biobanking; pilot studies; technical feasibility; biowallet

## Introduction

### Background

University biobanks collect, store, and distribute biospecimens such as tissue and blood, capitalizing on leftover clinical materials from affiliated hospitals to drive biomedical science and drug discovery [1-3]. Standard operating procedure for most biobanks in academic medical centers includes prospective broad consent for nonspecific, future research [4] coupled with deidentification, whereby identifiers are stripped before specimen allocation [5]. In this setting, patients do not learn what becomes of their donations, and scientists lack access to the donor, linked specimens, and evolving clinical data [4,6]. This disconnect, though the by-product of policies designed to protect privacy while promoting learning, promulgates a biobank ecosystem that permits problematic gaps in recognition, reciprocity, and return of results [7,8]. Simultaneously, vast yet siloed specimen collections have accumulated across most US academic medical centers, a widely underused and unsustainable “treasure trove” wherein frozen assets lay hidden from patients and scientists for whom they may be most valuable [3,9]. The lack of an efficient market for ensuring the use of donated materials deepens the crisis of faith in public health institutions and has prompted attempts at marketplace solutions [10,11].

We are advancing *decentralized biobanking* as a software platform predicated on blockchain technology’s democratic ethos, incentive alignment, transparency, and assurances of trust [12]. These key features are reflective of blockchains as permissionless, distributed, shared ledgers of digital transactions engineered to be mathematically concordant, accessible, and auditable [13], underscoring their first and most successful use to date for the creation of global digital currency such as Bitcoin, which makes them fit for purpose in efforts to decentralize ownership and governance of data through thoughtfully structured peer-to-peer networks [14]. One of the most promising innovations enabled by blockchains are nonfungible tokens (NFTs), digital record identifiers that serve as electronic deeds for provably unique digital or physical assets that may be represented “on-chain” [15]. The potential for blockchain and NFTs to play a role in restructuring control and ownership of data has been widely discussed, with several notable projects in the health care domain [16,17]. Although empowering patient ownership of health data is compelling in theory, full realization of such initiatives has been elusive in light of complex regulatory considerations, socioeconomic factors, and technical limitations for blockchain technologies and legacy systems [18,19].

Building on the success and diversity of blockchain applications for decentralized finance [20,21], decentralized biobanking applies human-centered design and innovative system mechanisms to empower patients to track donated biospecimens and engage in downstream research activities, outcomes, and products via a platform compatible with established privacy policies and workflows. Our approach provides patients with secure, direct access to personal specimen data housed in institutional databases via user-friendly mobile and web apps complemented with a privacy-preserving NFT digital twin framework [22]. This strategy may support stepwise adoption of increasingly autonomous and progressively decentralized collaborations among patients, scientists, and physicians in a dynamic biomedical metaverse, or “biomediverse.”

### Objectives

Successful implementation of decentralized biobanking will usher in a new standard for research transparency, foster institutional accountability to the patients and communities they serve, and create opportunities to unite siloed datasets, facilitate timely translation of precision medicine and enable structurally just marketplace solutions for improving efficiency and effectiveness in the management of one of our most precious human resources. In this paper, we explore the technical feasibility of decentralized biobanking through a description and quantitative analysis of a live pilot for a breast cancer biobank at a US academic medical center. We discuss system design, key features, and NFT functionality, illustrating how the platform provided transparency and recognition of patients’ contributions to a real-world biobank.

## Methods

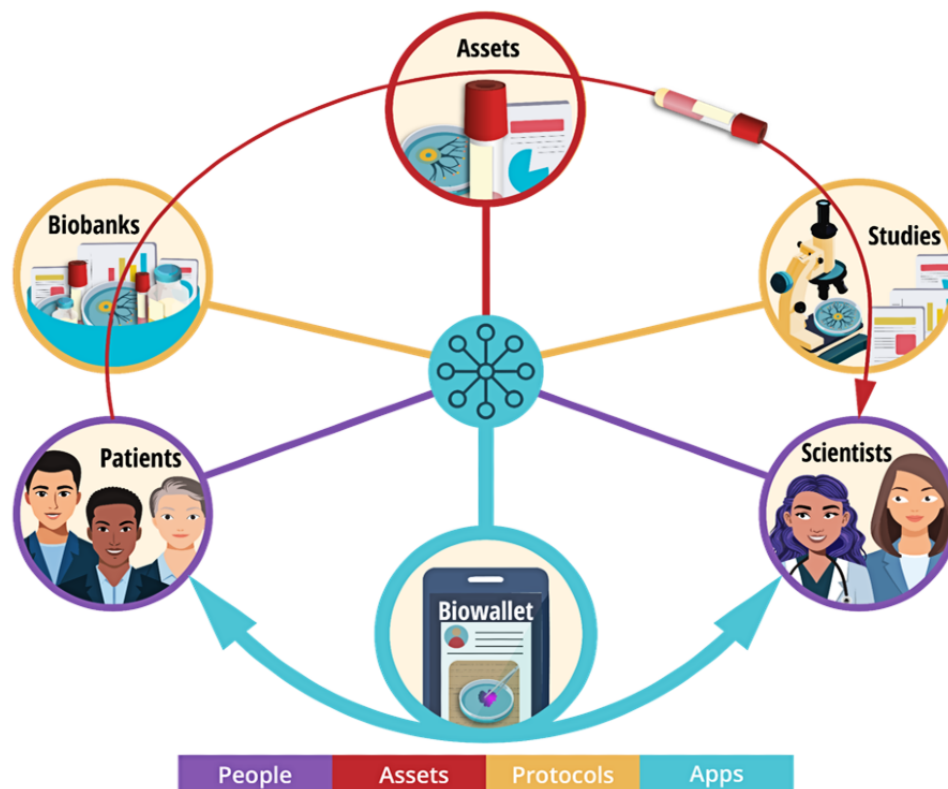
### Decentralized Biobanking System Design: NFT Digital Twin Framework

Decentralized biobanking builds digital bridges among patients, specimens, and scientists, connecting stakeholders based on real-world relationships predicated upon transactions within existing biobank infrastructure and research protocols (Figure 1). The system design represents all people, protocols, and assets in an NFT digital twin framework, creating a blockchain-backed overlay network on top of the established biospecimen ecosystem. Our approach presents a unique strategy for the progressive inclusion of patients, allowing for the implementation of a composable software platform with programmable, modular elements, mechanisms, and workflows that may be integrated with institutional biobank databases to provide durable transparency without requiring substantial time,

labor, or ongoing participation of physicians, biobankers, and scientists. This framework applies privacy by design throughout the engineering process, implementing techniques such as data minimization and innovative system architectures to ensure compliance with established biospecimen collection and research

protocols, institutional policies, and data structures. The core benefits of our approach are use case agnostic and can be applied for all biobanks, research protocols, and institutions with minor modifications at each new site.

**Figure 1.** Decentralized biobanking system design—nonfungible token (NFT) framework and software applications uniting patients, specimens, and scientists. This system diagram illustrates key entities of biobanking connected via a specimen supply chain (red arrow) yet presently lacking a unified platform for collaboration. The proposed decentralized biobanking NFT digital twin framework is designed to integrate with this established infrastructure, mapping the stakeholders, specimens, and studies in the biobanking ecosystem and enabling applications whereby they may be united for mutually beneficial collaboration, data exchange, and value-building activities.



## Pilot Setting

The Breast Disease Research Repository (BDRR; STUDY19060196) is a large breast cancer biobank platform at the intersection of the University of Pittsburgh, the University of Pittsburgh Medical Center, and Hillman Cancer Center that served as the pilot study use case. Broad prospective consent for the BDRR is embedded in the breast cancer service line, for example, concurrently with surgical consent. Once consented, “leftovers” from any clinical procedures may be collected by the biobank without further notice or engagement. From 2006 to 2023, more than 10,000 patients consented for the BDRR and specimens were collected from 4000 participants to date. In total, approximately 61,000 specimens were collected, and 6000 were distributed for research, with a mix of fresh and frozen distributions. The biobank operates via a hub-and-spoke model, allocating specimens chiefly to local investigators under designated research or subbiobanking protocols (eg, a flagship patient-derived organoid biobank that grows and distributes copies of living 3D cell cultures [approximately n=300]).

## Requirement Gathering

Foundational surveys, semistructured interviews, community engagement, and stakeholder alignment activities with

populations with breast cancer, physicians, advocates, and scientists informed our approach to designing a biobanking app for patients [23]. Broadly, we found that patients have an unmet demand for feedback about research on their specimens, with particular interest surrounding personal meaning or potential health benefits for the individual or their family members. For example, a survey respondent noted the following:

*Giving patients access to this type of information could decrease the lethal lag between research findings and actual clinical practice.*

One patient advocacy leader captured this sentiment, noting the following:

*We have been screaming for this, banging on pots and pans. Thank you for taking this on.*

Importantly, she alluded to the multifactorial challenge of enabling patients to track and learn about donated biospecimens [23], which would require novel, user-friendly interface designs as well as system architectures and pilot protocols compliant with regulatory norms, compatible with established workflows, and acceptable within the institutional milieu.



Thus, we interacted extensively with the breast cancer service line, the institutional biobanking platform, and institutional review board (IRB) and Office of Human Research Protections leadership, as well as research scientists, clinical and teaching faculty, IT staff, technology transfer teams, and cross-disciplinary institutional leadership. Concurrently, the ethnography of the specimen procurement supply chain allowed us to map the breast cancer biobank ecosystem [23]. We examined all contexts along the data pipeline, from population-level breast cancer screening to diagnostic biopsies and surgical treatments, clinical pathology, and specimen accessioning through the biobanking platform, where it may be stored for future use in  $-80^{\circ}\text{C}$  freezers or distributed fresh for next-generation biobanking applications such as patient-derived organoids, multi-omics, and high-throughput testing. Given the well-documented challenges for biobank sustainability, we took special interest in learning about economic and logistical challenges pertaining to this sector. Regulatory considerations, operational feasibility, and economic analyses will be reported elsewhere [23].

### Prototyping

The first decentralized biobanking prototype established the proof of concept, leveraging ERC-721 NFTs to keep patients connected to donated specimens throughout the research life cycle. The NFT platform was integrated with a novel mobile app for privacy-preserving collaboration among patients, scientists, and physicians in a model breast cancer organoid ecosystem. A second prototype advanced a comprehensive NFT digital twin framework with ERC-1155 modeled using a publicly available real-world organoid biobank dataset (National Cancer Institute Human Cancer Models Initiative) [24,25]. This web-based prototype focused on generating value for scientists, illustrating potential to enhance efficiency, effectiveness, and impact of biospecimen research. Third, no-code front-end mobile app prototypes were developed to demonstrate, test, and refine user interfaces and experiences for the engagement of donors in biobanking.

### User Interface and User Experience

We drafted wireframes using anonymous model biospecimen information from the institutional biobank database. App design processes sought to minimize cognitive effort for mobile app users, maximize accessibility across ages and educational levels, and adhere to rigorous privacy standards and customs in accordance with the established biospecimen collection protocols. We progressively simplified and iterated display text and content to make it as concise and concrete as possible and unified across decentralized biobanking app interfaces. To facilitate navigation, we streamlined presentation of content in each of the 4 core interfaces using accordion elements complemented with individual cards for each biospecimen, with pop-ups to guide transitions within and across interfaces. Unified color schemes, fonts, and item designs adhered to predetermined themes with a standardized format that was gradually refined.

The designs were tested and validated via further research surveys and interviews. Immersive design workshops solidified core app requirements. Initial usability testing included online and in-person sessions with clickable prototypes and functional

prototype demonstrations followed by usability testing and cognitive walk-throughs on users' personal devices.

### Front-End Development and Testing

Finalized mobile app designs were developed using Flutter so that iOS and Android users could participate in the pilot. The apps were tested and deployed to Apple TestFlight and the Google Play Store, allowing for download directly to participants' personal devices. From August 2022 to January 2023, feedback from 110 unique individuals was incorporated, including 45 (40.9%) BDRR members, 28 (25.5%) who downloaded and tested the app on their personal devices, and 14 (12.7%) who viewed personalized biospecimen content within the app interface. The result was a validated app facilitating interaction between donors and biospecimens within the breast cancer biobank, personalized collection content, and mappings from biobank database details.

### Blockchain Development

Initial decentralized biobanking prototypes were developed experimenting with different tokenization strategies using Ethereum's ERC-721 and ERC-1155 NFT standards for mapping dynamic relationships among patients, biospecimens, physicians, scientists, and corresponding biobanking and research protocols. However, variable costs of transaction fees (known as gas fees) on the Ethereum network and high friction for blockchain onboarding were major limitations for implementing a real-world pilot.

These constraints informed the design of a functional, blockchain-backed prototype suitable for the pilot population and setting, leveraging a fit-for-purpose blend of centralized and decentralized applications that would enable patients to track and learn about donated specimens appropriate to the highest-order objectives for the first live pilot of decentralized biobanking technology.

A nontransferable ERC-721 NFT, also referred to as a "soul-bound token" [26], was developed to represent each donor's immutable, inherently unique connection to their personal biospecimens. This token [26] was held within a single *externally owned account* that served as a custodial wallet. Of note, our previous decentralized biobanking prototype for organoid research networks, as described elsewhere, used ERC-1155 to advance a comprehensive digital twin ecosystem with NFTs representing patients, specimens, multigenerational derivatives (eg, patient-derived organoids), scientists, and physicians, as well as externally owned biobanker accounts, demonstrating the potential for a sophisticated solution [25]. However, while using the ERC-1155 standard would have offered savings for deploying multiple token collections representative of the entire biobanking ecosystem, applying them to a single soul-bound token collection for this use case would have yielded no additional benefits while adding unnecessary complexity [25].

Each biowallet NFT served as a customized yet anonymous "token of appreciation" for specimen donation coupled with a front-end user experience simulating token-gated access to personal biobank data. This token-gated process was performed manually, minting the tokens individually via the smart contract

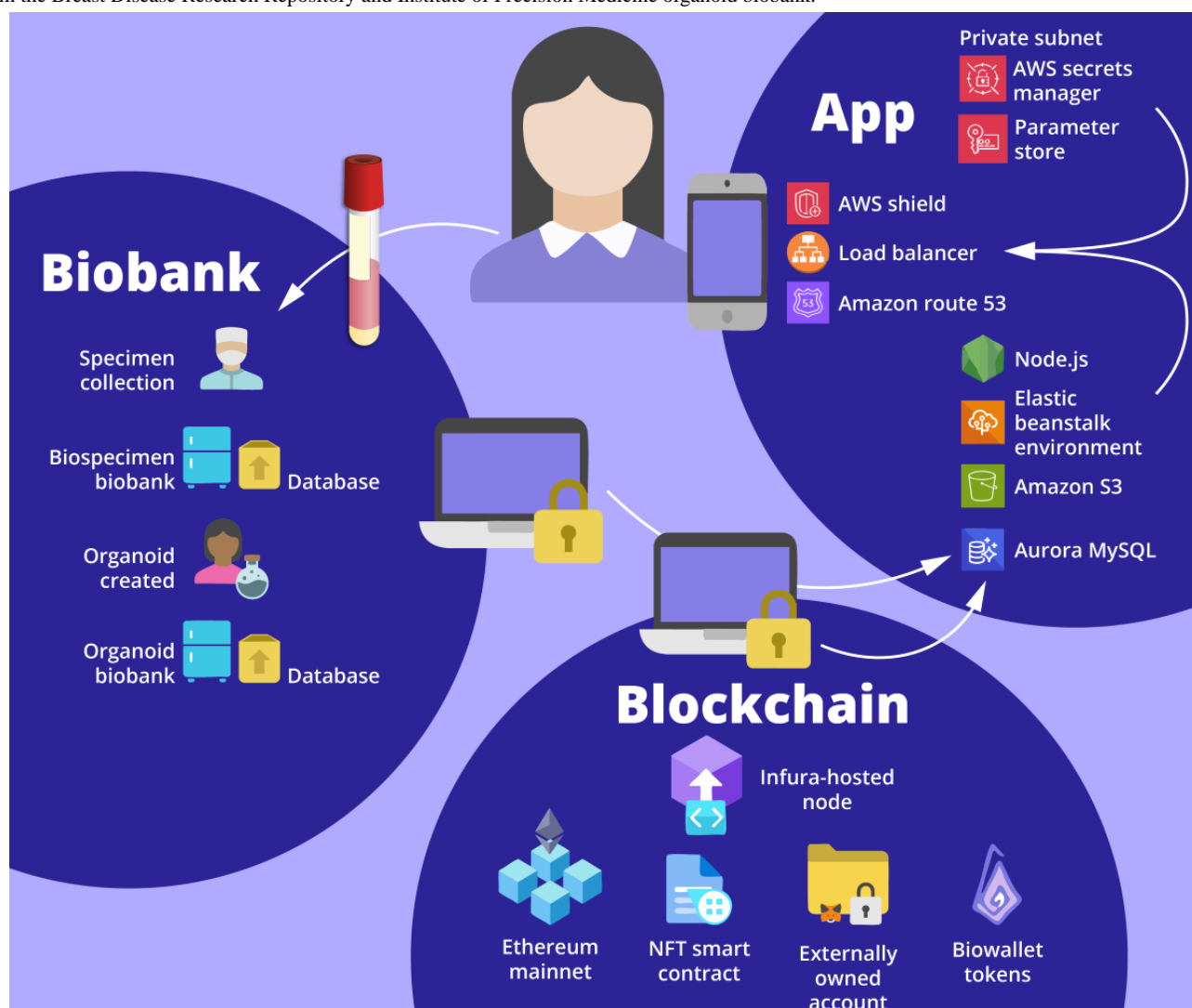
interface on Etherscan. Subsequently, the token metadata and transaction details were stored within a secure, IRB-approved database for the eligible user. This created a digital honest broker mechanism for managing in-app participant-specimen engagement without requiring further humans in the loop or revealing donor names or other personally identifiable information to third parties.

### System Architecture

The decentralized biobanking pilot system incorporated 3 core components: an *app* overlying *institutional biobank and research infrastructure* with a blockchain-backed *NFT digital twin framework* (Figure 2).

The app used an *n*-tier architecture pattern with interconnected workflows across distinct, modular components with varying responsibilities (Table 1). Our user-friendly mobile app, available on Android and iOS, was powered by applications built using Amazon Web Services. During this initial pilot phase, our system relied on external services and data sources that were not yet directly integrated with our deployed technology. Our NFT framework consisted of an ERC-721 smart contract designed to mint nontransferable, soul-bound biowallet tokens that were deployed to the Ethereum mainnet. Deidentified biospecimen data were provided by biobank personnel to authorized study team members, who would use a secure device to import the records into the pilot system's database. Both required manual processes for pilot implementation.

**Figure 2.** System architecture diagram—decentralized biobanking pilot app for breast cancer biobank. This system architecture diagram incorporates the decentralized biobanking mobile app powered by internal components that handle business logic, data storage, and data integrations built on a cloud-based infrastructure using Amazon Web Services (AWS); this is flanked by corresponding elements connected via secure authorized access devices for interacting with the nonfungible token (NFT) digital framework's biowallet tokens deployed on Ethereum and institutional data sources from the Breast Disease Research Repository and Institute of Precision Medicine organoid biobank.





**Table 1.** Key details of the decentralized biobanking pilot system architecture.

Component	Technical details
App	<ul style="list-style-type: none"><li>• Presentation tier: the Flutter mobile app built and deployed using Android Studio (Google) and Xcode (Apple Inc) to enable download to Android and iOS devices. The app provided front-end user interfaces for patients, enabling dynamic interactions, user inputs, and the presentation of queried information from institutional data sources through the app tier. Google’s Firebase Authentication services manage account creation and management, encrypting data in transit using HTTPS and at rest using the scrypt standard cryptographic protocol. Passwords are stored securely using encryption, salting, and 1-way hashing following NIST<sup>a</sup> 800-63b recommendations.</li><li>• App tier: used a Node.js (OpenJS Foundation) server to enable all core functionality and logic of the app, including specimen tracking with enhanced transparency into biobank activities and subsequent research. This layer is also responsible for enforcing security and access rules, handling connectivity to and communication with data sources and external services, and processing data to return to the presentation layer. Deployed on AWS<sup>b</sup> Elastic Beanstalk, the app instances sit behind load balancers for scalability, running in private subnets.</li><li>• Data tier: hosted by an Amazon Aurora database cluster using the MySQL engine. It hosts a secure, highly available database that stores and retrieves the information necessary for the app to run. This includes donated sample records housed on the BIOS<sup>c</sup> and corresponding biospecimen freezer repositories across 4 physical locations of the Pitt Biospecimen Core, as well as unique cryptographic IDs from Firebase and claimed biowallet NFTs<sup>d</sup> to establish privacy-preserving data linkages between donors and their deidentified biospecimens. As noted in the presentation tier, user credentials for accessing the app are stored separately on secure Firebase servers.</li><li>• Infrastructure tier: referenced within the app and data tiers, our AWS cloud infrastructure provides the foundation for networking and security, ensuring availability, scalability, and interoperability across system components <a href="#">Multimedia Appendix 1</a>.</li></ul>
Blockchain	<ul style="list-style-type: none"><li>• NFT framework: an ERC-721 smart contract designed to mint nontransferable, soul-bound biowallet tokens was deployed to the Ethereum mainnet via a transaction sent to an Infura-hosted node from a local Node.js runtime environment using Hardhat. The overarching framework incorporates NFTs representing all stakeholders, specimens, and protocols, allowing for composable layers of complexity, utility, and value to be built upon the PIO<sup>e</sup> architecture.</li></ul>
Biobank	<ul style="list-style-type: none"><li>• Institutional biospecimen and research databases: biobank personnel provided access to deidentified biospecimen data via OneDrive Microsoft Excel (Microsoft Corp) files to an authorized study team member, who would use a secure device to import the updated records into the Aurora database. Similarly, Microsoft Excel files containing biobank (BDRR<sup>f</sup>) registered members were provided by research staff as exported from OnCore. In addition, imaging and research data from an organoid biobank “spoke” were shared via OneDrive, and curated representative datasets were hosted on Dropbox (Dropbox, Inc).</li></ul>

<sup>a</sup>NIST: National Institute of Standards and Technology.

<sup>b</sup>AWS: Amazon Web Services.

<sup>c</sup>BIOS: Biospecimen Inventory and Operations System.

<sup>d</sup>NFT: nonfungible token.

<sup>e</sup>PIO: programmed input-output.

<sup>f</sup>BDRR: Breast Disease Research Repository.

**Pilot Study**

Participants were recruited via electronic and paper fliers for “Decentralized Biobanking “de-bi”: An App for Patient Feedback from Biobank Research Donation” (STUDY22020035). The pilot aimed to recruit 300 participants over 6 to 12 months. App download invites were distributed via email with Apple and Android instructions. IT support was provided as needed, with real-time bug fixes and improvements based on user feedback. App interfaces, design, and features were iterated in monthly sprints. Participatory research, user-centered design, and usability testing, as well as quantitative and qualitative assessments of patient, physician, and scientist acceptability, will be reported elsewhere. NFT minting for pilot performance took place from March 7, 2023, to May 8, 2023. [Multimedia Appendix 2](#) details the pilot recruitment to sample tracking process.

**Data Sources and Analysis**

The technical data reviewed included conceptual models, technical diagrams, product feature documentation, and

screenshots of user journeys as experienced by decentralized biobanking pilot participants using the Flutter app. We also consider biospecimen collection data from the institutional Biospecimen Inventory and Operations System via Microsoft Excel (Microsoft Corp) exports, in-app activity data recorded in a MySQL database, and blockchain transactions on the Ethereum network accessed via Etherscan. Technical feasibility was assessed from feature requirements, interface designs, and quantifiable user experiences from the live implementation. To further evaluate pilot outcomes, we provide simple descriptive statistics from the quantitative datasets and comparative cost analyses for alternative NFT design strategies calculated using values from tokens minted during the pilot. Patient experiences were captured via written feedback from a co-design workshop during the app development phase and a usability workshop session held with pilot participants.

**Ethical Considerations**

Research was performed under IRB-approved human subjects research protocols and a Quality Improvement protocol ([Textbox 1](#) provides protocol numbers, titles, and approving body).

Participants provided informed consent or the equivalent, in accordance with respective protocols. Conflict of interest disclosures were included in consent documents and verbal disclosures were provided for all online and in-person encounters. All data reported here are either de-identified or anonymized and privacy-by-design was utilized within the de-bi app to maintain confidentiality of participant identities.

Participants were not compensated for participation in the biobank, stakeholder interviews, quality improvement activities or de-bi app pilot study (STUDY19060196, IRB00019273, QRC 3958 and STUDY22020035, respectively). Our foundational research protocol (STUDY22010118) provided \$10 gift cards for surveys, with an additional \$20 for those who completed follow-up interviews.

**Textbox 1.** Human participants and quality improvement protocols for technology feasibility.

- STUDY22010118: patient views, preferences and engagement in next-generation biobank research (University of Pittsburgh)
- IRB00019273: nonfungible tokens for ethical, efficient and effective use of biosamples (Johns Hopkins University)
- STUDY19060196: Breast Disease Research Repository: tissue and bodily fluid and medical information acquisition protocol (04-162; Hillman Cancer Center)
- QRC 3958: patient-facing biobank platform development Quality Improvement proposal for Beckwith award–breast cancer supply chain analysis, biobank token model development, and initial pre-pilot testing with University of Pittsburgh Medical Center patients (University of Pittsburgh Medical Center)
- STUDY22020035: decentralized biobanking “de-bi”: exploring patients interests in feedback, education, follow-up, engagement and tokens of appreciation regarding biobank donation via mobile and web applications (University of Pittsburgh)

## Results

### Prepilot Results

A co-design session (n=15) was conducted before the pilot to characterize patient preferences and areas of confusion. This session was one in a series of extensive participatory design sessions, which we have reported elsewhere [23]. Participants were most excited about decentralized biobanking for feedback and recognition (“to see my own cells+know how those cells are advancing science”), community-engaged research (“to

connect with others through this app”), and precision medicine potential (“to get helpful results regarding my health”), suggesting acceptance of our vision and overall approach. At the conclusion of this phase, there was still confusion surrounding logistics and governance (“how we *find* our samples and approve their use”), technical concepts (“Why NFT’s?”), and unanswered big-picture questions (“Short+long-term—who benefits from this?”) regarding the decentralized biobanking platform. [Table 2](#) provides a thematic overview and representative quotes.

**Table 2.** A thematic overview of participant feedback gathered through a prepilot co-design session.

Theme	Prepilot participant feedback
<b>Aspects participants were “most excited about”</b>	
Personalized feedback and recognition	<ul style="list-style-type: none"><li>• “The opportunity to see my own cells+know how those cells are advancing science and clinical care.”</li><li>• “Having knowledge about [sample] types, research and current news about my tumors.”</li><li>• “To be able to follow where my personal donation goes, and what they are doing with it, and what they get out of it.”</li></ul>
Community-engaged re-search	<ul style="list-style-type: none"><li>• “Great for mutation studies with multiple primary cancer+tumors.”</li><li>• “Keeping up to date with genetic mutation research.”</li><li>• “I’m excited to connect with others through this app.”</li><li>• “That patients who invest their tissue in research are able to connect as co-investigators.”</li></ul>
Potential health benefits	<ul style="list-style-type: none"><li>• “I’m excited about the idea that there may be more ways to care for my family—better research practices may enable the medical field to work smarter—maybe ensuring that my children don’t need surgery, chemo, etc.”</li><li>• “I am very excited for anything that can improve my health and outcome (and of others).”</li><li>• “Being able to get helpful results regarding my health.”</li><li>• “I’m excited about the possibility to know how my tissue reacted to a treatment.”</li><li>• “Patient access to personal info/data; Personalized medicine potential.”</li></ul>
<b>Aspects participants “still found confusing”</b>	
Big picture	<ul style="list-style-type: none"><li>• “Why do people still get cancer, dammit!”</li><li>• “I don’t understand 1) How this may really help me+my family, 2) Short+long-term—who benefits from this? 3) Where does the \$ come from? 4) What are we giving up/sacrificing by saying ‘yes.’”</li><li>• “How will Dr. utilize?”</li></ul>
Logistics and governance	<ul style="list-style-type: none"><li>• “I don’t understand how we find our samples and approve their use—I also don’t understand what studies we could ‘suggest’ or enable through the samples we have provided.”</li><li>• “How likely is it that my samples will be used?”</li><li>• “Can you use it [de-bi app] even if your surgery already happened?”</li><li>• “How to get my tissue submitted to researchers.”</li></ul>
Unclear technical terms and concepts	<ul style="list-style-type: none"><li>• “Not really sure what an organoid is—is it a picture/video of my actual cells or is it a model of my cells?”</li><li>• “Why NFT’s?”</li><li>• “I am still learning about NFTs and how they will help breast cancer patients.”</li><li>• “How will patients interpret data—will it be translated?”</li></ul>

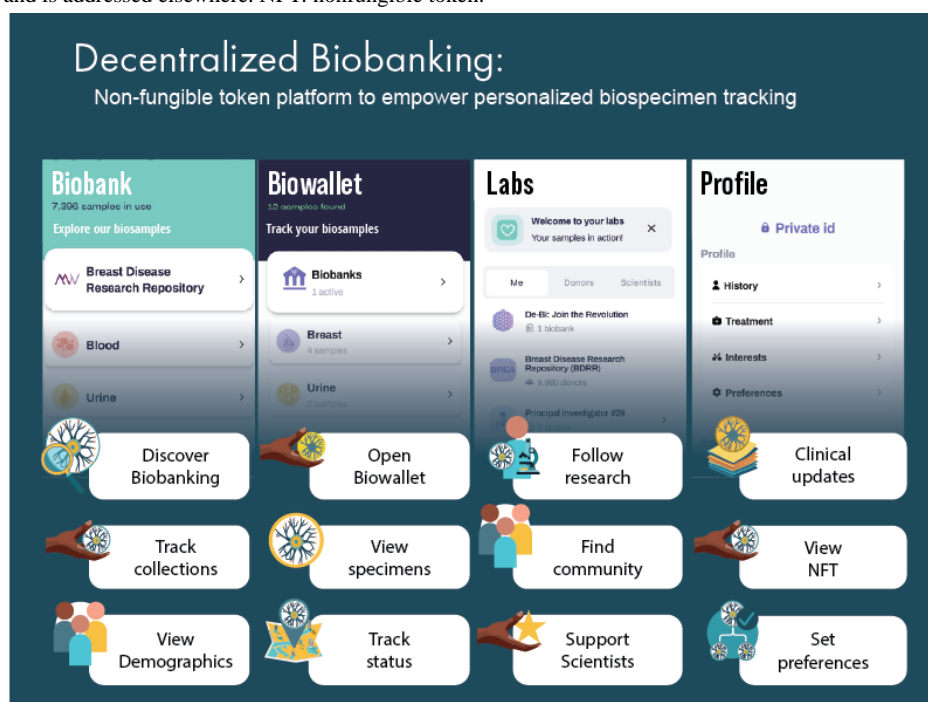
Overall Pilot Results

Overview

Over 10 weeks of active recruitment (February 16 to April 30, 2023), 1080 unique participants enrolled in the decentralized biobanking pilot, including 9.54% (930/9750) of confirmed biobank members (Multimedia Appendix 3). Approximately 600 app invites were distributed, and 405 participants downloaded and completed app registration, including 361 (89.1%) biobank members. All app users were female (405/405, 100%), and the mean age was 56 (SD 12.8; range 18-87) years,

making them younger than both the broader biobank membership and decentralized biobanking pilot participants (mean ages of 64, SD 13.6 and 58, SD 13.1 years, respectively). Multimedia Appendices 4 and 5 detail pilot participant and app user characteristics relative to those of the overall biobank membership. There were 4 key features of the piloted app, as shown in the user journey map (Figure 3). Biobank, biowallet, and profile features and quantified user journeys are illustrated in subsequent Journey sections, and laboratory features and respective user journeys for that context are also described in detail elsewhere.

**Figure 3.** Decentralized biobanking platform user journey. The user journey map demonstrates the status quo of the patient experience with biobank donation as well as the 4 key features of the decentralized biobanking mobile app that was piloted for a large breast cancer biobank member population from January 2023 to May 2023. Each of the columns represents primary activities within the different core screens of the decentralized biobanking mobile app, which the invited participants downloaded to personal iOS and Android devices. The Biobank, Biowallet, and Profile sections are illustrated with key activities and features. The Lab section on the far right is illustrated, although the journey for the community engagement feature is outside the scope of this study and is addressed elsewhere. NFT: nonfungible token.

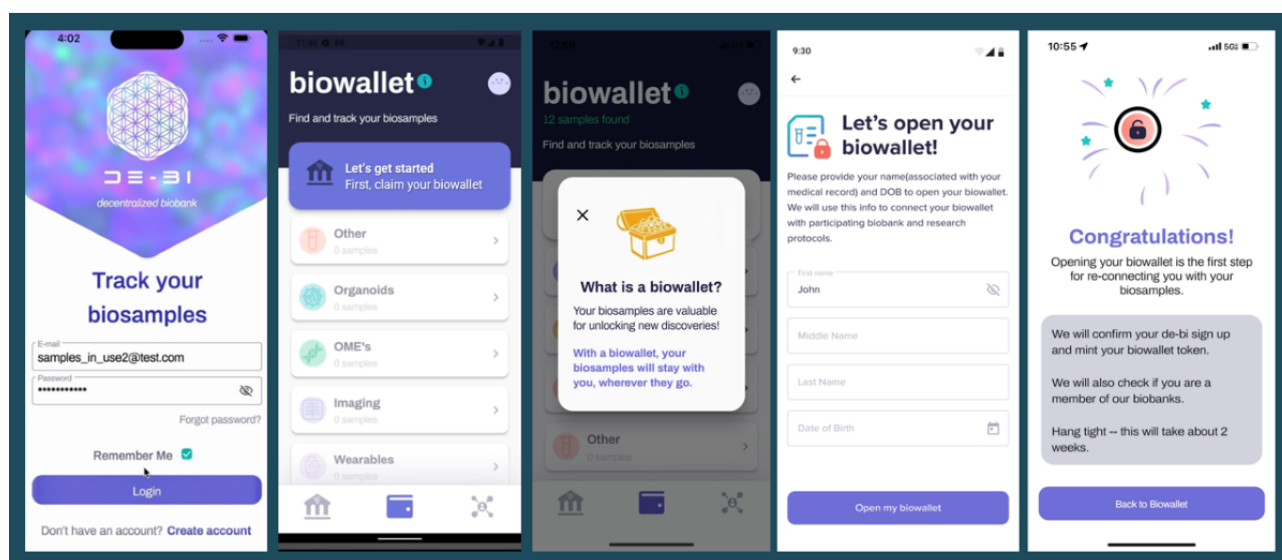


### *Journey 1: App Onboarding and Biowallet NFT Minting Process*

Upon downloading the app, users entered their name and birth date, triggering verification of biobank membership and sample

collections, with “biowallet NFT” minting, if applicable, serving as a digital representation of membership in the biobank donor community, delivering a user experience of a token-gated bridge between the user’s app and specimen data, if available (Figure 4).

**Figure 4.** Opening a biowallet—simulation of token-gated specimen access. The process of opening a biowallet required participants to enter their name and date of birth, triggering the system to match participants to corresponding members in the biobank (Breast Disease Research Repository). Once specimen status was established, biowallet nonfungible tokens were minted, specimens were linked to the account, and email notifications indicated to participants that their biowallet was available.



## Simulated Token Gating Workflows

Once users entered their name and date of birth into the decentralized biobanking app, a manual, coordinated effort involving biobank personnel and authorized study team members verified each user's biobank consent and matched donors to their respective biospecimens via a unique anonymous study ID linked to a Firebase (Google) unique ID associated with their decentralized biobanking app account. During this process, study team members would also mint a unique biowallet token for each verified donor with specimens. These tokens were held in a custodial wallet, but each token identifier was linked to donor records within the Amazon Aurora database to establish a second privacy-preserving mechanism for data linkage.

Firebase established the functional linkage to allow for proper access control and permission management within the app for this pilot, whereas the biowallet NFTs and the act of claiming were representative as a proof of concept as well as a token of appreciation for participating donors. This decision was made to limit excess complexity related to using web3 technologies

as a barrier to participation for this population while providing a comprehensible introduction to the concept of NFTs for establishing relationships between donors and their samples. Our aim was to ensure that donors were not excluded from engaging with the platform based on the extent of their blockchain expertise.

Various criteria for minting Biowallet tokens were considered for entire pilot and biobank deployment. Using variation in token minting costs observed throughout the pilot study to model minimum, average, and maximum costs (US \$1.84, US \$4.51, and US \$11.23, respectively), the selected model, minting tokens for all 272 pilot participants coenrolled in the biobank with one or more specimens collected, was projected to cost US \$1226.72 (SD US \$41.91; range US \$500.48-\$3054.56, [Figure 5A](#), left). Extended entire biobank implementation, this model is projected to cost US \$17,769.40 (SD US \$159.52; range US \$7265.62-\$44,229.27; [Figure 5A](#), right). Other models, such as specimen distribution to a research protocol or biobank membership were also considered.

**Figure 5.** Nonfungible token (NFT) minting costs and calculations for the breast cancer biobank pilot. (A) Pilot implementation—comparison of biowallet token minting criteria for the total cost of pilot deployment. Cost analysis used variation in token minting costs observed throughout the pilot study to model minimum, average, and maximum costs (US \$1.84, US \$4.51, and US \$11.23, respectively). \*Selected token minting criteria for the decentralized biobanking pilot. (B) Transaction costs in US \$ and ether (ETH) are illustrated for 151 NFTs minted during the decentralized biobanking pilot. (C) Timeline mapping variable cost of biowallet minting events and cumulative costs of minting 151 NFT biowallet tokens throughout the decentralized biobanking pilot.



## Token Minting Costs

The cost of deployment of the biowallet NFT protocol on Ethereum was US \$223.52. A total of 151 biowallet tokens were minted for US \$680.49 at an average of US \$4.51 per token (SD US \$2.54; range US \$1.84-\$11.23; [Figure 5B](#)). Biowallet tokens could be requested by decentralized biobanking pilot

participants who downloaded the app and had one or more specimens collected (148/405, 36.5%). For context, procurement, processing, storing, and disbursement of biospecimens in this institutional biobanking platform costs an estimated US \$1600 per case.



Biowallet tokens could be requested by decentralized biobanking pilot participants who downloaded the app and had one or more specimens collected (148/405, 36.5%). During the pilot, 140 total tokens were requested and minted for eligible participants. Minting events varied in cost based on fluctuating transaction fees and the number of participants who had requested biowallet tokens since the last token minting event. For instance, minting events ranged from US \$3.11 for minting one token, to US \$288.52 for minting 80 tokens in the first batch (Figure 5C).

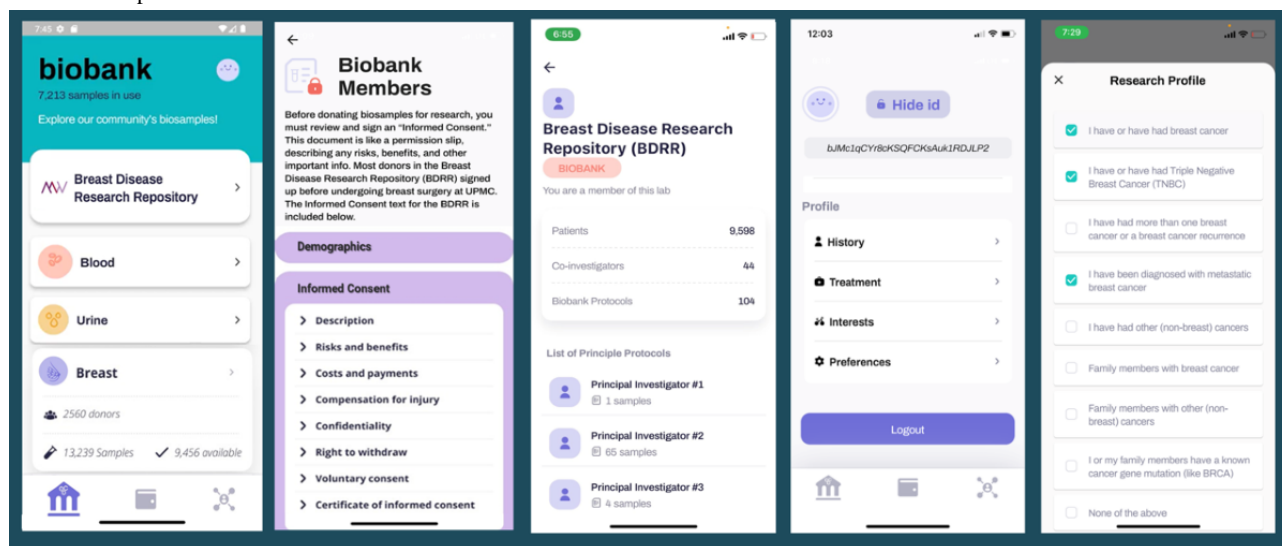
### Journey 2: Biobank Orientation and Research Profile

After requesting a *biowallet*, users were directed to visit the *biobank*, where they were oriented and learned about the overall biobank inventory and activities, including demographics of the consented donor population, framed as “biobank members”; informed consent content; principal investigators; and respective biobank operations and research activities for entire specimen collection (Figure 6). We included education about research protocol development, IRB oversight, procedures for specimen allocation, and investigator- and protocol-level transactions. The biobank displayed 60,973 biospecimens from 3940 unique donors collected from February 1995 to May 2023 and updated on a regular basis, with 318 new specimens added during the pilot. The feature tracked collection and distribution totals for the biobank, with breakdowns for each specimen type (Table 3).

The “profile” allowed participants to enter clinical history and treatments relevant for research on their specimens. We also assessed research interests, privacy preferences, engagement interest, and willingness to donate additional specimens to scientists as needed. In total, 37.8% (153/405) of the app users completed one or more portions of the profile, including 37.1% (134/361) of the biobank members. The profile also displayed the random “Private ID” number, which enabled users to remain deidentified while linking to their respective specimens. During the pilot, we experimented with the naming conventions, location, and order of presentation of biobank and profile features to assess impact on participants’ understanding of the biobank environment, affordances, constraints, and opportunities presented by the decentralized biobanking platform.

Nearly all participants who filled out the research profile (151/153, 98.7%) added one or more clinical details (eg, familial history of breast cancer; Multimedia Appendix 6). Profiles were completed by 39.9% (59/148) of the participants with samples, collectively annotating 886 specimens, including 760 (85.8%) available for future use, 36 (4.1%) “on hold” for a designated protocol, and 90 (10.2%) that were distributed for research, with information that was not contained within the institutional biobank database. In addition, participants added preferences regarding specimen use, willingness to provide further data and specimen donations, and future research engagement.

**Figure 6.** Biobank orientation journey, illustrating the biobank screen and user workflow introducing app users to biobank processes, what it means to be a biobank member, and regularly updated snapshots of investigator activities, protocols, and specimen allocations, at the level of the overall bank. The biobank also linked to participant's personal research profile, where they could provide key clinical details, interests, and preferences related to research on their specimens.





**Table 3.** Decentralized biobanking pilot population, app user and token claiming overview.

Pilot population	App users, n (%) <sup>a</sup>	Token claimed, n (%) <sup>b</sup>
<b>Total (N=1080)</b>	405 (37.5)	130 (12.04)
Biobank members (n=930) <sup>c</sup>	361 (38.82)	128 (13.76)
Biobank members with specimens (n=272) <sup>d</sup>	148 (54.41)	125 (46)
Collected specimens (n=3904) <sup>d</sup>	2133 (54.64)	1812 (46.41)
<b>Biobank members with specimens in use (n=165)<sup>d,e</sup></b>	88 (53.33)	74 (44.85)
Fresh (n=90)	46 (51.11)	40 (44.44)
Frozen (n=100)	50 (50)	40 (40)
<b>Specimens in use (n=377)<sup>d,e</sup></b>	202 (53.58)	177 (46.95)
Fresh (n=195)	104 (53.33)	95 (48.72)
Frozen (n=182)	98 (53.85)	82 (45.05)
<b>Number of donors with specimens available (n=242)<sup>d</sup></b>	132 (54.55)	110 (45.45)
Breast (n=147)	81 (55.1)	67 (45.58)
Blood (n=185)	97 (52.43)	82 (44.32)
Urine (n=166)	91 (54.82)	80 (48.19)
<b>Specimens available (n=3309)<sup>d</sup></b>	1757 (53.1)	1522 (46)
Breast (n=345)	205 (59.42)	178 (51.59)
Blood (n=2172)	1145 (52.72)	988 (45.55)
Urine (n=783)	406 (51.85)	355 (45.34)

<sup>a</sup>Specimen values and donor counts for all app engaged participants with specimens collected.

<sup>b</sup>Specimen values and donor counts for all app engaged participants with specimens collected who claimed biowallet tokens during the pilot study.

<sup>c</sup>Donor counts for all biobank consented pilot participants.

<sup>d</sup>Specimen values and donor counts for all biobank consented pilot participants with one or more specimens collected.

<sup>e</sup>Specimens considered in use if distributed to a research protocol as of May 4, 2023. A total of 218 specimens among all pilot participants with collected specimens designated “on hold” for future research use are not shown.

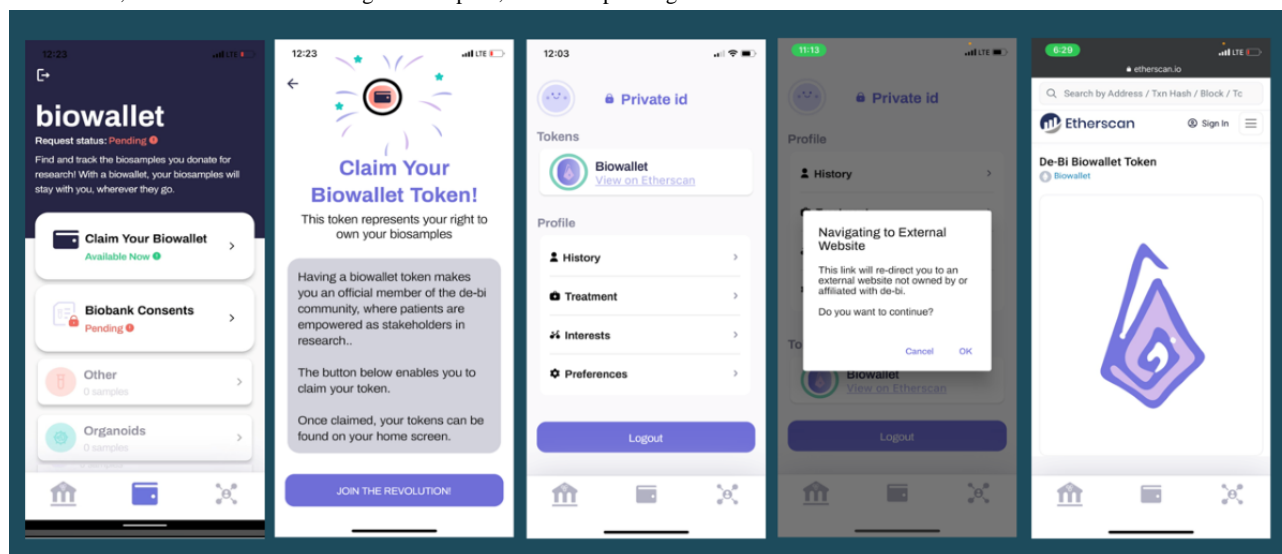
### ***Journey 3: Claiming and Viewing the Biowallet NFT***

#### **Overview**

Linking app accounts to biospecimen data occurred offline and took up to 2 weeks supported by software scripts and manual processes, including checks for false mismatches (eg, due to typos). Once biowallet NFTs were available, email notifications prompted participants to log in to their decentralized biobanking app to open their biowallet and access their personalized biospecimen data.

Once claimed, the “Biowallet token” appeared on the bottom of the screen with a link to view the corresponding Ethereum transaction data (Figure 7). The profile screen showed how patients could add clinical details that are not in the biobank database, making their biospecimens more readily discoverable by prospective users, reducing reliance on third-party chart review during study planning. The biowallet NFT signified membership in a collective committed to breast cancer research. Once claimed, the individual’s unique biowallet NFT could be viewed via an in-app Etherscan display. The app user experience represented this process as a symbolic “token of appreciation” as a form of reciprocity for biobank contributions.

**Figure 7.** Claiming and viewing the biowallet nonfungible token (NFT). The figure illustrates the biowallet NFT claiming process, first showing the appearance of the biowallet when the token is available to be claimed. Next, the claiming process is shown, which invites donors to “join the revolution!” Once claimed, the user’s personal NFT is represented on the profile page, which is connected via a hyperlink and an in-app display of the Etherscan view of the NFT, a customized biowallet logo for the pilot, and corresponding blockchain transaction data.



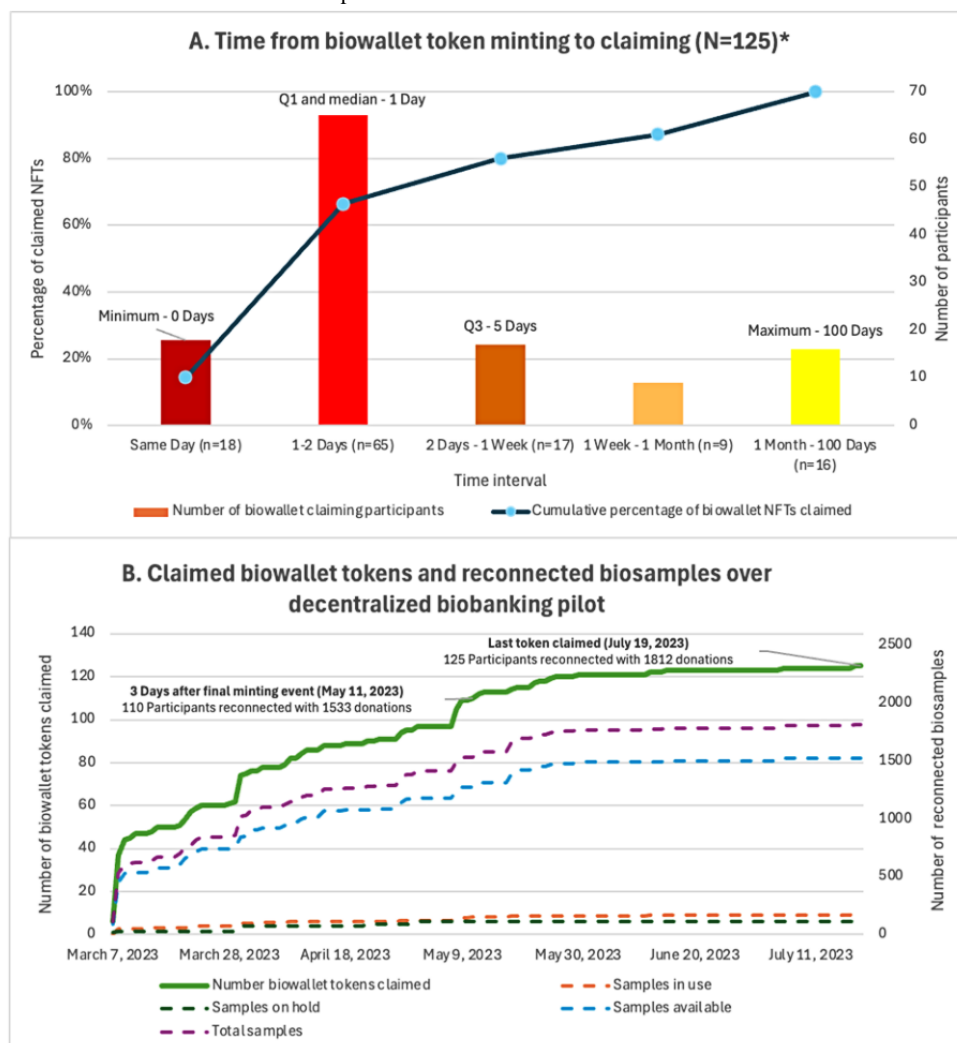
### Proof of Concept for Blockchain-Backed Biobanking App

The initial round of minting included “tokens of appreciation” for participants who were active in the demonstration phase of the app design and usability testing. The blockchain mechanism was initially tested with 4 test mints followed by minting “tokens of appreciation” for 7 demonstration phase participants. In total, 71% (5/7) of the demonstration users successfully completed the token minting claiming process, illustrating the use of the “biowallet” NFT as a representation of the individual’s membership in the biobank community. After validating functional integration of the blockchain simulation, eligibility for biowallet tokens was limited to those with confirmed

specimens in the breast cancer biobank, enabling us to simulate use of the NFTs to establish token-gated access to deidentified specimen accounts.

Of 148 app users with specimens, 140 (94.6%) initiated the biowallet token minting process during the pilot. Of 140 tokens minted, 125 (89.3%) were claimed by users, with an average of 10 (median 1, IQR 1-5, range 0-100) days between token minting and token claiming (Figure 8). Compared to individuals who did not claim their biowallet, those who did claim their biowallet were slightly younger (average of 58.9, SD 10.8 vs 61.9, SD 14.3 years) and had a similar time since biobank consent (7.8, SD 5.0 years since consent for claimants vs 7.7, SD 5.3 years for nonclaimants; Multimedia Appendix 7).

**Figure 8.** Nonfungible token claiming details for the decentralized biobanking breast cancer biobank pilot. Participant engagement and timing illustrates (A) interest in biospecimen tracking and receptiveness to email notification to facilitate the token claiming process and (B) the effective reconnecting specimens to participants that occurred during the pilot as tokens were claimed. In total, 89.3% (125/140) of tokens minted for app users with specimens were claimed during the pilot. Tokens were considered unclaimed after ~2.5 months following the final token minting event. A total of 15 participants had not yet claimed their token as of the conclusion of the pilot.



Ethnography of the US cancer specimen supply chain, including engagement with industry and academic stakeholders, generated the following conservative estimates for the commercial value of cancer tissue, blood, and urine specimens with well-annotated clinical data: US \$1000 for cancer tissue, US \$500 for blood, and US \$300 for urine. Hypothetically, this equates to US \$1 million of “available” specimens being populated into app users’ biowallets during the pilot. Similarly, this corresponds with a total value of approximately US \$30 million for unused specimens in frozen storage, with roughly US \$7000 in value per specimen contributor. Additional details of the scalability and economic feasibility of the proposed blockchain solution will be addressed elsewhere.

#### Journey 4: Viewing Personal Specimen Details

The “biowallet” was where participants could view details about when they consented for biobank donation (Figure 9). Once linkage between the user’s app and respective biobank data was established, individuals were able to track and learn about their own biospecimens. Details available via an interactive accordion feature included their biosample collection date, sample type and medium, if and when each sample was shared for a

particular research protocol, and similar sample-level information within the institutional database. The biowallet also includes a taxonomy of physical and digital biospecimen data types that may, in the future, be trackable by individual participants.

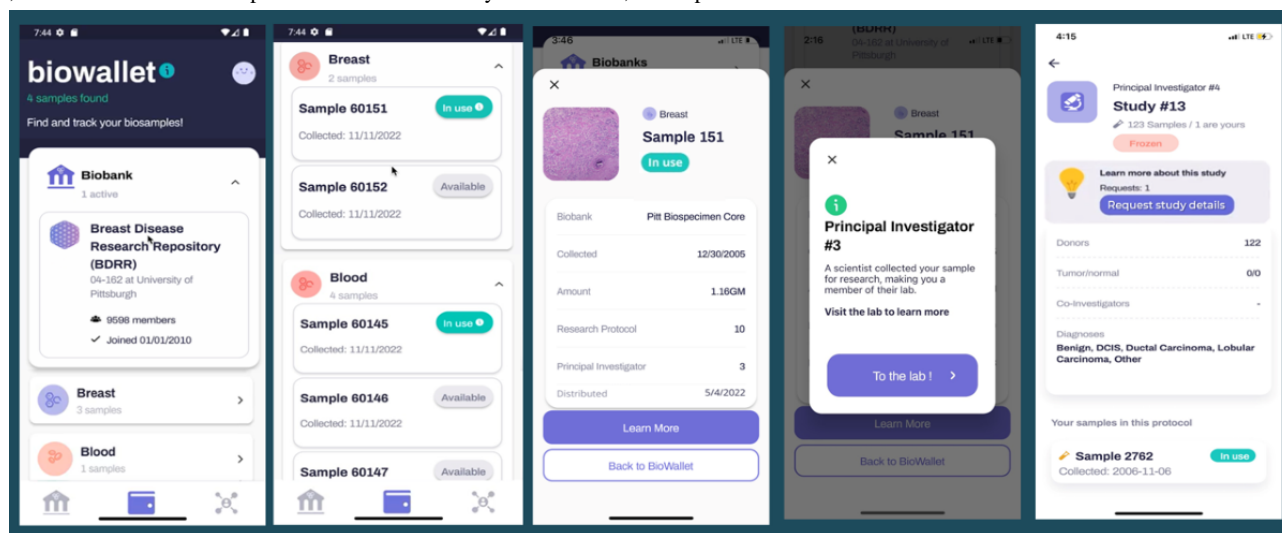
Further details regarding specimen distribution and availability were indicated via additional pop-ups, providing users with an opportunity to navigate to an app-based laboratory. Here app users could learn how many donors had contributed specimens of similar types, or had specimens distributed to the same research protocol. Of the biobank members using the app, 41% (148/361) had their “biowallet” populated with a total of 2113 specimens (mean 14.4, SD 12.1; range 1-84), including 1414 (66.9%) blood specimens, 419 (19.8%) urine specimens, and 296 (14%) breast tissues. In total, 70.9% (105/148) of sample holders had one or more breast tissue specimens. A total of 59.5% (88/148) had one or more specimens “in use” (mean 2.3, SD 1.6 per person; range 1-8), 40.5% (60/148) of the participants with specimens had none “in use,” and 4.7% (7/148) of the participants had specimens “on hold” (mean 24.9, SD 16.3; range 10-61). Individuals who had no specimens available

received a digital biobank membership card (Figure 9, panel 2) and in-app text notifying the participant that no specimens had been located (yet), with a range of possible explanations.

Collectively, 202 of app users' specimens were "in use," including 104 (51.5%) that were delivered "fresh" the day of donation (eg, for organoid development) and 98 (4%) from a frozen collection. A total of 8.2% (174/2113) were "on hold"

for a designated study, and 83.15% (1757/2113) were "available." App users' specimens were distributed to 22 different investigators under 42 research protocols. Between February 15, 2023, and May 4, 2023, users donated 39 new specimens, which appeared on the app, 2 (5%) of which were distributed fresh. In addition, 18% (7/39) were distributed from frozen storage, and 54% (21/39) were placed "on hold" during the pilot.

**Figure 9.** Biowallet sample tracking journey. This figure illustrates the participant experience learning about their personal specimen donations via an interactive biowallet landing page. Pop-up and accordion features enabled participants to learn about their specimens, including the type, collection date, distribution to a research protocol versus availability for future use, and explore further details about similar donations and distributions.



### Participant Feedback During the Pilot

During the pilot, cognitive walk-throughs with participants illuminated areas of interest along with potential opportunities for design improvement. Key areas of excitement included seeing how their samples were used. One participant stated the following:

*I will [otherwise] never know anything about my cells.*

Areas for improvement included improving technological accessibility (eg, making it iPad compatible) and clarifying the information presented (eg, "Will there be a way to learn more about each study?"). Table 4 provides a detailed thematic overview and representative quotes.

**Table 4.** A thematic overview of participant feedback gathered through cognitive walk-throughs conducted during the pilot.

Theme	Pilot participant feedback
<b>Things they liked</b>	
Big picture	<ul style="list-style-type: none"><li>• “This is cool on so many levels.”</li><li>• “Incredible concept to learn about.”</li><li>• “There are endless possibilities and uses for this.”</li><li>• “There is hope for others by giving my cells.”</li></ul>
Personalized feedback	<ul style="list-style-type: none"><li>• “I can’t wait to see what’s being done with my samples!”</li><li>• “Loved the idea of having access to my tissue info+how the two cancers are connected.”</li><li>• “I will [otherwise] never know anything about my cells.”</li><li>• “I’ll get to see the process.”</li></ul>
Empowerment	<ul style="list-style-type: none"><li>• “Information I could never access before.”</li><li>• “Give patients more control and information.”</li><li>• “Profile preferences—great idea.”</li><li>• “Private ID+Ability to connect w/ others in similar diagnosis.”</li></ul>
User interfaces and user experience	<ul style="list-style-type: none"><li>• “Menus under biowallet are clear+concise.”</li><li>• “Look of the app.”</li><li>• “Easy to navigate.”</li><li>• “Easy to use/menus good.”</li><li>• “Love the status of ‘in use’ and ‘available.’”</li></ul>
<b>Things they did not like or that did not meet their expectations</b>	
Information provided	<ul style="list-style-type: none"><li>• “Where are investigators that have my tissue or samples.”</li><li>• “Unclear when no samples (needs explanation).”</li><li>• “Will there be a way to learn more about each study?”</li><li>• “I need a little more background before fooling around with the app.”</li></ul>
Accessibility	<ul style="list-style-type: none"><li>• “Needed tutorial.”</li><li>• “Are there options for people who do not have email on their phone.”</li><li>• “Under personal history, other than TNBC (triple negative breast cancer) other breast cancers should be identified.”</li><li>• “Need to be able to use on an iPad for larger screen.”</li><li>• “Possible to put app on android tablet?”</li><li>• “Being older I’m not a techie and it takes a while.”</li></ul>
Functionality and user navigation	<ul style="list-style-type: none"><li>• “Biowallet should be first icon.”</li><li>• “Make biobank/wallet first tab.”</li><li>• “Add search bar in connect.”</li><li>• “Some functions are more intuitive than others—more prompts are needed.”</li><li>• “What was the purpose behind ‘home’ icon community samples.”</li></ul>

## Discussion

### Principal Findings

The decentralized biobanking pilot demonstrated the technical feasibility of design, development, and implementation of a user-friendly app to deliver transparency and engagement for donors to a well-established biospecimen collection protocol at a US academic medical center. Over 400 participants downloaded and tested the decentralized biobanking app during the pilot, asserting interest in tracking their biospecimens, demonstrating the usability of a patient interface for institutional biobanking data. “Biowallet” tokens (ERC-721) were minted for app users with confirmed specimens, and 89.3% (125/140) successfully claimed their NFTs on the app, with over half (72/125, 57.6%) of the population achieving the task within 1 day of token minting.

Pilot participants’ biowallet token claiming process symbolically asserted their right to know what happens to their inherently unique biospecimens, to which they are immutably linked via a nontransferable, one-of-a-kind relationship. The user experience simulated an NFT-gated process, functionally reconnecting app users to >1800 deidentified specimens, providing visibility of affiliated community members and related research activities all while preserving confidentiality. Critically, this was achievable with data architecture, interfaces, and workflows that maintained compliance with preexisting deidentification standards and specimen collection and distribution protocols.

Similarly, we showed how integration with institutional biobank infrastructure can passively provide transparency for donors without imposing undue burdens on investigators or relying on individual research programs to sustain community engagement. Transparency in biobanking has the potential to rebuild donor



trust in biobanks and improve accountability in biomedical research [27–29]. Consequently, transparency may be a driver to improve biobank donations, particularly among communities with historically rooted distrust of biomedical research [30,31]. The decentralized biobanking framework also allowed for the retrospective and prospective onboarding of donors, demonstrating the potential to convert existing biobanks to a progressively decentralized, patient-centered model.

Minting biowallet NFTs averaged US \$4.51 (SD US \$2.54; range US \$1.84–\$11.23) per token, with a projected total cost of US \$17,769.40 (SD US \$159.52) for all biobank members with specimens. Importantly, a 1-time minting expense of <US \$5 per patient may be considered marginal, especially in view of the cost of specimen procurement, storage, and distribution. A workshop on biospecimen economics found the cost of operating a large biobank to be US \$861 per patient [32]. The value of the specimens themselves is also substantial relative to minting expenses; academic researchers may pay up to US \$200 per sample, whereas commercial entities may pay up to US \$20,000 per sample [32]. When biospecimens are converted into living models (eg, organoids), the expenses of both processing and development increase, but the value is multiplied several-fold as 1-mL aliquots of the model may cost upward of several thousand dollars per copy for academic and commercial users alike [33,34].

Importantly, we also demonstrated how empowering patients may in turn help scientists by allowing them to annotate their biospecimens with relevant data that may not be represented in the institutional biobank database or may be otherwise not directly available to prospective or current specimen users. Over 37% (150/405) of the participants demonstrated how longitudinal donor involvement might be leveraged to improve biosample curation and discoverability, creating opportunities to enrich research; link siloed datasets; and drive more efficient, community-driven use of biobank resources. Enhanced annotation of biospecimens with clinical data reflects increasing demand among the biobanking community to gain more contextual biospecimen data [35]. Project LUNGBANK is an example of ongoing efforts to provide more comprehensive clinical data to enrich biospecimens [36]. In LUNGBANK, clinically relevant findings collected through manual chart review of patient medical records were used to annotate biospecimens [36]. For the decentralized biobanking app, more intuitive, strategic placement of the profile feature and improved framing of its functionality and benefits for donors and scientists will be essential to optimize the utility of this feature.

Although relatively limited in functionality compared to the NFT framework advanced in our preclinical prototypes, the blockchain aspect of the piloted app was significant for several reasons. First, it represents the first time that most of our participants, including several octogenarians, had ever interacted with blockchain technologies. Second, persistence in overcoming the friction of onboarding related to the blockchain elements served as further evidence of the high value that patients place on tracking their specimens, to the point that they were willing to participate in a cumbersome, multistage process that, in some cases, took weeks. Third, the blockchain aspect of the piloted app remains a permanent, institution-agnostic

record of the relationship between specific donors and their respective biospecimens, highlighting the potential to reunite individuals with these deeply personal assets, with yet unmet potential for assurances of trust and shared rewards of research. Finally, the biowallet NFT represents a foundational gateway to a composable and progressively decentralized biobanking ecosystem. That which starts with 1 biowallet token per participant who contributes specimens may be built upon in a stepwise manner, forging an interconnected overlay network that recognizes and unlocks value across today's siloed biobank landscape.

## Limitations

The pilot relied on manual data workflows to enable demonstration of a functional decentralized biobanking platform without requiring full integration of the patient-facing apps with the enterprise system. Such manual workflows are impractical for sustainability and scalability. The exponential growth of health information and advanced computing makes workflow automation increasingly fundamental [37]. Thus, application programming interface (API) integration and automated processes will be necessary for future apps. In view of the volume of requests received during the pilot as well as interest in expanding the program to other institutional biobanks, hospital leadership approved API development to facilitate such integrations for the next stages of the pilot program. In addition to being essential for technical feasibility, this approval was critical as it demonstrated that the manual aspects of our workflows were not material for the acceptability of our strategy for reconnecting donors with their deidentified specimens within institutional biobanks.

Notifications based on in-app activity event triggers were not fully implemented during the pilot, and a number of manual steps were required, including substantial coordination across study team members and email-based messaging to notify participants about critical changes such as token availability and biosample status updates. Automated communications must be incorporated into future pilots with accommodation for a range of patient preferences and values. Subsequent development will also make a web-based version to avoid exclusion of participants for whom smartphone apps may not be preferred or accessible, particularly with respect to age and household income [38].

Furthermore, the piloted app interfaces and user journeys were designed for patient users, whereas engagement with physicians, biobankers, and scientists occurred via alternative channels (eg, email and institutional platforms). This limited the functionality and value within the app as research content was high level, limited to the scope of the biobank database. Ongoing work is advancing real-world applications of decentralized biobanking for scientists and other stakeholders within the NFT digital twin ecosystem. Inclusion of professional users directly within the decentralized biobanking platform will be key for unlocking the ongoing value and network effects of our framework.

Regarding the blockchain elements, the high and highly variable costs of token mints on Ethereum illustrate the importance of more cost-efficient strategies, such as layer-2 solutions, for full-scale implementation. Importantly, our focus on the primary



NFT digital twin framework centers the stakeholders and their relational mappings within the ecosystem. This allowed us to focus on tokenizing the individual participants, in this case, 1 token per biospecimen donor rather than 1 per biospecimen, which would have increased costs 10- to 20-fold. This was sensible, especially considering limitations on functionality of a specimen-representing NFT in the setting of our pilot app; that is, it was not necessary to tokenize specimens for implementing transparency and our study did not provide additional permissions relevant to potential tokenized specimen utility for shared governance or profit sharing regarding the underlying biobank assets. Moreover, ensuring the long-term economic sustainability of biobanks is already a salient concern, with high costs driven by human resources, equipment, and sample handling [39-41]. Cost-effectiveness will be essential for broader adoption of decentralized biobanking technology, and blockchain solutions in themselves must be complemented with social, cultural, and legal innovations to enact meaningful progress [40,42,43].

In addition, NFTs were minted for individual participants, and personal NFTs were rendered via an in-app Etherscan display, although the token-gated aspect of the app leveraged Firebase Unique Identifiers rather than NFTs to minimize complexity and potential points of failure. Simulation of the user interface and user experience of blockchain interactions was necessary to overcome barriers to onboarding inherent to contemporary avoidances and constraints of decentralized apps, particularly as our patient population was older and almost exclusively from non-digital native generations and many were actively grappling with cancer. This was especially critical given concurrent educational barriers surrounding the simultaneous introduction of patients to both biobanking and blockchain for the first time. For example, a knowledge assessment on biobanking administered to biospecimen donors found that approximately half of all questions were answered either incorrectly or with “I don’t know.” Similarly, most patients we engaged with during app design, development, and pilot-testing were not familiar with the term “biobank,” illustrating the fundamental challenge of delivering a patient-friendly biobanking app. These findings underscore the gap between providing information during the prospective informed consent process and achieving true comprehension via enduring transparency and ongoing feedback [44,45]. To this end, we prioritized orientation to biobanking and developed lexicon and app design features that make data within biobank databases accessible to donors via a decentralized biobanking platform that coheres with the ethos of decentralization at its core.

For future implementations, we aim to advance blockchain-backed solutions with seamless onboarding experiences through the exploration of newer standards such as ERC-4337 for account abstraction, which awards the programmable flexibility to remove complex barriers to entry such as the current requirement for users to create their own third-party wallets to interact with the decentralized app. Advancement of these technologies may provide seamless integration of decentralized biobanking platforms with both institutional databases and blockchain overlay networks, with future potential to unite participants, specimens, and scientists

across various institutions. Transparency and engagement in biospecimen management is a necessary step toward institutional transformation to achieve community partnership, shared decisions, and progressive democratization. More research is needed to test our hypotheses about the role of blockchain technology in a comprehensive and universal decentralized biobanking solution [46].

The success of our pilot inspired potential to revolutionize biobanking via a decentralized platform but also revealed challenges and limitations for current biospecimen collection workflows, standard operating procedures, and data management strategies [47]. Implementation of transparency for past, present, and future biospecimen collection and distribution will require innovative system designs that overcome idiosyncrasies of individual biobank databases coupled with incentive structures and governance models that promote trust and ensure that biobanking practice optimizes individual and collective interests for patients, scientists, and society [48-50]. While the principles and techniques demonstrated in this study theoretically translate to any other research biobanking context, our technical approach must be validated across a variety of clinical and socioeconomic settings, institutional and regional cultures, and biomedical research contexts.

Critically, this pilot addressed a single, disease-focused university biobank with a largely White, female, and geographically localized population. Technology acceptance must be confirmed for diverse patients, diseases, and contexts [51]. Both iOS and Android users were included, yet some did not use smartphones, and others preferred not to download apps. We have since developed a web-based platform, expanding availability to anyone with internet access, although disparities persist. Ongoing research is exploring the impact of age, race, time elapsed since surgery, and stage of disease on technology acceptability, as well as how to optimize recruitment and trustworthiness for underserved populations [51,52]. Current work is also addressing populations such as those with prostate and lung cancer in which male individuals are more heavily represented, and we have incorporated socioeconomic assessments into our data collection to ensure that we advance solutions that are broadly accessible and applicable, especially for economically and educationally marginalized groups.

Looking ahead beyond feasibility, the practical implementation of scalable, decentralized biobanking solutions requires technical enhancements to overcome the discussed challenges and limitations of this pilot. User interfaces must prioritize usability, comprehensibility, and accessibility by leveraging new standards for account abstraction to reduce the complexity of interacting with blockchain components in our solution. Similarly, ongoing research should inform iterative refinement of different strategies for effective presentation of research-related information curated for diverse patient populations. Efforts toward long-term sustainability should include app cost optimization techniques such as deployment on layer-2 networks for major reductions in blockchain transaction costs and the automation of key workflows and processes through proper integration with institutional software and databases. Because each new environment can be quite nuanced, the application of our technology to new use cases will still require custom

configurations when onboarding, but some of these efforts may be streamlined by standardizing integration patterns with widely used laboratory information management systems and research tools.

Finally, our privacy-by-design approach requires due diligence in execution to mitigate risks to users. Abiding by security best practices in development and thorough vulnerability testing are essential measures in protecting against critical security risks. Intentional disaster recovery plans with detailed incident response protocols for specific events are important for prompt threat containment, recovery of system resources with minimal downtime, and communication to affected users and stakeholders. Proactive preparation to set up comprehensive monitoring, automated backups with manual snapshots across system resources and environments, and pre-emptively programmed functionality for pausing and redeploying compromised system components or deployed smart contracts are crucial for the effective execution of incident response plans.

## Conclusions

This pilot demonstrates the technical capacity and resources for a functional decentralized biobanking software app that empowers patients to track specimens donated to a real-world breast cancer research biobank with a novel implementation of blockchain technology. The patient-friendly mobile app renders institutional biobank inventory and transactions in a meaningful, personalized biowallet context, providing a rewarding user experience. We demonstrated the app's readiness for API integrations, which would allow for sustainable and scalable implementation across multiple biobank protocols by seamlessly and dynamically displaying biobanking activities to donors. Pilot participants successfully claimed NFTs within the app, restoring provenance for personal biospecimens and related data. This advancement introduces a new paradigm for ethical biobanking, fostering donor engagement and inclusion in personalized research networks appropriate to contemporary learning health systems and mobile computing capabilities while maintaining deidentification and compliance with established protocols.

## Acknowledgments

The authors would like to thank the pilot participants, physicians, scientists, institutional review board members, and biobankers in the pilot setting who made this work possible. They are especially grateful to Drs Balaji Palanisamy, Dimitriy Babichenko, Adam Lee, Mylynda Massart, Adrian Lee, Adam Brufsky, Peter Allen, Eric Dueweke, Rajiv Dhir, and Suzanne Gollin, each of whom contributed significantly to the technical and operational design, development, deployment, approval, and oversight of the pilot described herein. Foundational research on decentralized biobanking is generously supported by a grant from Emerson Collective and Yosemite to Johns Hopkins Berman Institute of Bioethics. The pilot feasibility study described in this paper was enabled by grants from the University of Pittsburgh Medical Center Beckwith Institute, which supported app production and integration with the institutional biobank, and the Pitt Chancellor's Gap Fund, which supported blockchain designs and technical development. Additional labor, materials, and resources needed to execute this study were provided by the Pitt Biospecimen Core (RRID: SCR\_025229) as supported in part by the Office of the Senior Vice Chancellor for the Health Sciences of the University of Pittsburgh, the University of Pittsburgh and University of Pittsburgh Medical Center-affiliated Institute for Precision Medicine, Magee-Womens Research Institute, and David Berg Center for Ethics and Leadership at the Katz Graduate School of Business. This manuscript reflects the independent research of the authors, whose scholarship, reputations, and commercial activities reflect consistent, recognized commitments to advancing the state of the art for ethical biobanking.

## Data Availability

The datasets generated or analyzed during this study are available from the corresponding author upon reasonable request.

## Conflicts of Interest

WS, RCM, JK, MM, and MG are shareholders in de-bi, co, a company created to advance decentralized biobanking technology to empower patients, accelerate science, and realize precision medicine, and ME is a consultant thereof. The pilot study described in this paper was not sponsored by de-bi, co; however, detailed conflict of interest disclosures were incorporated into all informed consent forms and quality improvement procedures, and verbal disclosures were made throughout pilot engagement. Conflicts of interest were disclosed in accordance with the pilot protocol under Conflict Management Plans, and the principal investigator on the decentralized biobanking app pilot protocol was nonconflicted.

## Multimedia Appendix 1

Architecture details of the deployed cloud infrastructure, encompassing networking setup, front-end and back-end deployment configurations, database architecture, continuous integration and continuous delivery processes, domain management, and monitoring systems.

[[DOCX File, 17 KB](#) - [bioinform\\_v6ile70463\\_app1.docx](#) ]

## Multimedia Appendix 2

Workflow for the de-bi pilot, including consent for the pilot, downloading the app, and minting biowallet tokens to link personal biospecimen data to their biowallet.

[DOCX File, 179 KB - [bioinform\\_v6i1e70463\\_app2.docx](#)]

#### Multimedia Appendix 3

Decentralized biobanking pilot study participation rates among eligible biobank members by age, race, and time from initial biobank consent.

[DOCX File, 31 KB - [bioinform\\_v6i1e70463\\_app3.docx](#)]

#### Multimedia Appendix 4

Decentralized biobanking pilot study and breast cancer biobank age distributions, and comparison of pilot enrollment rates by time from initial biobank consent.

[DOCX File, 136 KB - [bioinform\\_v6i1e70463\\_app4.docx](#)]

#### Multimedia Appendix 5

Demographics of the breast cancer biobank and decentralized biobanking pilot populations.

[DOCX File, 35 KB - [bioinform\\_v6i1e70463\\_app5.docx](#)]

#### Multimedia Appendix 6

Decentralized biobanking pilot participant age ranges and engagement metrics among app onboarded participants by sample collection status, including age, biobank membership, years since initial biobank consent, and research profile completion rates.

[DOCX File, 2726 KB - [bioinform\\_v6i1e70463\\_app6.docx](#)]

#### Multimedia Appendix 7

Characteristics of pilot participants who did versus did not complete research profiles and claim biowallet tokens on decentralized biobanking app.

[DOCX File, 15 KB - [bioinform\\_v6i1e70463\\_app7.docx](#)]

## References

1. Kongsholm NC, Christensen ST, Hermann JR, Larsen LA, Minssen T, Pedersen LB, et al. Challenges for the sustainability of university-run biobanks. *Biopreserv Biobank* 2018 Aug;16(4):312-321. [doi: [10.1089/bio.2018.0054](#)] [Medline: [30016130](#)]
2. Kinkorová J. Biobanks in the era of personalized medicine: objectives, challenges, and innovation: overview. *EPMA J* 2015;7(1):4 [FREE Full text] [doi: [10.1186/s13167-016-0053-7](#)] [Medline: [26904153](#)]
3. Rush A, Matzke L, Cooper S, Gedye C, Byrne JA, Watson PH. Research perspective on utilizing and valuing tumor biobanks. *Biopreserv Biobank* 2019 Jun;17(3):219-229. [doi: [10.1089/bio.2018.0099](#)] [Medline: [30575428](#)]
4. Klingstrom T, Bongcam-Rudloff E, Reichel J. Legal and ethical compliance when sharing biospecimen. *Brief Funct Genomics* 2018 Jan 01;17(1):1-7 [FREE Full text] [doi: [10.1093/bfpg/elx008](#)] [Medline: [28460118](#)]
5. Hallinan D, Friedewald M. Open consent, biobanking and data protection law: can open consent be 'informed' under the forthcoming data protection regulation? *Life Sci Soc Policy* 2015 Jan 24;11(1):1 [FREE Full text] [doi: [10.1186/s40504-014-0020-9](#)] [Medline: [26085311](#)]
6. Sobel ME, Dreyfus JC, Dillehay McKillip K, Kolarcik C, Muller WA, Scott MJ, et al. Return of individual research results: a guide for biomedical researchers utilizing human biospecimens. *Am J Pathol* 2020 May;190(5):918-933 [FREE Full text] [doi: [10.1016/j.ajpath.2020.01.014](#)] [Medline: [32201265](#)]
7. Elger BS, De Clercq E. Returning results: let's be honest!. *Genet Test Mol Biomarkers* 2017 Mar;21(3):134-139 [FREE Full text] [doi: [10.1089/gtmb.2016.0395](#)] [Medline: [28306398](#)]
8. Wolf SM. Return of results in genomic biobank research: ethics matters. *Genet Med* 2013 Feb;15(2):157-159 [FREE Full text] [doi: [10.1038/gim.2012.162](#)] [Medline: [23386184](#)]
9. Scudellari M. Biobank managers bemoan underuse of collected samples. *Nat Med* 2013 Mar;19(3):253. [doi: [10.1038/nm0313-253a](#)] [Medline: [23467224](#)]
10. AminoChain homepage. AminoChain. URL: <https://aminochain.io/> [accessed 2025-03-27]
11. iSpecimen. URL: <https://www.ispecimen.com/> [accessed 2025-03-27]
12. Hasselgren A, Hanssen Rensaa JA, Kralevska K, Gligoroski D, Faxvaag A. Blockchain for increased trust in virtual health care: proof-of-concept study. *J Med Internet Res* 2021 Jul 30;23(7):e28496 [FREE Full text] [doi: [10.2196/28496](#)] [Medline: [34328437](#)]
13. Velmovitsky PE, Bublitz FM, Fadrique LX, Morita PP. Blockchain applications in health care and public health: increased transparency. *JMIR Med Inform* 2021 Jun 08;9(6):e20713 [FREE Full text] [doi: [10.2196/20713](#)] [Medline: [34100768](#)]

14. Bayyapu S. Blockchain healthcare: redefining data ownership and trust in the medical ecosystem. *Int J Adv Res Eng Technol* 2020 Nov;11(11):2748-2755 [[FREE Full text](#)]
15. Alshater MM, Nasrallah N, Khoury R, Joshapura M. Deciphering the world of NFTs: a scholarly review of trends, challenges, and opportunities. *Electron Commer Res* 2024 Jul 30. [doi: [10.1007/s10660-024-09881-y](#)]
16. Gross M, Hood AJ, Sanchez WL. Blockchain technology for ethical data practices: decentralized biobanking pilot study. *Am J Bioeth* 2023 Nov 25;23(11):60-63. [doi: [10.1080/15265161.2023.2256286](#)] [Medline: [37879029](#)]
17. Mamo N, Martin GM, Desira M, Ellul B, Ebejer JP. Dwarna: a blockchain solution for dynamic consent in biobanking. *Eur J Hum Genet* 2020 May;28(5):609-626 [[FREE Full text](#)] [doi: [10.1038/s41431-019-0560-9](#)] [Medline: [31844175](#)]
18. McGhin T, Choo KR, Liu CZ, He D. Blockchain in healthcare applications: research challenges and opportunities. *J Netw Comput Appl* 2019 Jun;135:62-75. [doi: [10.1016/j.jnca.2019.02.027](#)]
19. Attaran M. Blockchain technology in healthcare: challenges and opportunities. *Int J Healthc Manag* 2020 Nov 08;15(1):70-83. [doi: [10.1080/20479700.2020.1843887](#)]
20. Schär F. Decentralized finance: on blockchain- and smart contract-based financial markets. *Fed Reserve Bank St. Louis Rev* 2021 Apr 15:153-174 [[FREE Full text](#)] [doi: [10.20955/r.103.153-74](#)]
21. Chen Y, Bellavitis C. Blockchain disruption and decentralized finance: the rise of decentralized business models. *J Bus Ventur Insights* 2020 Jun;13:e00151. [doi: [10.1016/j.jbvi.2019.e00151](#)]
22. Sanchez W, Linder L, Miller RC, Hood A, Gross MS. Non-fungible tokens for organoids: decentralized biobanking to empower patients in biospecimen research. *Blockchain Healthc Today* 2024;7 [[FREE Full text](#)] [doi: [10.30953/bhty.v7.303](#)] [Medline: [38715762](#)]
23. Dewan A, Eifler M, Hood A, Sanchez W, Gross M. Building a decentralized biobanking app for research transparency and patient engagement: participatory design study. *JMIR Hum Factors* 2025 Mar 05;12:e59485 [[FREE Full text](#)] [doi: [10.2196/59485](#)] [Medline: [40053747](#)]
24. Human Cancer Models Initiative (HCMI). National Institutes of Health National Cancer Institute Center for Cancer Genomics. URL: <https://www.cancer.gov/ccg/research/functional-genomics/hcmi> [accessed 2025-03-27]
25. Sanchez W, Dewan A, Budd E, Eifler M, Miller RC, Kahn J, et al. Decentralized biobanking applications empower personalized tracking of biospecimen research: technology feasibility. *JMIR Bioinform Biotechnol* 2025 Apr 14:70463. [doi: [10.2196/70463](#)]
26. Singh P, Sagar S, Singh S, Alshahrani HM, Getahun M, Soufiene BO. Blockchain-enabled verification of medical records using soul-bound tokens and cloud computing. *Sci Rep* 2024 Oct 22;14(1):24830. [doi: [10.1038/s41598-024-75708-3](#)] [Medline: [39438519](#)]
27. Gille F, Axler R, Blasimme A. Transparency about governance contributes to biobanks' trustworthiness: call for action. *Biopreserv Biobank* 2021 Feb 01;19(1):83-85. [doi: [10.1089/bio.2020.0057](#)] [Medline: [33124891](#)]
28. Weil CJ, Nanyonga S, Hermes A, McCarthy A, Gross M, Nansumba H, et al. Experts speak forum: community engagement in research biobanking. *Biopreserv Biobank* 2024 Oct 01;22(5):535-539. [doi: [10.1089/bio.2024.0131](#)] [Medline: [39431940](#)]
29. Gross MS, Hood AJ, Miller RC. Nonfungible tokens as a blockchain solution to ethical challenges for the secondary use of biospecimens: viewpoint. *JMIR Bioinform Biotechnol* 2021 Oct 22;2(1):e29905 [[FREE Full text](#)] [doi: [10.2196/29905](#)] [Medline: [38943235](#)]
30. Statler M, Wall BM, Richardson JW, Jones RA, Kools S. African American perceptions of participating in health research despite historical mistrust. *ANS Adv Nurs Sci* 2023;46(1):41-58. [doi: [10.1097/ANS.0000000000000435](#)] [Medline: [35984948](#)]
31. Scharff DP, Mathews KJ, Jackson P, Hoffsuemmer J, Martin E, Edwards D. More than Tuskegee: understanding mistrust about research participation. *J Health Care Poor Underserved* 2010 Aug;21(3):879-897 [[FREE Full text](#)] [doi: [10.1353/hpu.0.0323](#)] [Medline: [20693733](#)]
32. Compton CC. Making economic sense of cancer biospecimen banks. *Clin Transl Sci* 2009 Jun 29;2(3):172-174 [[FREE Full text](#)] [doi: [10.1111/j.1752-8062.2008.00108.x](#)] [Medline: [20443887](#)]
33. HUB Organoids homepage. HUB Organoids. URL: <https://www.huborganoids.nl/> [accessed 2025-03-27]
34. American Type Culture Collection homepage. American Type Culture Collection. URL: <https://www.atcc.org/> [accessed 2025-03-27]
35. Reihs R, Proynova R, Maqsood S, Ataian M, Lablans M, Quinlan PR, et al. BBMRI-ERIC negotiator: implementing efficient access to biobanks. *Biopreserv Biobank* 2021 Oct 01;19(5):414-421. [doi: [10.1089/bio.2020.0144](#)] [Medline: [34182766](#)]
36. Ceker D, Baysungur V, Evman S, Kolbas I, Gordebil A, Nalbantoglu S, et al. LUNGBANK: a novel biorepository strategy tailored for comprehensive multi-omics analysis and P-medicine applications in lung cancer. *Research Square Preprint* posted online on January 24, 2024 [[FREE Full text](#)] [doi: [10.21203/rs.3.rs-3816689/v1](#)]
37. Zayas-Cabán T, Okubo TH, Posnack S. Priorities to accelerate workflow automation in health care. *J Am Med Inform Assoc* 2022 Dec 13;30(1):195-201 [[FREE Full text](#)] [doi: [10.1093/jamia/ocac197](#)] [Medline: [36259967](#)]
38. Sidoti O, Dawson W, Gelles-Watnick R, Favereio M, Atske S, Radde K, et al. Mobile fact sheet. Pew Research Center. 2024 Nov 13. URL: <https://www.pewresearch.org/internet/fact-sheet/mobile/> [accessed 2025-03-27]
39. Doucet M, Yuille M, Georgiou L, Dagher G. Biobank sustainability: current status and future prospects. *J Biorepository Sci Appl Med* 2017 Jan;Volume 5:1-7. [doi: [10.2147/bsam.s100899](#)]



40. Odeh H, Miranda L, Rao A, Vaught J, Greenman H, McLean J, et al. The biobank economic modeling tool (BEMT): online financial planning to facilitate biobank sustainability. *Biopreserv Biobank* 2015 Dec;13(6):421-429 [FREE Full text] [doi: [10.1089/bio.2015.0089](https://doi.org/10.1089/bio.2015.0089)] [Medline: [26697911](https://pubmed.ncbi.nlm.nih.gov/26697911/)]
41. Simeon-Dubach D, Henderson MK. Sustainability in biobanking. *Biopreserv Biobank* 2014 Oct;12(5):287-291. [doi: [10.1089/bio.2014.1251](https://doi.org/10.1089/bio.2014.1251)] [Medline: [25314050](https://pubmed.ncbi.nlm.nih.gov/25314050/)]
42. Racine V. Can blockchain solve the dilemma in the ethics of genomic biobanks? *Sci Eng Ethics* 2021 Jun 01;27(3):35. [doi: [10.1007/s11948-021-00311-y](https://doi.org/10.1007/s11948-021-00311-y)] [Medline: [34061257](https://pubmed.ncbi.nlm.nih.gov/34061257/)]
43. Sabharwal K, Hutler B, Eifler M, Gross M. Decentralized biobanking for transparency, accountability, and engagement in biospecimen donation. *J Health Care Law Policy* 2025 [FREE Full text]
44. Kasperbauer TJ, Schmidt KK, Thomas A, Perkins SM, Schwartz PH. Incorporating biobank consent into a healthcare setting: challenges for patient understanding. *AJOB Empir Bioeth* 2021 Dec 04;12(2):113-122 [FREE Full text] [doi: [10.1080/23294515.2020.1851313](https://doi.org/10.1080/23294515.2020.1851313)] [Medline: [33275086](https://pubmed.ncbi.nlm.nih.gov/33275086/)]
45. Dewan A, Eifler M, Hood A, Sanchez W, Gross M. Building a decentralized biobanking app for research transparency and patient engagement: participatory design study. *JMIR Hum Factors* 2025 Mar 05;12:e59485 [FREE Full text] [doi: [10.2196/59485](https://doi.org/10.2196/59485)] [Medline: [40053747](https://pubmed.ncbi.nlm.nih.gov/40053747/)]
46. El-Gazzar R, Stendal K. Blockchain in health care: hope or hype? *J Med Internet Res* 2020 Jul 10;22(7):e17199 [FREE Full text] [doi: [10.2196/17199](https://doi.org/10.2196/17199)] [Medline: [32673219](https://pubmed.ncbi.nlm.nih.gov/32673219/)]
47. Ellis H, Joshi MB, Lynn AJ, Walden A. Consensus-driven development of a terminology for biobanking, the Duke experience. *Biopreserv Biobank* 2017 Apr;15(2):126-133 [FREE Full text] [doi: [10.1089/bio.2016.0092](https://doi.org/10.1089/bio.2016.0092)] [Medline: [28338350](https://pubmed.ncbi.nlm.nih.gov/28338350/)]
48. Rogers J, Carolin T, Vaught J, Compton C. Biobankonomics: a taxonomy for evaluating the economic benefits of standardized centralized human biobanking for translational research. *J Natl Cancer Inst Monogr* 2011;2011(42):32-38. [doi: [10.1093/jncimonographs/lgr010](https://doi.org/10.1093/jncimonographs/lgr010)] [Medline: [21672893](https://pubmed.ncbi.nlm.nih.gov/21672893/)]
49. Catchpoole D. 'Biohoarding': treasures not seen, stories not told. *J Health Serv Res Policy* 2016 Apr 05;21(2):140-142. [doi: [10.1177/1355819615599014](https://doi.org/10.1177/1355819615599014)] [Medline: [26248620](https://pubmed.ncbi.nlm.nih.gov/26248620/)]
50. Dewan A, Rubin JC, Gross MS. Informed consensus: the future of respect for persons in biomedical research. *Am J Bioeth* 2025 (forthcoming). [doi: [10.1080/15265161.2025.2470695](https://doi.org/10.1080/15265161.2025.2470695)]
51. Hiatt RA, Kobetz EN, Paskett ED. Catchment areas, community outreach and engagement revisited: the 2021 guidelines for cancer center support grants from the National Cancer Institute. *Cancer Prev Res (Phila)* 2022 Jun 02;15(6):349-354. [doi: [10.1158/1940-6207.CAPR-22-0034](https://doi.org/10.1158/1940-6207.CAPR-22-0034)] [Medline: [35652232](https://pubmed.ncbi.nlm.nih.gov/35652232/)]
52. Wilkowska W, Ziefle M. Perception of privacy and security for acceptance of e-health technologies: exploratory analysis for diverse user groups. In: *Proceedings of the 5th International Conference on Pervasive Computing Technologies for Healthcare and Workshops*. 2011 Presented at: PervasiveHealth 2011; May 23-26, 2011; Dublin, Ireland. [doi: [10.4108/icst.pervasivehealth.2011.246027](https://doi.org/10.4108/icst.pervasivehealth.2011.246027)]

## Abbreviations

**API:** application programming interface  
**BDRR:** Breast Disease Research Repository  
**IRB:** institutional review board  
**NFT:** nonfungible token

*Edited by Z Yue; submitted 22.12.24; peer-reviewed by E Gillette, N Godwin, T David, T Church; comments to author 13.02.25; revised version received 27.02.25; accepted 04.03.25; published 10.04.25.*

### *Please cite as:*

Sanchez W, Dewan A, Budd E, Eifler M, Miller RC, Kahn J, Macis M, Gross M  
*Decentralized Biobanking Apps for Patient Tracking of Biospecimen Research: Real-World Usability and Feasibility Study*  
*JMIR Bioinform Biotech* 2025;6:e70463  
URL: <https://bioinform.jmir.org/2025/1/e70463>  
doi:[10.2196/70463](https://doi.org/10.2196/70463)  
PMID:[40208659](https://pubmed.ncbi.nlm.nih.gov/40208659/)

©William Sanchez, Ananya Dewan, Eve Budd, M Eifler, Robert C Miller, Jeffery Kahn, Mario Macis, Marielle Gross. Originally published in *JMIR Bioinformatics and Biotechnology* (<https://bioinform.jmir.org>), 10.04.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in *JMIR*



Bioinformatics and Biotechnology, is properly cited. The complete bibliographic information, a link to the original publication on <https://bioinform.jmir.org/>, as well as this copyright and license information must be included.

# Designing a Finite Element Model to Determine the Different Fixation Positions of Tracheal Catheters in the Oral Cavity for Minimizing the Risk of Oral Mucosal Pressure Injury: Comparison Study

Zhiwei Wang<sup>1,2</sup>, MSc; Zhenghui Dong<sup>1</sup>, MSc; Xiaoyan He<sup>2</sup>, BSc; ZhenZhen Tao<sup>3</sup>, MSc; Jinfang QI<sup>1</sup>, MSc; Yatian Zhang<sup>4</sup>, MSc; Xian Ma<sup>4</sup>, MSc

<sup>1</sup>The Sixth Affiliated Hospital of Xinjiang Medical University, No. 39 Wuxing South Road, Tianshan District, Urumqi, Xinjiang, China

<sup>2</sup>Department of Critical Care Medicine, Qinghai Fifth People's Hospital, Xining, China

<sup>3</sup>Emergency Department, The Fourth Affiliated Hospital of Xinjiang Medical University, Xin Jiang, China

<sup>4</sup>School of Nursing, Xinjiang Medical University, Xin Jiang, China

## Corresponding Author:

Zhenghui Dong, MSc

The Sixth Affiliated Hospital of Xinjiang Medical University, No. 39 Wuxing South Road, Tianshan District, Urumqi, Xinjiang, China

## Abstract

**Background:** Despite being an important life-saving medical device to ensure smooth breathing in critically ill patients, the tracheal tube causes damage to the oral mucosa of patients during use, which increases not only the pain but also the risk of infection.

**Objective:** This study aimed to establish finite element models for different fixation positions of tracheal catheters in the oral cavity to identify the optimal fixation position that minimizes the risk of oral mucosal pressure injury.

**Methods:** Computed tomography data of the head and face from healthy male subjects were selected, and a 3D finite element model was created using Mimics 21 and Geomagic Wrap 2021 software. A pressure sensor was used to measure the actual pressure exerted by the oral soft tissue on the upper and lower lips, as well as the left and right mouth corners of the tracheal catheter. The generated model was imported into Ansys Workbench 22.0 software, where all materials were assigned appropriate values, and boundary conditions were established. Vertical loads of 2.6 N and 3.43 N were applied to the upper and lower lips, while horizontal loads of 1.76 N and 1.82 N were applied to the left and right corners of the mouth, respectively, to observe the stress distribution characteristics of the skin, mucosa, and muscle tissue in four fixation areas.

**Results:** The mean (SD) equivalent stress and shear stress of the skin and mucosal tissues were the lowest in the left mouth corner (28.42 [0.65] kPa and 6.58 [0.16] kPa, respectively) and progressively increased in the right mouth corner (30.72 [0.98] kPa and 7.05 [0.32] kPa, respectively), upper lip (35.20 [0.99] kPa and 7.70 [0.17] kPa, respectively), and lower lip (41.79 [0.48] kPa and 10.02 [0.44] kPa, respectively;  $P < .001$  for both stresses). The equivalent stress and shear stress of the muscle tissue were the lowest in the right mouth angle (34.35 [0.52] kPa and 5.69 [0.29] kPa, respectively) and progressively increased in the left mouth corner (35.64 [1.18] kPa and 5.74 [0.30] kPa, respectively), upper lip (43.17 [0.58] kPa and 8.91 [0.55] kPa, respectively), and lower lip (43.17 [0.58] kPa and 11.96 [0.50] kPa, respectively;  $P < .001$  for both stresses). The equivalent stress and shear stress of muscle tissues were significantly greater than those of skin and mucosal tissues in the four fixed positions, and the difference was statistically significant ( $P < .05$ ).

**Conclusions:** Fixation of the tracheal catheter at the left and right oral corners results in the lowest equivalent and shear stresses, while the lower lip exhibited the highest stresses. We recommend minimizing the contact time and area of the lower lip during tracheal catheter fixation, and to alternately replace the contact area at the left and right oral corners to prevent oral mucosal pressure injuries.

(JMIR Bioinform Biotech 2025;6:e69298) doi:[10.2196/69298](https://doi.org/10.2196/69298)

## KEYWORDS

tracheal catheter; fixed position; oral mucosal pressure injury; finite element; biomechanical analysis

Introduction

The primary method of respiratory support for critically ill patients in the intensive care unit (ICU) is oral tube intubation, which ensures airway patency, increases ventilation volume, and enhances lung function. However, the use of oral tube intubation may lead to oral mucosal pressure injury (OMPI) due to excessive or prolonged pressure, friction, and shear forces [1]. OMPI can increase patient pain, elevate the risk of infection, impose a financial burden on health care, increase staff workload, and even result in medical disputes. The incidence of OMPI in patients in the ICU ranges from 2.95% to 49.2%, with different fixation positions and methods of tracheal catheterization influencing its occurrence [2]. While numerous factors contribute to OMPI, including patient-related factors, physiological conditions, the use of specific medications, and nursing-related aspects, there are limited reports addressing the mechanical factors that cause OMPI [3-5]. The International Guidelines for the Clinical Prevention and Treatment of Stress Injuries suggest that finite element models can be employed to evaluate mechanical factors by assessing stress distribution characteristics within tissue structures and predicting the risk of cellular and tissue damage [6].

The purposes of this study were to use the finite element theory contact algorithm to simulate and analyze the compression process of the oral soft tissue when the endotracheal tube is fixed in different fixed positions in the oral cavity, and to explore the stress distribution characteristics of the oral soft tissue under the force of the endotracheal tube. This would help to more realistically and accurately evaluate the actual force on

the oral soft tissue structure and to clarify the reasonable fixed position of the endotracheal tube when it is fixed in the oral cavity, so as to prevent the occurrence of OMPI.

Methods

Finite Element Model

A finite element model of the tracheal catheter positioned at various locations within the mouth was established. The selected participant for the head and facial computed tomography scan was a 28-year-old male volunteer with a normal BMI, measuring 175 cm in height and weighing 72 kg. A total of 512 images, each with a thickness of 0.625 mm, were obtained. The DICOM format data were imported into the 3D reconstruction software Mimics (version 21.0; Materialise) and Geomagic Wrap (version 2021; Raindrop) for model fitting and structural segmentation, respectively. A resistive film pressure sensor was employed to measure the actual pressure exerted by the tracheal catheter in different areas of the patient’s mouth, with each measurement being repeated 100 times to calculate an average value using the gravitational formula. Subsequently, using the measured pressures from solid models as the input data, the Ansys software (version 22.0; ANSYS) was used to import the optimized model, define material properties, remesh the model, and generate an accurate finite element model to conduct finite element analysis based on the defined elastic modulus, Poisson ratio, boundary conditions, and simulated loads for various tissues (skin mucosa and muscle tissue), as well as the tracheal catheter and bone [7,8]. The properties of each material are shown in Table 1; the skin and mucosa are set as nonlinear materials, and the bones are set as isotropic materials

Table . Material properties of the finite element model.

Material	Modulus of elasticity (Mpa)	Young modulus (Mpa)	Shear modulus (Mpa)	Poisson ratio (%)
Tracheal catheter	3	— <sup>a</sup>	1500	0.38
Skeleton	13,400	18,000	—	0.25
Muscle	0.045	0.25	—	0.49
Cutaneous mucosa	—	3	2	0.49

<sup>a</sup>not available.

Ethical Considerations

This study was approved by the Ethics Committee of the Sixth Affiliated Hospital of Xinjiang Medical University (approval number: LFYLLSC20220905-01). All procedures in this study are in line with the ethical standards of the Human Experiments Responsible Committee (Institution and State) and the Declaration of Helsinki.

Setting of Boundary Conditions

In this study, four models representing the upper lip, lower lip, left mouth corner, and right mouth corner were established. The fixed support areas of the models were designated as the top and bottom, allowing for rigid support to be simulated through fixed constraints. A sliding friction contact was implemented between the lip and the tracheal tube, with a friction coefficient set at 1 [9]. A bonded connection was established among the

skin, mucous membrane, and muscle tissue. The model accounted for the effects of gravity in a vertical downward direction, with a gravitational acceleration of 9.8 m/s².

Measurement Indicators

The equivalent stress and shear stress of the skin mucosa and muscle tissue were measured under different fixed positions of the tracheal catheter within the mouth. The stress distribution characteristics of the pressure injury model were analyzed for the fixed positions of the upper lip, lower lip, left mouth corner, and right mouth corner. The stress measurement for each part was conducted 10 times to obtain an average value.

Statistical Analysis

Statistical analysis was performed using SPSS (version 25.0; IBM Corp). Measurement data were expressed as mean (SD). One-way ANOVA was employed for comparisons between

groups, while the *t* test was used for intragroup comparisons. A *P* value of less than .05 was considered statistically significant.

Results

Model Verification

A finite element model of the tracheal catheter was established with a total of 14,635 nodes and 8267 elements at various fixed positions within the oral cavity. This model included the ilium of the upper and lower jaws, as well as the skin, mucosa, and muscle tissues of the oral cavity. The extreme values and

distribution trends of stress at the mouth angle and lower lip were consistent with the findings of Amrani et al [9], indicating the effectiveness of the modeling approach employed in this study.

Equivalent Stress

The equivalent stress of the skin mucosa was the lowest in the left mouth corner, and then progressively increased in the right mouth corner, upper lip, and lower lip. In contrast, the equivalent stress of muscle tissue was the highest in the right mouth corner, followed by the left mouth corner, upper lip, and lower lip. Notably, the equivalent stress of muscle tissue was significantly greater than that of the skin mucosal tissue (*P*<.001; Table 2).

Table . Comparison of equivalent stress results between skin mucosa and muscle tissue (kPa, n=10).

Position	Cutaneous mucosa, mean (SD)	Muscle tissue, mean (SD)	<i>t</i> test ( <i>df</i> )	<i>P</i> value	95% CI
Upper lip	35.20 (0.99)	43.59 (0.84)	−20.371 (9)	<.001	−9.252 to −7.522
Lower lip	41.82 (0.92)	48.35 (0.92)	−15.927 (9)	<.001	−7.389 to −5.667
Left mouth corner	28.42 (0.65)	35.64 (1.18)	−16.924 (9)	<.001	−8.118 to −6.325
Right mouth corner	30.72 (0.99)	34.34 (0.38)	−10.789 (9)	<.001	−3.420 to −2.912
<i>F</i> <sub>1</sub> -score	430.942	573.406	N/A <sup>a</sup>	N/A	N/A
<i>P</i> value	<.001	<.001	N/A	N/A	N/A

<sup>a</sup>not available.

Shear Stress

The shear stress of the skin mucosal tissue was the lowest in the left mouth corner, and progressively increased in the right mouth corner, upper lip, and lower lip. In contrast, the shear stress of the muscle tissue was the lowest in the right mouth

corner, and progressively increased in the left mouth corner, upper lip, and lower lip. At the four fixed positions, the shear stress of the left and right oral muscle tissue was lower than that of the skin mucosa, while the shear stress of the upper and lower lip muscle tissue was higher than that of the skin mucosal tissue (*P*<.005; Table 3)

Table . Comparison of shear stress results between the skin mucosa and muscle tissue (kPa, n=10).

Position	Cutaneous mucosa, mean (SD)	Muscle tissue, mean (SD)	<i>t</i> test ( <i>df</i> )	<i>P</i> value	95% CI
Upper lip	7.60 (0.21)	8.91 (0.39)	−8.959 (9)	<.001	−1.613 to −0.998
Lower lip	10.17 (0.16)	11.69 (0.78)	−5.057 (9)	<.001	−2.145 to −0.882
Left mouth corner	6.58 (0.17)	5.79 (0.33)	6.799 (9)	.001	0.543 to 1.030
Right mouth corner	7.45 (0.36)	5.69 (0.29)	11.972 (9)	<.001	1.450 to 2.068
<i>F</i> <sub>1</sub> -score	244.363	126.411	N/A <sup>a</sup>	N/A	N/A
<i>P</i> value	<.001	<.001	N/A	N/A	N/A

<sup>a</sup>not available.

Comparison of Equivalent Stress and Shear Stress in the Mucosal Tissue of the Upper and Lower Lips and the Left and Right Mouth Corners

Equivalent stress was found to be lower in the upper lip compared to the lower lip, and the left mouth corner exhibited

lower stress than the right mouth corner (*P*<.001; Table 4-5). In terms of shear stress, the upper lip also showed significantly lower values than the lower lip (*P*<.001;Table5), while the left mouth corner had lower shear stress than the right mouth corner (*P*<.001; Table 5).

**Table .** Comparison of the results of equivalent stress and shear force in the left and right mouth corners (kPa, n=10).

Position	Left side mouth corner, mean (SD)	Right side mouth corner, mean (SD)	<i>t</i> test ( <i>df</i> )	<i>P</i> value	95% CI
Equivalent stress	28.42 (0.65)	30.72 (0.99)	−6.160 (9)	<.001	−3.094 to −1.520
Shear stress	6.58 (0.17)	7.45 (0.36)	−6.984 (9)	<.001	−1.125 to −0.605

**Table .** Comparison of the results of equivalent stress and shear force in the skin mucosal tissue of the upper and lower lip (kPa, n=10).

Position	Upper lip, mean (SD)	Lower lip, mean (SD)	<i>t</i> test ( <i>df</i> )	<i>P</i> value	95% CI
Equivalent stress	35.20 (0.99)	41.82 (0.92)	−15.472 (9)	<.001	−7.519 to 5.721
Shear stress	7.60 (0.21)	10.17 (0.16)	−16.769 (9)	<.001	−2.931 to −2.279

**Comparison of Equivalent Stress and Shear Stress in the Muscle Tissue of the Upper and Lower Lips and Left and Right Mouth Corners**

The equivalent stress was the lower in the upper lip than in the lower lip (*P*<.001), and higher in the left mouth corner than in

the right mouth corner (*P*=.004; [Table 6](#)). The shear stress was lower in the upper lip than in the lower lip (*P*<.001), and lower in the left mouth angle than in the right mouth angle (*P*=.298; [Table 7](#))

**Table .** Comparison of equivalent stress and shear force results in the left and right mouth corners (kPa, n=10).

Position	Left side mouth corner, mean (SD)	Right side mouth corner, mean (SD)	<i>t</i> test ( <i>df</i> )	<i>P</i> value	95% CI
Equivalent stress	35.64 (1.18)	34.34 (0.38)	3.308 (9)	.004	0.474 to 2.124
Shear stress	5.74 (0.30)	5.69 (0.29)	1.071 (9)	.50	−0.221 to 0.435

**Table .** Comparison of equivalent stress and shear force results in the muscle tissue of the upper and lower lips (kPa, n=10).

Position	Upper lip, mean (SD)	Lower lip, mean (SD)	<i>t</i> test ( <i>df</i> )	<i>P</i> value	95% CI
Equivalent stress	43.59 (0.84)	48.35 (0.92)	−12.115 (9)	<.001	−5.587 to −3.935
Shear stress	8.91 (0.39)	11.69 (0.78)	−12.477 (9)	<.001	−3.561 to −2.545

**Stress Distribution Rules of the Four Groups of Models**

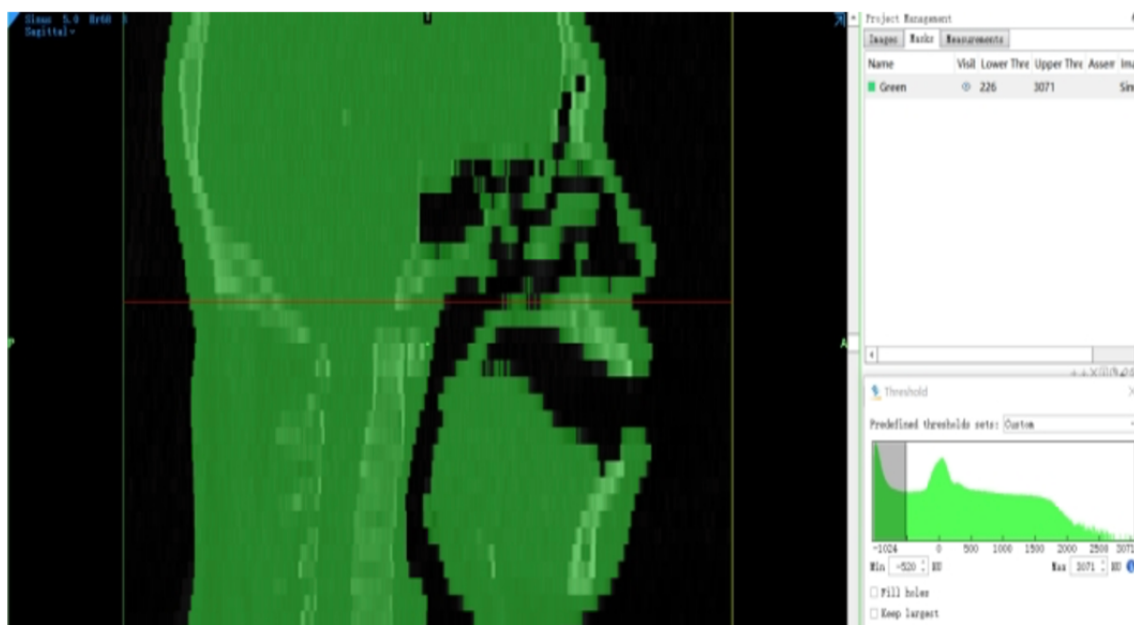
The equivalent stress range of the skin mucosa and muscle tissue gradually extends from the stress center to the periphery. In this study, the application direction of the forces on the upper and lower lips is vertical, with the maximum peak values of both equivalent stress and shear stress occurring at the stress point and subsequently radiating outward in the vertical direction. Conversely, the forces applied at the left and right mouth corners are horizontal, causing the stress range to spread horizontally, with the highest stress values appearing at the direct contact point between the tracheal catheter and the mucosal tissue. The distribution of shear stress is centered on the soft tissue stress point and encompasses the entire lip, mandibular region, and both sides of the face, resulting in a broader range of stress. The

equivalent stress and shear stress at the mouth corners are significantly lower than those at the upper and lower lips.

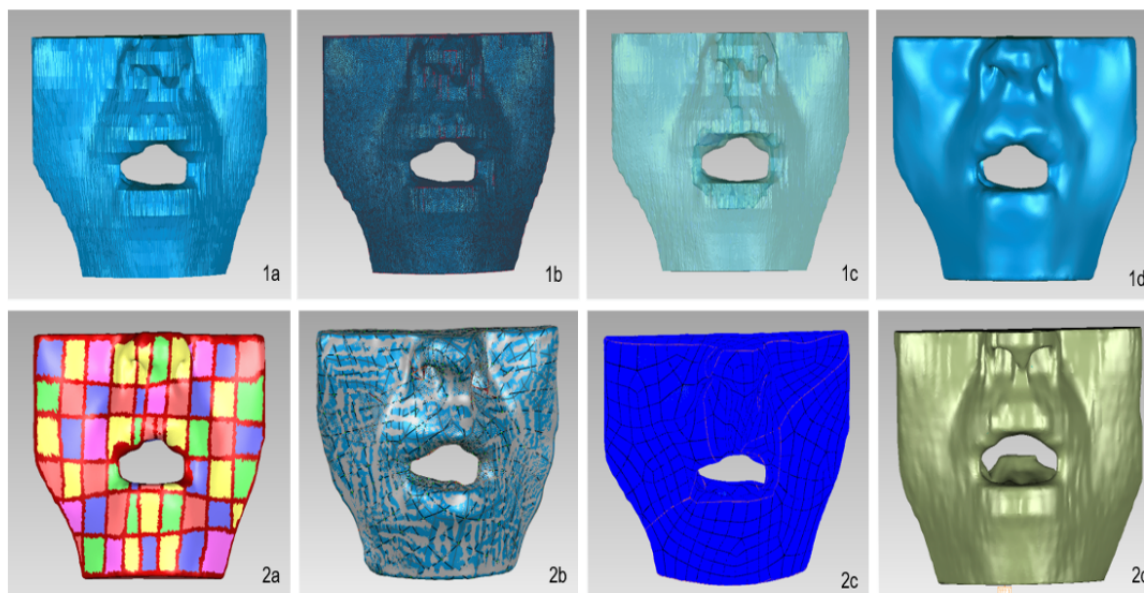
To explore the underlying reasons, when the tracheal catheter is fixed at the corner of the mouth, it makes contact with the corner, the upper lip, and the lower lip. The pressure, shear force, and friction generated by this contact are dispersed across the three contact surfaces of the mouth and the upper and lower lips. The contact surface between the tracheal tube and the upper and lower lips serves as the primary stress point, leading to greater stress values at the upper and lower lips compared to the corners of the mouth, with the lower lip experiencing the highest stress. The results of the finite element analysis indicate that the stress at the corners of the mouth is lower, followed by that at the upper lip ([Figures 1-4](#)).



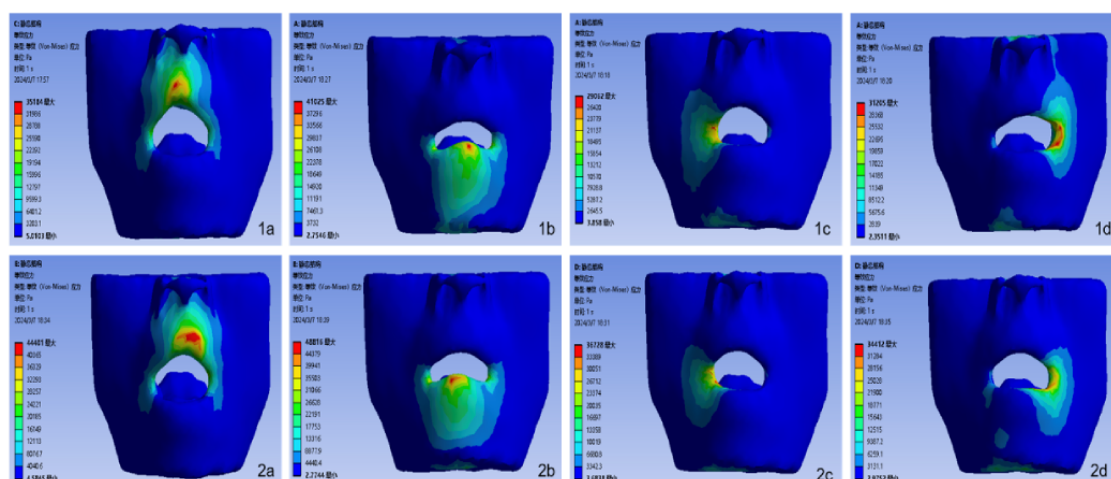
**Figure 1.** Mimics21.0 software was used to reconstruct the patient's head, face and oral tissues in 3D with an interval of 0.25 mm, and the contour range of the skin mucosa and muscle tissue was constructed through the thresholds of different tissues.



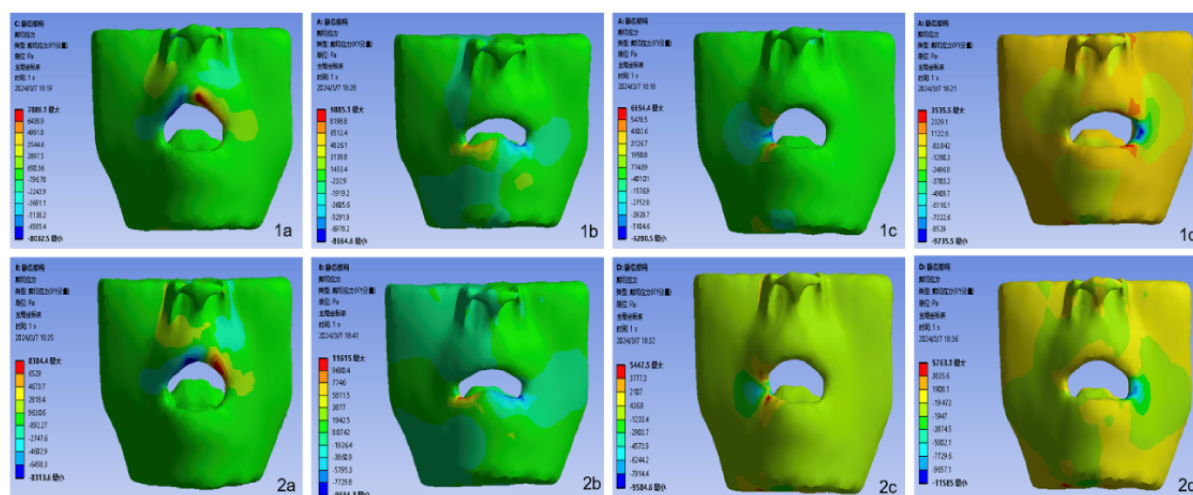
**Figure 2.** The probe contour line is used to redraw the contour line of the model, so that the surface pieces are more extensible and the concave and convex surfaces are reduced. The structural patch trims the model patch again to make the patch smoother and smoother, which is consistent with the characteristics of the skin tissue. Construct grids, and optimize and adjust all patch nodes and elements. Finally, the fitting surface constructs a model that is similar to the actual oral and facial features of the human body.



**Figure 3.** (1a-1d) Equivalent stress nephogram of two tissues at 4 fixed locations: upper lip, lower lip, left mouth corner, and right mouth corner. (2a-2d) Equivalent stress nephogram of the muscle tissue of the upper lip, lower lip, left mouth corner, and right mouth corner.



**Figure 4.** Shear stress nephogram of two tissues at 4 fixed locations. (1a-1d) Shear stress nephogram of the mucosa tissue of the upper lip, lower lip, left mouth corner, and right mouth corner. (2a-2d) Shear stress nephogram of the muscle tissue of the upper lip, lower lip, left mouth corner, and right mouth corner.



## Discussion

The results of this study showed that when the tracheal tube was in contact with the lower lip, the equivalent stress and shear stress values of muscle tissue and mucosal tissue were the largest, followed by the upper lip, and the left and right mouth angles were lower than those of the upper and lower lip. Finite element analysis modeling is a powerful bioengineering technique employed to assess tissue loading, encompassing the interactions between tissues, objects, and medical devices. This numerical method effectively addresses mechanical problems [10]. It enables rapid and accurate stress-strain analysis of the structure, shape, load, and mechanical properties of materials in any given model [11]. Moreover, finite element analysis objectively and accurately reflects the distribution of stress, strain, and deformation, and has gained widespread application in oral biomechanics research in recent years [12].

The tracheal catheter is a critical instrument for mechanical ventilator-assisted therapy in patients in the ICU; however, the

catheter itself and improper fixation methods may lead to OMPI [6]. From a biomechanical perspective, the OMPI associated with tracheal catheters primarily results from vertical pressure, shear forces, and friction [13]. Continuous mechanical loading on soft tissues is the main contributor to stress injuries, typically occurring at bony prominences or in areas contacting medical devices. When skin or deep tissue deformation persists for a certain duration owing to the pressure from medical devices, pressure injuries may develop [14]. In this study, the mechanical load originated from the force exerted by the tracheal catheter on the oral soft tissue. Contact between the tracheal catheter and the oral mucosal tissue resulted in continuous pressure, leading to tissue deformation in the mucosa. Research indicates that tracheal catheters and their fixation devices are stiffer than oral soft tissues. When the mechanical properties of these instruments do not align with those of the soft tissues, deformation occurs in the latter, concentrating mechanical stress and strain at the points of direct contact, which then gradually extends to the surrounding areas [15,16].

Continuous vertical pressure on soft tissues is a significant factor in the occurrence of stress injuries. The incidence of OMPI correlates with the intensity and duration of pressure; the greater the pressure and the longer its application, the higher the risk of developing OMPI is [17]. Furthermore, when the tracheal tube is improperly fitted and fixed too tightly, the pressure and shear force exerted will increase [14]. Shear forces applied to deep skin tissues can obstruct capillaries, leading to localized ischemia and hypoxia, which may result in deep tissue necrosis. Consequently, damage from shear forces is often undetected in the early stages and is more challenging to heal than damage from typical wounds [13]. Friction arises from the movement between the oral mucosal tissue and the surface of the tracheal tube; while it does not directly cause OMPI, it can compromise the epidermal cuticle, leading to the shedding of the mucosal surface layer and heightened sensitivity to pressure injuries. Once the compromised oral mucosal tissue is subjected to stimuli from saliva and other secretions, the risk of pressure injury escalates. Additionally, friction raises the temperature of the local mucosal tissue, disrupts the local microenvironment, alters pH levels, and increases tissue oxygen consumption, further exacerbating tissue ischemia and heightening the risk of OMPI [16].

The magnitude of the internal mechanical load required to cause tissue damage depends on the duration of the applied force and the specific biomechanical tolerance of the stressed tissue, which is influenced by factors such as age, shape, health status, and the functional capacity of the body systems, including tissue repair ability [18]. Both high loads applied for short durations and low loads sustained over extended periods can lead to tissue damage [18-20]. Continuous loading is one of the primary contributors to this damage; it refers to loads applied over prolonged periods (ranging from a few minutes to several hours or even days), also known as quasi-static mechanical loading. Research indicates that when soft tissues come into contact with the support surfaces of medical devices, pressure and shear forces are generated between the soft tissues and these surfaces [21]. This interaction results in distortion and deformation of the soft tissues under pressure, affecting both the skin and deeper tissues (including fat, connective tissue, and muscle), leading to stress and strain within the tissues [21]. Excessive internal

stress in the tissues can disrupt intracellular material transport by damaging cellular structures (such as the cytoskeleton or plasma membrane) or by hindering the transport process itself (for example, by reducing blood perfusion, impairing lymphatic function, and affecting material transport in the interstitial space), which can ultimately result in cell death and trigger an inflammatory response. Concurrently, the emergence of endothelial cell spacing increases vascular permeability, leading to inflammatory edema, which further exacerbates the mechanical load on cells and tissues due to elevated tissue pressure, thus contributing to the development of pressure injuries [22-24].

According to the results of finite element analysis, the stress experienced by the lower lip is the highest, followed by the upper lip, with levels significantly exceeding those at the corners of the mouth. Therefore, in clinical practice, when fixing a tracheal catheter, it is advisable to select the mouth corner to maximize the contact surface area between the catheter and this region. Placing the tracheal catheter in the middle of the mouth minimizes the contact time between the catheter and the oral mucosa. Additionally, regular changes in the fixation position can help redistribute pressure, thereby reducing pressure, shear forces, and friction on the oral mucosa, ultimately lowering the risk of OMPI.

This study analyzed alterations in the stress experienced by oral soft tissue under pressure at various fixation positions of the tracheal catheter within the mouth, from a biomechanical perspective. It provides a theoretical foundation for preventing OMPI in patients with tracheal catheters in the ICU. While this study effectively simulates the biomechanical effects of contact between oral soft tissue and the tracheal catheter, it does not fully replicate the actual forces experienced by oral soft tissue in real-life situations, as the area of contact between the tracheal catheter and the oral soft tissue cannot be completely simulated. Additionally, the study included only one young adult male, which limits the generalizability of the findings. Therefore, it is essential to include participants of varying genders and ages to enhance the scientific validity of the research. Furthermore, improvements in the identification rate and curvature of the 3D grid of the model should be pursued to generate higher-quality 3D models, thereby enhancing data accuracy.

## Data Availability

The data supporting the findings of this study are available from the corresponding author upon reasonable request.

## Conflicts of Interest

None declared.

## References

1. Maofan Y, Huilan Z, Keyu C, et al. Research progress on oral mucosal pressure injuries in patients with oral tube intubation in the ICU. *Journal of Nursing* 2023;38(2):21-24. [doi: [10.3870/j.issn.1001-4152.2023.02.021](https://doi.org/10.3870/j.issn.1001-4152.2023.02.021)]
2. Qian L, Lizhu W, Yirong Z, et al. Research progress on oral mucosal pressure injuries in patients with oral tube intubation in the ICU. *Chinese Journal of Acute and Critical Care* 2023;4(5):473-477. [doi: [10.3761/j.issn.2096-7446.2023.05.019](https://doi.org/10.3761/j.issn.2096-7446.2023.05.019)]
3. Li T, Min L, Qian Y. Research progress on nursing care for oral mucosal pressure injuries in patients with oral catheterization. *Nurs Res* 2023;37(20):3682-3686. [doi: [10.12102/j.issn.1009-6493.2023.20.013](https://doi.org/10.12102/j.issn.1009-6493.2023.20.013)]



4. Choi BK, Kim MS, Kim SH. Risk prediction models for the development of oral-mucosal pressure injuries in intubated patients in intensive care units: A prospective observational study. *J Tissue Viability* 2020 Nov;29(4):252-257. [doi: [10.1016/j.jtv.2020.06.002](https://doi.org/10.1016/j.jtv.2020.06.002)]
5. Tian-kuang L, Yu-Juan L, Huan L, et al. Research progress on the characteristics and nursing care of mucosal pressure injuries in different parts of ICU patients. *Journal of Nursing* 2022;29(8):35-39. [doi: [10.16460/j.issn1008-9969.2022.08.-035](https://doi.org/10.16460/j.issn1008-9969.2022.08.-035)]
6. Jia L, Deng Y, Xu Y, et al. Development and validation of a nomogram for oral mucosal membrane pressure injuries in ICU patients: A prospective cohort study. *J Clin Nurs* 2024 Oct;33(10):4112-4123. [doi: [10.1111/jocn.17296](https://doi.org/10.1111/jocn.17296)] [Medline: [38797947](https://pubmed.ncbi.nlm.nih.gov/38797947/)]
7. Zhang K, Chen Y, Feng C, et al. Machine learning based finite element analysis for personalized prediction of pressure injury risk in patients with spinal cord injury. *Comput Methods Programs Biomed* 2025 Apr;261:108648. [doi: [10.1016/j.cmpb.2025.108648](https://doi.org/10.1016/j.cmpb.2025.108648)] [Medline: [39922124](https://pubmed.ncbi.nlm.nih.gov/39922124/)]
8. Keenan BE, Evans SL, Oomens CWJ. A review of foot finite element modelling for pressure ulcer prevention in bedrest: Current perspectives and future recommendations. *J Tissue Viability* 2022 Feb;31(1):73-83. [doi: [10.1016/j.jtv.2021.06.004](https://doi.org/10.1016/j.jtv.2021.06.004)]
9. Amrani G, Gefen A. Which endotracheal tube location minimises the device-related pressure ulcer risk: The centre or a corner of the mouth? *Int Wound J* 2020 Apr;17(2):268-276. [doi: [10.1111/iwj.13267](https://doi.org/10.1111/iwj.13267)] [Medline: [31724822](https://pubmed.ncbi.nlm.nih.gov/31724822/)]
10. Welch-Phillips A, Gibbons D, Ahern DP, Butler JS. What Is Finite Element Analysis? *Clin Spine Surg* 2020 Oct;33(8):323-324. [doi: [10.1097/BSD.0000000000001050](https://doi.org/10.1097/BSD.0000000000001050)] [Medline: [32675684](https://pubmed.ncbi.nlm.nih.gov/32675684/)]
11. Wang CX, Rong QG, Zhu N, Ma T, Zhang Y, Lin Y. Finite element analysis of stress in oral mucosa and titanium mesh interface. *BMC Oral Health* 2023 Jan 17;23(1):25. [doi: [10.1186/s12903-022-02703-3](https://doi.org/10.1186/s12903-022-02703-3)] [Medline: [36650512](https://pubmed.ncbi.nlm.nih.gov/36650512/)]
12. Guo R, Lam XY, Zhang L, Li W, Lin Y. Biomechanical analysis of miniscrew-assisted molar distalization with clear aligners: a three-dimensional finite element study. *Eur J Orthod* 2024 Jan 1;46(1):cjad077. [doi: [10.1093/ejo/cjad077](https://doi.org/10.1093/ejo/cjad077)] [Medline: [38134411](https://pubmed.ncbi.nlm.nih.gov/38134411/)]
13. Zhijun R, Xinhua X, Anqi C, et al. New progress in the prevention of stress injuries induced by mechanical factors. *Nurs Res* 2017;31(10):1167-1170. [doi: [10.3969/j.issn.1009-6493.2017.10.005](https://doi.org/10.3969/j.issn.1009-6493.2017.10.005)]
14. Na W, Yuan-Ting L, Yin-shi X, et al. Summary of evidence for the prevention of medical device-related stress injuries in ICU patients. *Chinese Journal of Practical Nursing* 2022;38(13):992-997. [doi: [10.3760/cma.j.cn211501-20210710-01863](https://doi.org/10.3760/cma.j.cn211501-20210710-01863)]
15. Gefen A. The aetiology of medical device-related pressure ulcers and how to prevent them. *Br J Nurs* 2021 Aug 12;30(15):S24-S30. [doi: [10.12968/bjon.2021.30.15.S24](https://doi.org/10.12968/bjon.2021.30.15.S24)] [Medline: [34379465](https://pubmed.ncbi.nlm.nih.gov/34379465/)]
16. Mak AFT, Zhang M, Tam EWC. Biomechanics of pressure ulcer in body tissues interacting with external forces during locomotion. *Annu Rev Biomed Eng* 2010 Aug 15;12:29-53. [doi: [10.1146/annurev-bioeng-070909-105223](https://doi.org/10.1146/annurev-bioeng-070909-105223)] [Medline: [20415590](https://pubmed.ncbi.nlm.nih.gov/20415590/)]
17. Lustig A, Margi R, Orlov A, Orlova D, Azaria L, Gefen A. The mechanobiology theory of the development of medical device-related pressure ulcers revealed through a cell-scale computational modeling framework. *Biomech Model Mechanobiol* 2021 Jun;20(3):851-860. [doi: [10.1007/s10237-021-01432-w](https://doi.org/10.1007/s10237-021-01432-w)] [Medline: [33606118](https://pubmed.ncbi.nlm.nih.gov/33606118/)]
18. Grigatti A, Gefen A. The biomechanical efficacy of a hydrogel-based dressing in preventing facial medical device-related pressure ulcers. *Int Wound J* 2022 Aug;19(5):1051-1063. [doi: [10.1111/iwj.13701](https://doi.org/10.1111/iwj.13701)] [Medline: [34623741](https://pubmed.ncbi.nlm.nih.gov/34623741/)]
19. Bogie KM, Zhang GQ, Roggenkamp SK, et al. Individualized Clinical Practice Guidelines for Pressure Injury Management: Development of an Integrated Multi-Modal Biomedical Information Resource. *JMIR Res Protoc* 2018 Sep 6;7(9):e10871. [doi: [10.2196/10871](https://doi.org/10.2196/10871)] [Medline: [30190252](https://pubmed.ncbi.nlm.nih.gov/30190252/)]
20. Morrow MM, Hughes LC, Collins DM, Vos-Draper TL. Clinical Remote Monitoring of Individuals With Spinal Cord Injury at Risk for Pressure Injury Recurrence Using mHealth: Protocol for a Pilot, Pragmatic, Hybrid Implementation Trial. *JMIR Res Protoc* 2024 Apr 10;13:e51849. [doi: [10.2196/51849](https://doi.org/10.2196/51849)] [Medline: [38598267](https://pubmed.ncbi.nlm.nih.gov/38598267/)]
21. Gawlitta D, Li W, Oomens CWJ, Baaijens FPT, Bader DL, Bouten CVC. The relative contributions of compression and hypoxia to development of muscle tissue damage: an in vitro study. *Ann Biomed Eng* 2007 Feb;35(2):273-284. [doi: [10.1007/s10439-006-9222-5](https://doi.org/10.1007/s10439-006-9222-5)] [Medline: [17136445](https://pubmed.ncbi.nlm.nih.gov/17136445/)]
22. Caulk AW, Chatterjee M, Barr SJ, Contini EM. Mechanobiological considerations in colorectal stapling: Implications for technology development. *Surg Open Sci* 2023 Jun;13:54-65. [doi: [10.1016/j.sopen.2023.04.004](https://doi.org/10.1016/j.sopen.2023.04.004)] [Medline: [37159635](https://pubmed.ncbi.nlm.nih.gov/37159635/)]
23. Pan Y, Yang D, Zhou M, et al. Advance in topical biomaterials and mechanisms for the intervention of pressure injury. *iScience* 2023 Jun 16;26(6):106956. [doi: [10.1016/j.isci.2023.106956](https://doi.org/10.1016/j.isci.2023.106956)] [Medline: [37378311](https://pubmed.ncbi.nlm.nih.gov/37378311/)]
24. Peko Cohen L, Ovadia-Blechman Z, Hoffer O, Gefen A. Dressings cut to shape alleviate facial tissue loads while using an oxygen mask. *Int Wound J* 2019 Jun;16(3):813-826. [doi: [10.1111/iwj.13101](https://doi.org/10.1111/iwj.13101)] [Medline: [30838792](https://pubmed.ncbi.nlm.nih.gov/30838792/)]

## Abbreviations

**ICU:** intensive care unit

**OMPI:** oral mucosal pressure injury

*Edited by H Yan; submitted 26.11.24; peer-reviewed by SB Shenoy, YH Shash; revised version received 30.03.25; accepted 29.04.25; published 11.07.25.*

*Please cite as:*

Wang Z, Dong Z, He X, Tao Z, QI J, Zhang Y, Ma X

*Designing a Finite Element Model to Determine the Different Fixation Positions of Tracheal Catheters in the Oral Cavity for Minimizing the Risk of Oral Mucosal Pressure Injury: Comparison Study*

*JMIR Bioinform Biotech* 2025;6:e69298

URL: <https://bioinform.jmir.org/2025/1/e69298>

doi: [10.2196/69298](https://doi.org/10.2196/69298)

© Zhiwei Wang, Zhenghui Dong, Xiaoyan He, ZhenZhen Tao, Jinfang QI, Yatian Zhang, Xian Ma. Originally published in JMIR Bioinformatics and Biotechnology (<https://bioinform.jmir.org>), 11.7.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Bioinformatics and Biotechnology, is properly cited. The complete bibliographic information, a link to the original publication on <https://bioinform.jmir.org/>, as well as this copyright and license information must be included.



# Extracting Knowledge From Scientific Texts on Patient-Derived Cancer Models Using Large Language Models: Algorithm Development and Validation Study

Jiarui Yao<sup>1,2\*</sup>, PhD; Zinaida Perova<sup>3\*</sup>, PhD; Tushar Mandloi<sup>3</sup>, MSc; Elizabeth Lewis<sup>3</sup>, MSc; Helen Parkinson<sup>3</sup>, PhD; Guergana Savova<sup>1,2</sup>, PhD

<sup>1</sup>Computational Health Informatics Program, Boston Children's Hospital, 401 Park Drive, Boston, MA, United States

<sup>2</sup>Harvard Medical School, Boston, MA, United States

<sup>3</sup>European Molecular Biology Laboratory, European Bioinformatics Institute, Hinxton, United Kingdom

\* these authors contributed equally

## Corresponding Author:

Jiarui Yao, PhD

Computational Health Informatics Program, Boston Children's Hospital, 401 Park Drive, Boston, MA, United States

## Abstract

**Background:** Patient-derived cancer models (PDCMs) have become essential tools in cancer research and preclinical studies. Consequently, the number of publications on PDCMs has increased significantly over the past decade. Advances in artificial intelligence, particularly in large language models (LLMs), offer promising solutions for extracting knowledge from scientific literature at scale.

**Objective:** This study aims to investigate LLM-based systems, focusing specifically on prompting techniques for the automated extraction of PDCM-related entities from scientific texts.

**Methods:** We explore 2 LLM-prompting approaches. The classic method, direct prompting, involves manually designing a prompt. Our direct prompt consists of an instruction, entity-type definitions, gold examples, and a query. In addition, we experiment with a novel and underexplored prompting strategy—soft prompting. Unlike direct prompting, soft prompts are trainable continuous vectors that learn from provided data. We evaluate both prompting approaches across state-of-the-art proprietary and open LLMs.

**Results:** We manually annotated 100 abstracts of PDCM-relevant papers, focusing on PDCM papers with data deposited in the CancerModels.Org platform. The resulting gold annotations span 15 entity types for a total 3313 entity mentions, which we split across training (2089 entities), development (542 entities) and held-out, eye-off test (682 entities) sets. Evaluation includes the standard metrics of precision or positive predictive value, recall or sensitivity, and  $F_1$ -score (harmonic mean of precision and recall) in 2 settings: an exact match setting, where spans of gold and predicted annotations have to match exactly, and an overlapping match setting, where the spans of gold and predicted annotations have to overlap. GPT4-o with direct prompting achieved  $F_1$ -scores of 50.48 and 71.36 for exact and overlapping match settings, respectively. In both evaluation settings, LLaMA3 soft prompting improved performance over direct prompting ( $F_1$ -score from 7.06 to 46.68 in the exact match setting; and 12.0 to 71.80 in the overlapping evaluation setting). Results with LLaMA3 soft prompting are slightly higher than GPT4-o direct prompting in the overlapping match evaluation setting.

**Conclusions:** We investigated LLM-prompting techniques for the automatic extraction of PDCM-relevant entities from scientific texts, comparing the traditional direct prompting approach with the emerging soft prompting method. In our experiments, GPT4-o demonstrated strong performance with direct prompting, maintaining competitive results. Meanwhile, soft prompting significantly enhanced the performance of smaller open LLMs. Our findings suggest that training soft prompts on smaller open models can achieve performance levels comparable to those of proprietary very large language models.

(*JMIR Bioinform Biotech* 2025;6:e70706) doi:[10.2196/70706](https://doi.org/10.2196/70706)

## KEYWORDS

patient-derived cancer models; large language models; knowledge extraction; in-context learning; soft prompting; prompt tuning; information extraction

## Introduction

Patient-derived cancer models (PDCMs) are created from a patient's own tumor sample and capture the complexity of human tumors to enable more accurate, personalized drug development and treatment selection. These models, including patient-derived xenografts (PDXs), organoids, and cell lines, allow researchers to test treatments and identify the most effective therapies, and have emerged as indispensable tools in both cancer research and precision medicine. The US National Institutes of Health (NIH) have made significant investments in the generation and characterization of these models, with more than US \$3 billion dedicated to active grants referencing PDCMs with a component of their research based on data extracted from the NIH RePORTER [1] for fiscal year 2024 alone. The number of publications using PDCMs continues to increase generating substantial and rich metadata and data that require standardization, harmonization, and integration to maximize the impact of these models and their associated data within the research and clinical communities. CancerModels.Org platform [2] serves as a unified gateway to the largest collection of PDCMs and related data. It empowers researchers and clinicians to discover suitable models for testing research hypotheses, conducting large-scale drug screenings, and advancing precision medicine initiatives. Extraction of PDCM-relevant knowledge and its harmonization within CancerModels.Org is essential to ensure that basic and translational researchers, bioinformaticians, and tool developers have access to PDCM knowledge. While manual curation of publications ensures high accuracy when performed by domain experts, it is time-consuming and labor-intensive. Thus, a more streamlined and efficient knowledge acquisition method is needed to address the growing demand within the scientific community for the timely availability of the PDCM metadata and its associated data.

In parallel, large language models (LLMs) [3-5] often referred to as generative artificial intelligence (AI) systems are trained on vast amounts of data and have demonstrated impressive capabilities in the health care domain [6-8]. Researchers have studied the use of LLMs in addressing various tasks related to health care such as diagnosing conditions [9,10], clinical decision support [11], answering patient questions [12], and medical education [13,14]. It has been shown that LLMs can extract meaningful information from texts [15-17].

In this work, we explore LLM-prompting techniques with the goal of extracting knowledge from PDCM-relevant scholarly publications. We focus on the classic direct prompting [4] and the underexplored soft prompting [18] with state-of-the-art (SOTA) proprietary and open LLMs. Our experimental results provide insights into selecting the optimal prompting methods for specific tasks. The contributions of this paper are:

1. Studying the feasibility of SOTA LLMs as oncology knowledge extractors for PDCM-relevant information from scholarly scientific literature.
2. Creating a manually curated gold dataset spanning 15 entity types for a total 3313 entity mentions from 100 abstracts of PDCM-relevant papers.
3. Researching and comparing, to our knowledge for the first time, direct versus soft prompting techniques for oncology knowledge extraction, specifically PDCM-relevant information from scholarly scientific literature.

## Methods

### Concepts

We define “knowledge” as entities of interest to researchers working with PDCMs in the cancer research field. For example, the patient's diagnosis provides a reference point to confirm that a PDCM faithfully recapitulates the biology of the original tumor and is essential for ensuring the model's relevance and reliability in studies of cancer progression or treatment response. Thus, “diagnosis” is important to understand the model's characteristics in the context of patient's disease. The patient's age can significantly affect the molecular and genetic characteristics of the tumor. For example, pediatric cancers often have distinct genetic drivers and tumor microenvironments compared to cancers in older adults. In addition, age-related biological factors, such as immune system, metabolism, and hormone levels, influence how a tumor responds to treatments. Thus, knowing the patient's age is imperative for predictive accuracy of the model in preclinical testing and relevance of research findings. Therefore, we selected 15 most commonly used CancerModels.Org data model attributes (Table 1), which include the attributes defined in the minimal information standard for patient-derived xenograft models [19] and the draft minimal information standard for in vitro models [20].

**Table .** Entity definitions based on the CancerModels.Org data model with examples and interannotator agreement  $F_1$ -scores in the exact match setting that requires the spans of the annotators to match exactly.

Entity type	Definition	Example	IAA <sup>a</sup>
diagnosis	Diagnosis at the time of collection of the patient tumor used in the cancer model	TNBC <sup>b</sup>	61.67
age_category	Age category of the patient at the time of tissue sampling	Adult, pediatric	60
genetic_effect	Any form of chromosomal rearrangement or gene-level changes	Missense, amplification	57.67
model_type	Type of patient-derived model	PDX <sup>c</sup> , organoid	53.33
molecular_char	Data or assay generated from or performed on the model in this study	RNA sequencing, whole-exome sequencing	54.33
biomarker	Gene, protein or biological molecule identified in or associated with patient's/model's tumor	BRCA1 <sup>d</sup> , IDH <sup>e</sup> , epidermal growth factor receptor 2	61.33
treatment	Treatment received by the patient or tested on the model	Surgery, chemotherapy, PARP-inhibitor	55.67
response_to_treatment	Effect of the treatment on the patient's tumor or model	Progression-free survival, reduced tumor growth	55
sample_type	The type of material used to generate the model or how this material was obtained	Tissue fragment, autopsy	49
tumor_type	Collected tumor type used for generating the model	Primary, recurrent	49.67
cancer_grade	Quantitative or qualitative grade reflecting how quickly the cancer is likely to grow	Grade 1, low-grade	42
cancer_stage	Information about the cancer's extent in the body according to specific type of cancer staging system	TNM <sup>f</sup> system, T0, stage I	59.33
clinical_trial	The type of clinical trial or ClinicalTrials.org identifier	Phase II, prospective randomized clinical trials	60.67
host_strain	The name of the mouse host strain where the tissue sample was engrafted for generating the PDX model	NOD-SCID <sup>g</sup>	61.67
model_id	ID of the patient-derived cancer model generated in this study	PHLC402	100

<sup>a</sup>IAA: interannotator agreement.<sup>b</sup>TNBC: triple-negative breast cancer.<sup>c</sup>PDX: patient-derived xenograft.<sup>d</sup>BRCA1: breast cancer gene 1.<sup>e</sup>IDH: isocitrate dehydrogenase.<sup>f</sup>TNM: tumor node metastasis.<sup>g</sup>NOD-SCID: nonobese diabetic severe combined immunodeficiency.

## Corpus

We used 100 abstracts to develop the gold-standard corpus annotated for the 15 entities (Table 1). The abstracts were chosen from publications linked to the PDCMs submitted to CancerModels.Org platform. They were selected to cover all 3 types of models in the resource-PDXs, organoids, and cell lines. The final corpus is available on GitHub (see Data and Code Availability section).

Three annotators (ZP, TM, and EL) independently labeled entities in all 100 abstracts for a total of 40 hours. The annotation quality was tracked through interannotator agreement (IAA), a measure of agreement between each annotation produced by different annotators working on the same dataset. The IAA is an indication of how difficult the task is for humans and it becomes the target for system development. We used pairwise  $F_1$ -score as the IAA metric [21] in the exact match setting that

requires the spans of the annotators to match exactly. We computed the agreement between each pair of annotators and averaged across the 3 sets of scores. The final IAA for each entity type is reported in Table 1. The IAA range is 42 - 100 indicating moderate agreement. Note that the lowest agreement is for low occurrence entity types, for example, cancer\_grade has only 8 instances with 42 IAA. These low-frequency entity types are more likely to be overlooked by the human experts as annotation is a cognitively demanding task. Thus, to ensure a high-quality gold-standard dataset, we overlaid the single

annotations with an adjudication step, where the annotators discussed annotation disagreements and potential missed annotations to come to final joint decisions. The resulting gold dataset spans 15 entity types for a total 3313 entity mentions (refer Table 2 for distributions) was split into training, development, and test sets in the standard 60:20:20 ratio. The train set was used for creating entity extraction algorithms, the development set for refining the algorithms, and the test set for the final evaluation.

Table . Distribution of entity type annotations across training, development, and test sets.

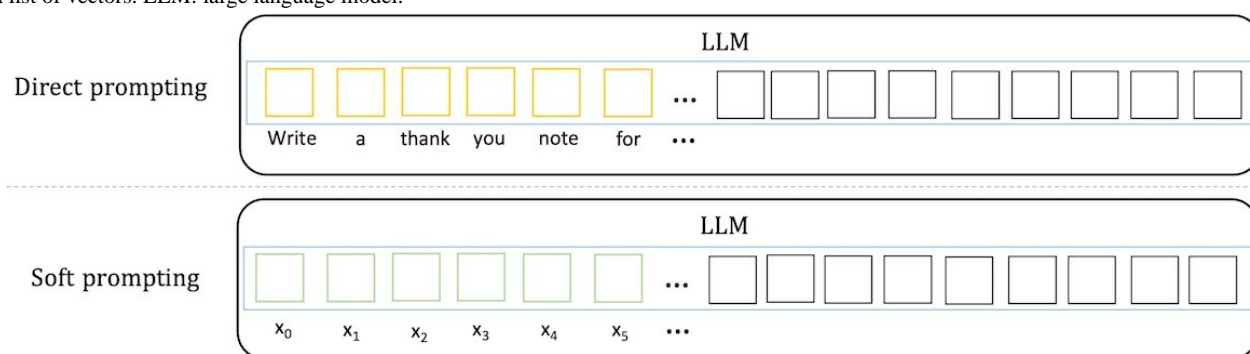
Entity type	Training, n	Development, n	Test, n	Total, n
diagnosis	362	122	114	598
age_category	19	0	0	19
genetic_effect	69	20	33	122
model_type	326	114	110	550
molecular_char	128	37	46	211
biomarker	503	118	163	784
treatment	426	77	130	633
response_to_treatment	99	21	28	148
sample_type	22	8	7	37
tumor_type	61	19	28	108
cancer_grade	6	1	1	8
cancer_stage	7	1	4	12
clinical_trial	35	2	4	41
host_strain	9	0	7	16
model_id	17	2	7	26
Total	2089	542	682	3313

Prompting Methods

Various prompting techniques have been proposed since the emergence of LLMs [22-25]. At a high level, these prompting techniques can be divided into 2 categories, direct prompting [4] and soft prompting [18,24,26] . The main difference between the two methods is the prompt representation, that is whether the prompt consists of human language words or vectors (Figure 1). Direct prompting (or discrete prompting) is the most intuitive and now classic prompting method where users directly interact with LLMs using natural language. For example, a user may ask ChatGPT to “Write a thank you note to an old friend of my parents”; in this case, the text within the quotation marks is a discrete prompt. Soft prompting (or continuous prompting) uses

a machine learning approach to train a sequence of continuous vectors, which are the “virtual tokens” of the prompt. It is worth noting that soft prompting differs from fine tuning. With soft prompting, the LLM parameters are not updated, only the soft prompt parameters are adjusted. In contrast, finetuning requires to update the parameters of the entire LLM, and therefore needs more computation resources. Both prompting techniques have their advantages and disadvantages. Compared to direct prompting, soft prompting does not require the tedious process of manually creating prompts; however, it requires some labeled data to train the prompt. In this work, we explore both direct and soft prompting as we aim to explore the latest developments in LLMs and prompting techniques for the task of extracting PDCM entities from abstracts of academic papers.

**Figure 1.** Illustration of the 2 prompting methods. In direct prompting, a prompt contains a sequence of words. In soft prompting, a prompt consists of a list of vectors. LLM: large language model.



## Direct Prompting

When asking LLMs to extract entities such as diagnoses or biomarkers, the most intuitive way is to ask LLMs to output the entities directly. In example 1 below, “ALK” is a biomarker entity. One may expect the model to output  $\{“biomarker” [ALK]\}$ . However, we note that the string “ALK” is mentioned multiple times in this example text, therefore it is not clear which “ALK” the model refers to. To get the most precise extraction to facilitate a more fine-grained analysis, we instruct the model to output the offsets of the specific mentions in the text (ie, the spans). For instance, if the model gives us  $[(48, 51, “ALK,” biomarker), (323, 326, “ALK,” biomarker), \dots]$ , we know that from character 48 to character 51, there is a biomarker entity, “ALK.” Similarly, we can find another biomarker entity “ALK” at position 323 - 326.

### Example 1:

*Oncogenic fusion of anaplastic lymphoma kinase (ALK) with echinoderm microtubule associated protein like 4 protein or other partner genes occurs in 3 to 6% of lung adenocarcinomas. Although fluorescence in situ hybridization (FISH) is the accepted standard for detecting anaplastic lymphoma receptor tyrosine kinase gene (ALK) gene rearrangement that gives rise to new fusion genes, not all ALK FISH-positive patients respond to ALK inhibitor therapies.*

We started our exploration by designing prompts with an explicit instruction to specify the character offsets of each entity along with the entity text and type (eg, 48, 51, “ALK”, biomarker). However, our experiments show that it was challenging for the LLM to output the correct character offsets, a seemingly straightforward task (all the model needs to do is to count the number of characters); however, the complexity of this seemingly straightforward task is likely due to the LLM’s way of breaking words outside its vocabulary into so-called word pieces, for example, “organoid” is broken down into 2 word pieces “organ” and “-oid.” Considering that LLMs were trained as generative models [3,4], we subsequently cast the entity extraction task as a generation task, where we instructed the model to mark the entities with XML tags. For instance, if the model outputs “Oncogenic fusion of anaplastic lymphoma kinase (<biomarker>ALK</biomarker>) with echinoderm microtubule ...,” then postprocessing the output with regular expressions would find the exact position of “ALK” in the text. Specifically,

we asked the LLMs to mark the start and end of an entity with <entity\_type> and </entity\_type> tags, where entity\_type is a placeholder for the specific entity type, such as biomarker or treatment (refer Table 1 for the full list).

## Soft Prompting

Designing the direct prompts manually could be time-consuming and minor changes in the prompt language could lead to drastic changes in the model performance [24,27]. On the other hand, soft prompting requires some amount of gold data for its training and annotating gold data by domain experts could also be time-consuming. Fortunately, only a small set of labeled data are needed to train soft prompts. As described above, we created a gold dataset, which we used for training and evaluating our soft prompting approach.

There are a few soft prompting methods, the difference usually lies in how the prompt vectors are initialized and learned. Prompt-tuning [18] is a technique that learns the prompt by adding a list of virtual tokens (ie, vectors) in front of the input, where the virtual tokens can be randomly initialized, or drawn from a pretrained word embedding [28] set. Another method is P-tuning [24], which uses small neural networks such as feedforward neural networks [29] (multilayer perceptron) or recurrent neural networks [30] (eg, long-short term memory) as the prompt encoder to learn the prompt. Only the parameters in the prompt encoder are updated during training, while the weights in the LLMs remain frozen. In our experiments, we found P-tuning did not always converge to an optimal solution for our task perhaps due to the random initialization of the vectors rather than using carefully pretrained word embeddings. Therefore, we focused on prompt-tuning in this work. Following Lester et al [18], we initialized the vectors in the prompt with the embeddings of the label words in the entity type set (Table 1).

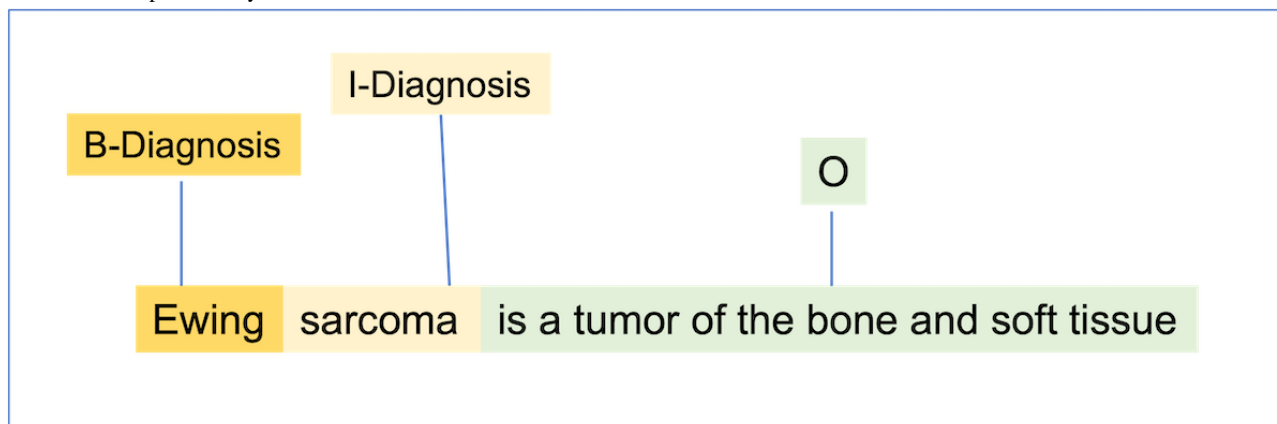
The standard approach for entity extraction in natural language processing is via token classification [31]. Concretely, a classifier is trained to predict the label for each token in a sentence according to a predefined label set. Additionally, each label is prepended with a B or I prefix to indicate the entity’s Beginning or Inside mention, respectively. An example is provided in Figure 2. “Ewing sarcoma” is an entity mention of the diagnosis type. Thus “Ewing” and “sarcoma” are labeled as “Diagnosis,” while all other tokens are labeled as “O,” meaning they are Outside of an entity. To be more precise, “Ewing” is at the beginning of the diagnosis entity, and “sarcoma” is inside



of the entity, so they are labeled as “B-Diagnosis” and “I-Diagnosis,” respectively.

To summarize, we trained a multiclass classifier for the soft-prompting training step. There are 15 entity types in our dataset, therefore there are  $15 \times 2 + 1 = 31$  labels for token classification, with one extra label for “O.”

**Figure 2.** An example of entity extraction as token classification.

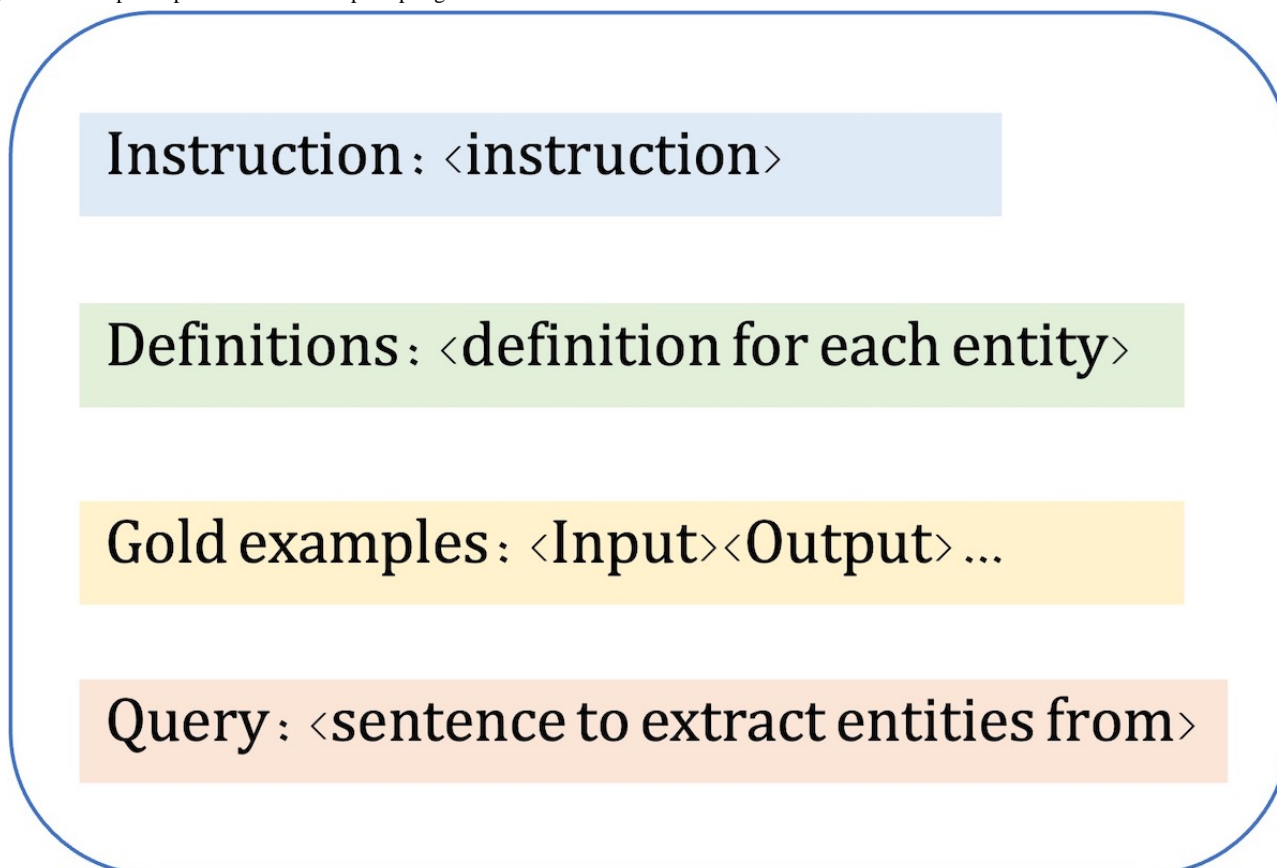


### Experimental Set-Up

For efficiency purposes, we used Apache cTAKES [32] to split an abstract into sentences which were then passed to the LLMs to extract entities one sentence at a time. Our direct prompt included the instruction, the definition of each entity type, 5

examples (few-shot in-context learning) and the query (the sentence). The in-context learning [4] is a common practice in LLM prompting and has consistently shown improved results as the examples guide the LLM onto an optimal path [33,34]. Figure 3 presents our prompt template, and examples are in Multimedia Appendix 1.

**Figure 3.** Prompt template used in direct prompting.



When choosing the LLMs, we used GPT-4o [35], one of the most powerful proprietary LLMs at the time of this study, and SOTA open LLMs from the LLaMA3 family [36], including LLaMA3.1 70B, LLaMA3.1 8B, LLaMA3.2 1B, and LLaMA3.2

3B. We did not use GPT-4o or LLaMA3.1 70B to train the soft prompts due to computational resource limitations; thus, our work here is representative of the computational environment in the vast majority of academic medical centers and research

labs. We set the soft prompt length to 30. We trained the soft prompt on the training set for 50 epochs with a learning rate of 0.001. Hyperparameters were tuned on the development set using the LLaMA3.1 8B model.

We report the evaluation results on the test set in the next section. In addition, we apply 5-fold cross-validation and report the average scores with SDs. For the 5-fold cross-validation, we excluded the 3 abstracts used to sample the gold examples for direct prompting and split the remaining 97 abstracts into 5 folds with a 20:20:20:17 ratio. For direct prompting, we ran the model on each fold and reported the average scores. For soft prompting, we set aside one fold as the test set and trained the soft prompts on the remaining 4 folds.

Results

We used the standard evaluation metrics of precision or positive predictive value, recall or sensitivity, and  $F_1$ -score (the harmonic mean of precision and recall) with 2 evaluation settings: “exact match” setting requires the span output from the model to exactly match the span of the gold annotation, and “overlapping

match” setting allows the model to get partial credit if its extraction overlaps the spans in the gold annotation. For example, the model may extract “patient-derived tumor xenograft (PDX)” as a model\_type entity, while the gold annotation is “patient-derived tumor xenograft (PDX) models.” Under the “exact match” setting, “patient-derived tumor xenograft (PDX)” is NOT a match to “patient-derived tumor xenograft (PDX) models;” while under the “overlapping match” setting, it is a match since the spans overlap.

Tables 3 and 4 show the evaluation results on the test set. In Table 3, we can see that under the “exact match” setting, GPT-4o direct prompting achieves the highest  $F_1$ -score of 50.48. The performances of the LLaMA3 family models drop as the model size decreases, with  $F_1$ -score from 38.40 for the 70B model to 6.78 for the 1B model. However, there is a consistent improvement in  $F_1$ -scores with soft prompting over direct prompting. For the LLaMA3.2 models, the performance of the 3B model improves significantly, with  $F_1$ -score rising from 7.06 to 46.68  $F_1$ -score—more than 8 points higher than the LLaMA3.1-70B model with direct prompting ( $F_1$ -score=38.40), despite the substantial difference in model size.

**Table .** Evaluation results on the test set (exact match) as precision or positive predictive value, recall or sensitivity, and  $F_1$ -score (harmonic mean of precision and recall).

Exact match	Precision	Recall	$F_1$ -score
Direct prompting			
GPT-4o	56.09	45.89 <sup>a</sup>	50.48 <sup>a</sup>
LLaMA3.1-70B	57.27 <sup>a</sup>	28.89	38.40
LLaMA3.1-8B	35.80	18.48	24.37
LLaMA3.2-3B	25.23	4.10	7.06
LLaMA3.2-1B	23.48	3.96	6.78
Soft prompting			
LLaMA3.1-8B	47.17	45.75	46.44
LLaMA3.2-3B	47.30 <sup>a</sup>	46.09 <sup>a</sup>	46.68 <sup>a</sup>
LLaMA3.2-1B	46.19	45.01	45.59

<sup>a</sup>These are the best results.

**Table .** Evaluation results on the test set (overlapping match) as precision or positive predictive value, recall or sensitivity, and  $F_1$ -score (harmonic mean of precision and recall).

Overlapping match	Precision	Recall	$F_1$ -score
Direct prompting			
GPT-4o	76.96	66.52 <sup>a</sup>	71.36 <sup>a</sup>
LLaMA3.1-70B	77.95 <sup>a</sup>	43.99	56.24
LLaMA3.1-8B	50.54	27.49	35.61
LLaMA3.2-3B	41.03	7.03	12.00
LLaMA3.2-1B	35.34	6.01	10.28
Soft prompting			
LLaMA3.1-8B	71.19	70.53	70.86
LLaMA3.2-3B	72.05 <sup>a</sup>	71.55 <sup>a</sup>	71.80 <sup>a</sup>
LLaMA3.2-1B	70.38	70.48	70.42

<sup>a</sup>These are the best results.

Similar trends are observed in Table 4 under the “overlapping match” evaluation. GPT4-o shows an  $F_1$ -score performance of 71.36, maintaining its position as the top performer for direct prompting. The 3 smaller LLaMA3 models continue to benefit from soft prompting, with the LLaMA3.2 3B model achieving slightly higher score than GPT4-o with direct prompting ( $F_1$ -scores of 71.80 vs 71.36 ).

Tables 5 and 6 present the results with 5-fold cross-validation under “exact match” and “overlapping” match respectively. Once again, our observations indicate that with soft prompting, the smaller LLaMA models attain performance levels comparable to GPT-4o.

**Table .** Five-fold cross-validation results (exact match) as precision or positive predictive value, recall or sensitivity, and  $F_1$ -score (harmonic mean of precision and recall).

Exact match	Precision	Recall	$F_1$ -score
Direct prompting, mean (SD)			
GPT-4o	60.73 (2.69)	49.92 (3.46)	54.75 (2.84)
LLaMA3.1-70B	57.56 (1.53)	31.70 (1.24)	40.87 (1.25)
LLaMA3.1-8B	38.29 (3.29)	20.57 (2.18)	26.75 (2.61)
LLaMA3.2-3B	27.01 (3.20)	5.25 (0.80)	8.80 (1.29)
LLaMA3.2-1B	9.84 (5.98)	0.74 (0.47)	1.38 (0.87)
Soft prompting, mean (SD)			
LLaMA3.1-8B	51.76 (3.09)	50.21 (2.24)	50.94 (2.55)
LLaMA3.2-3B	50.99 (2.43)	49.54 (2.98)	50.24 (2.53)
LLaMA3.2-1B	49.34 (3.47)	49.98 (3.19)	49.13 (3.10)

**Table .** Five-fold cross-validation results (overlapping match) as precision or positive predictive value, recall or sensitivity, and  $F_1$ -score (harmonic mean of precision and recall).

Overlapping match	Precision	Recall	$F_1$ -score
Direct prompting, mean (SD)			
GPT-4o	77.82 (2.54)	67.52 (2.17)	72.28 (1.88)
LLaMA3.1-70B	78.01 (1.14)	47.77 (0.71)	59.25 (0.81)
LLaMA3.1-8B	52.75 (3.02)	29.78 (2.60)	38.04 (2.84)
LLaMA3.2-3B	42.42 (2.89)	8.64 (1.09)	14.34 (1.64)
LLaMA3.2-1B	22.09 (5.74)	1.67 (0.54)	3.10 (0.99)
Soft prompting, mean (SD)			
LLaMA3.1-8B	73.78 (3.09)	73.77 (1.25)	73.75 (2.06)
LLaMA3.2-3B	73.48 (1.97)	73.51 (1.11)	73.48 (1.31)
LLaMA3.2-1B	71.51 (3.43)	73.25 (2.46)	72.34 (2.63)

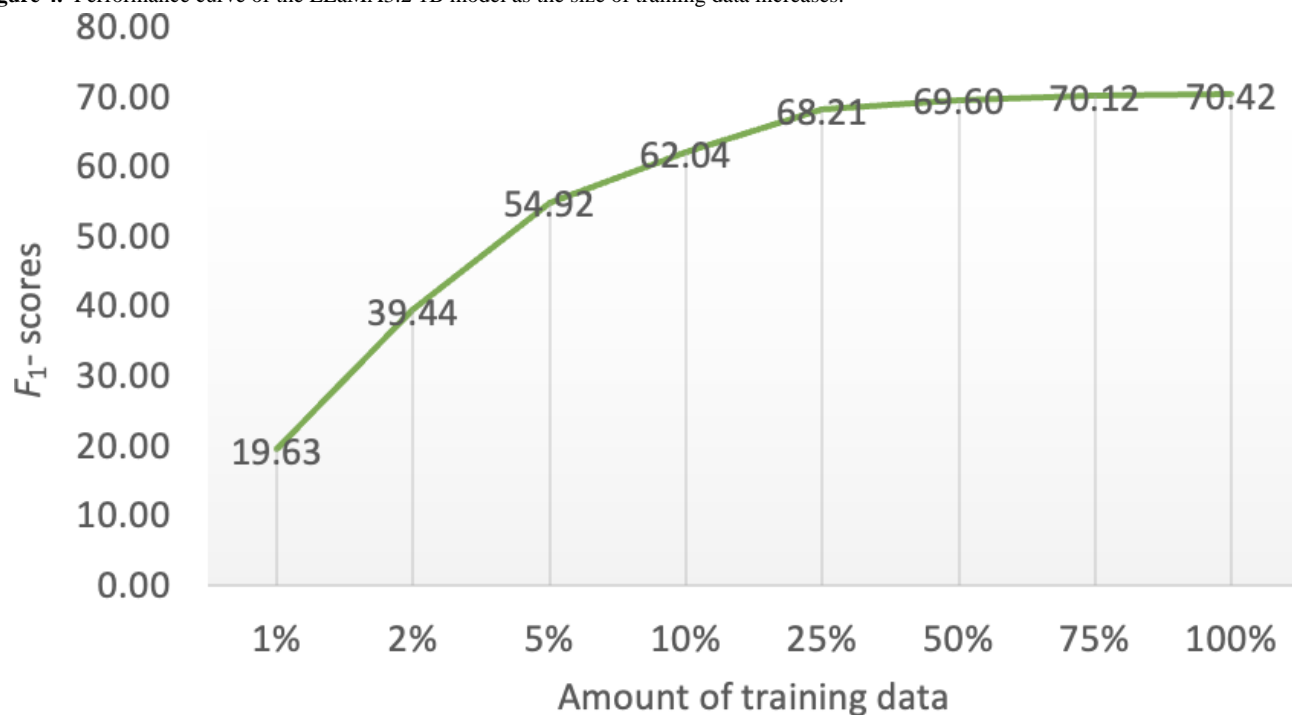
Discussion

Principal Findings

Our experiments demonstrate that soft prompting, a relatively underexplored aspect of LLM prompting, can significantly enhance the performance of smaller LLMs. The 3 LLaMA models exhibit comparable performance under soft prompting (an  $F_1$ -score of 46 in the exact match setting, and 70 in the overlapping match setting). These results are particularly promising results given the limited training data, consisting of 60 abstracts with 2089 entity mentions. Please note that all  $F_1$ -scores mentioned in this section refer to the  $F_1$ -scores on the test set.

How much data is needed to train the soft prompt? To answer this question, we trained LLaMA3.2 1B model, the smallest model used in this work, with different amounts of training data.

Figure 4 shows the relation between the proportion of training data and the  $F_1$ -scores on the test set (overlapping match). Solid performance was achieved with only 5% of the training data (26 sentences from 3 abstracts). With 25% of the training data (129 sentences from 15 abstracts), the model achieved an  $F_1$ -score of 68.21, only 2 points lower than using the entire training set, and only 3 points lower than GPT4-o with direct prompting. Despite the impressive performance of GPT4-o direct prompting, one potential issue is that not all data used in biomedical research can be sent to proprietary models such as GPT or the Gemini family models [8] via public application programming interfaces. That is, for applications using real patient data that require Health Insurance Portability and Accountability Act-compliant platforms, our findings demonstrate that achieving performance comparable to proprietary LLMs such as GPT4-o remains feasible through soft prompting. However, this approach necessitates a tradeoff, requiring a small set of labeled data for optimal effectiveness.

**Figure 4.** Performance curve of the LLaMA3.2 1B model as the size of training data increases.

Some entities appear more frequently than other entities in our dataset. For example, diagnosis and treatment mentions are more frequent than mentions of cancer\_grade. In Table 7, we present the number of instances of each entity type in our dataset and the corresponding performance of GPT4-o direct prompting. We can see that GPT4-o performs the best for the entity types that have the most instances—diagnosis, model type, and treatment entities. Of these frequent entity types, biomarker is the one with the lowest performance. Our error analysis points to several factors that could have contributed to these results, including ambiguous and inconsistent mentions and contextual dependencies. In this task, we defined a biomarker as “gene, protein or biological molecule identified in or associated with patient’s/model’s tumor.” Thus, biomarker entities can be mentioned using their full names (eg, epidermal growth factor

receptor, Inc-RP11-536 K7.3, echinoderm microtubule-associated protein-like 4), standardized gene or protein symbols (*NPM1*, KRAS, PTEN) or abbreviations of metabolites (NADPH, D2HG). Moreover, a biomarker entity (eg, “MEK”) often overlaps with a treatment entity (eg, “MEK inhibitor”). The ambiguity in biomarker entity mentions might interfere with the model’s ability to recognize them consistently. In addition, biomarker entities are often mentioned as lists (see Example 2) resulting in a different frequency within and across the abstracts and patterns of entity mentions, in comparison with other entities. Overall, ambiguity emerges as the primary source of error. More precise definitions, accompanied by examples illustrating the distinct meanings, might present a solution. Table S2 in Multimedia Appendix 1 provides the breakdown of errors per entity type along with examples.



**Table .** Evaluation results of GPT4-o with direct prompts for each entity type as precision or positive predictive value, recall or sensitivity, and  $F_1$ -score (harmonic mean of precision and recall). Results are overlapping match setting on the test set.

Entity type	Training instances, n	Development instances, n	Test instances, n	Precision	Recall	$F_1$ -score	IAA <sup>a</sup>
diagnosis	362	122	114	92.47	75.44	83.09 <sup>b</sup>	61.67
age_category	19	0	0	0.0	0.0	0.0	60
genetic_effect	69	20	33	45.71	47.06	46.38	57.67
model_type	326	114	110	88.07	84.21	86.10 <sup>b</sup>	53.33
molecular_char	128	37	46	65.22	63.83	64.52 <sup>b</sup>	54.33
biomarker	503	118	163	85.05	55.49	67.16 <sup>b</sup>	61.33
treatment	426	77	130	81.74	70.15	75.50 <sup>b</sup>	55.67
response_to_treatment	99	21	28	38.64	60.71	47.22	55
sample_type	22	8	7	45.45	71.43	55.56 <sup>b</sup>	49
tumor_type	61	19	28	66.67	57.14	61.54 <sup>b</sup>	49.67
cancer_grade	6	1	1	50.0	100	66.67 <sup>b</sup>	42
cancer_stage	7	1	4	33.33	25.0	28.57	59.33
clinical_trial	35	2	4	80.0	100	88.89 <sup>b</sup>	60.67
host_strain	9	0	7	100	28.57	44.44	61.67
model_id	17	2	7	66.67	28.57	40.0	100

<sup>a</sup>IAA: interannotator agreement.  
<sup>b</sup> $F_1$ -scores exceeding interannotator agreement.

*Example 2:*  
*Genomic alterations involved RB1 (55%), TP53 (46%), PTEN (29%), BRCA2 (29%), and AR (27%), and there was a range of androgen receptor signaling and NEPC marker expression.*

The moderate performance of entity types such as genetic\_effect, molecular\_char, and response\_to\_treatment, and tumour\_type is due to the number of training instances ranging from 61 to 128 as well as the IAA ranging from 49.67 to 57.67. The moderate IAA scores of those entity types underscore the need for refined annotation protocols and modeling strategies that better capture domain-specific knowledge. Furthermore, the lower performance observed for entity types with smaller sample sizes (eg, model\_id) highlights the need for enhancing model performance on low-frequency labels. Future research could

explore strategies such as data augmentation to improve the model’s generalizability for underrepresented entities.

The extraction of PDCM-relevant knowledge is not an easy task for the domain experts as indicated by the IAA ( $F_1$ -score below 65 for all entity types except for model\_id). In 9 out of 15 entity types, the system performance in an overlapping match setting exceeds the IAA (last two columns of Table 7). This is the case for categories with plentiful training instances (eg, diagnosis, model\_type) as well as for categories with fewer training instances (eg, sample\_type, cancer\_grade). For the exact match setting, in 6 out of 15 entity types, the system performance exceeds the IAA (last two columns in Table 8). Therefore, the LLM could be a viable assistant, with its outputs reviewed by a domain expert to ensure the accuracy of the finalextraction. We believe such human-in-the-loop approaches present a promising direction for future research and application.



**Table .** Evaluation results of GPT4-o with direct prompts for each entity type as precision or positive predictive value, recall or sensitivity, and  $F_1$ -score (harmonic mean of precision and recall). Results are exact match setting on the test set.

Entity type	Training in- stances, n	Development in- stances, n	Test instances, n	Precision	Recall	$F_1$ -score	IAA <sup>a</sup>
diagnosis	362	122	114	77.17	62.28	68.93 <sup>b</sup>	61.67
age_category	19	0	0	0.0	0.0	0.0	60.0
genetic_effect	69	20	33	25.71	27.27	26.47	57.67
model_type	326	114	110	56.88	56.36	56.62 <sup>b</sup>	53.33
molecular_char	128	37	46	54.35	54.35	54.35 <sup>b</sup>	54.33
biomarker	503	118	163	46.74	26.38	33.73	61.33
treatment	426	77	130	72.34	52.31	60.71 <sup>b</sup>	55.67
response_to _treatment	99	21	28	27.91	42.86	33.80	55
sample_type	22	8	7	45.45	71.43	55.56 <sup>b</sup>	49
tumor_type	61	19	28	50.0	39.29	44.0	49.67
cancer_grade	6	1	1	50.0	100	66.67 <sup>b</sup>	42
cancer_stage	7	1	4	33.33	25.0	28.57	59.33
clinical_trial	35	2	4	40.0	50.0	44.44	60.67
host_strain	9	0	7	100	14.29	25.0	61.67
model_id	17	2	7	66.67	28.27	40.0	100

<sup>a</sup>IAA: interannotator agreement.

<sup>b</sup> $F_1$ -scores exceeding the interannotator agreement.

We would like to note that the work presented in the paper was done in a computational environment representative of the vast majority of academic medical centers and nonindustry research labs. Although we have access to SOTA Graphics Processing Units, we still found ourselves constrained as to the extent to which we could use very large language models. The larger community needs to address the growing gap in computational resources between big tech and the rest of the research community.

Limitations

As this is a feasibility study, we limited ourselves to the extraction of entity mentions of 15 entity types chosen from attributes in the descriptive standards for PDCMs. While these are recognized by the PDCM and oncology community, they do not cover all knowledge in the PDCM-relevant texts. Some refinement of the entity types will be beneficial to improve prompting results.

We limited our corpus to 100 abstracts from papers associated with PDCMs deposited in CancerModels.Org. We did not assess the abstracts for the presence and equal distribution of all the entities. Thus, there were very few mentions of some entities in the corpus (eg, cancer\_stage), negatively affecting our overall  $F_1$ -score. We decided not to exclude those entities as these results could guide refinements of future studies. The computational methods discussed here are applicable to other studies requiring the extraction of textual information from

scientific papers. Future work could involve extending this method to extract knowledge from the main body of the papers.

Conclusions

This study investigates the potential of LLMs as powerful tools for extracting PDCM-relevant knowledge from scientific literature—an essential task for advancing cancer research and precision medicine. By comparing direct and soft prompting across both proprietary and open LLMs, we provide valuable insights into the most effective strategies for PDCM-relevant knowledge extraction. Our findings indicate that GPT-4o, when used with direct prompting, maintains competitive performance, while soft prompting significantly enhances the effectiveness of smaller LLMs. In conclusion, our results demonstrate that training soft prompts on smaller open models can achieve performance levels comparable to those of proprietary LLMs.

To our knowledge, this is the first study to implement SOTA LLMs prompting for knowledge extraction in the PDCM domain and the first to explore the emerging topic of soft prompting in this context. Our findings demonstrate that LLMs can effectively streamline the extraction of complex cancer model metadata, potentially reducing the burden of manual curation and accelerating the integration of PDCMs into research and clinical workflows. Additionally, this study lays the foundation for future research aimed at optimizing LLMs for large-scale knowledge extraction tasks. Efficiently extracting and harmonizing PDCM-relevant knowledge will ultimately drive progress in cancer research and precision oncology, equipping researchers and clinicians with better tools to improve patient

outcomes. More broadly, our study contributes to the ongoing discourse on the applicability of LLMs, acknowledging that while they offer transformative potential, they are not a universal solution for all tasks.

## Acknowledgments

Funding was provided by the US National Institutes of Health (U24CA248010, R01LM013486, U24CA253539) and European Bioinformatics Institute (EMBL-EBI) Core Funds.

## Data Availability

The data and code will be available upon publication in the CancerModels.Org Github repository [37].

## Conflicts of Interest

None declared.

## Multimedia Appendix 1

Prompts used in direct prompting experiments and detailed error analysis.

[DOCX File, 20 KB - [bioinform\\_v6i1e70706\\_app1.docx](#)]

## References

1. RePORTER. National Institutes of Health. URL: <https://reporter.nih.gov/> [accessed 2024-12-16]
2. Perova Z, Martinez M, Mandloi T, et al. PDCM Finder: an open global research platform for patient-derived cancer models. *Nucleic Acids Res* 2023 Jan 6;51(D1):D1360-D1366. [doi: [10.1093/nar/gkac1021](#)] [Medline: [36399494](#)]
3. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. arXiv. Preprint posted online on Jun 12, 2017. [doi: [10.48550/arXiv.1706.03762](#)]
4. Brown TB, Mann B, Ryder N, et al. Language models are few-shot learners. arXiv. Preprint posted online on May 28, 2020. [doi: [10.48550/arXiv.2005.14165](#)]
5. OpenAI, Achiam J, Adler S, et al. GPT-4 technical report. arXiv. Preprint posted online on Mar 15, 2023. [doi: [10.48550/arXiv.2303.08774](#)]
6. Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. *N Engl J Med* 2023 Mar 30;388(13):1233-1239. [doi: [10.1056/NEJMs2214184](#)] [Medline: [36988602](#)]
7. Omiye JA, Gui H, Rezaei SJ, Zou J, Daneshjou R. Large language models in medicine: the potentials and pitfalls: a narrative review. *Ann Intern Med* 2024 Feb;177(2):210-220. [doi: [10.7326/M23-2772](#)] [Medline: [38285984](#)]
8. Saab K, Tu T, Weng WH, et al. Capabilities of Gemini models in medicine. arXiv. Preprint posted online on Apr 29, 2024. [doi: [10.48550/arXiv.2404.18416](#)]
9. Kanjee Z, Crowe B, Rodman A. Accuracy of a generative artificial intelligence model in a complex diagnostic challenge. *JAMA* 2023 Jul 3;330(1):78-80. [doi: [10.1001/jama.2023.8288](#)] [Medline: [37318797](#)]
10. Savage T, Nayak A, Gallo R, Rangan E, Chen JH. Diagnostic reasoning prompts reveal the potential for large language model interpretability in medicine. *NPJ Digit Med* 2024 Jan 24;7(1):20. [doi: [10.1038/s41746-024-01010-1](#)] [Medline: [38267608](#)]
11. Williams CYK, Miao BY, Kornblith AE, Butte AJ. Evaluating the use of large language models to provide clinical recommendations in the emergency department. *Nat Commun* 2024 Oct 8;15(1):8236. [doi: [10.1038/s41467-024-52415-1](#)] [Medline: [39379357](#)]
12. Ayers JW, Poliak A, Dredze M, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA Intern Med* 2023 Jun 1;183(6):589-596. [doi: [10.1001/jamainternmed.2023.1838](#)] [Medline: [37115527](#)]
13. Lucas HC, Upperman JS, Robinson JR. A systematic review of large language models and their implications in medical education. *Med Educ* 2024 Nov;58(11):1276-1285. [doi: [10.1111/medu.15402](#)] [Medline: [38639098](#)]
14. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. *PLOS Digit Health* 2023 Feb;2(2):e0000198. [doi: [10.1371/journal.pdig.0000198](#)] [Medline: [36812645](#)]
15. Perot V, Kang K, Luisier F, et al. LMDX: language model-based document information extraction and localization. In: Ku LW, Martins A, Srikumar V, editors. *Findings of the Association for Computational Linguistics ACL 2024: Association for Computational Linguistics*; 2024:15140-15168. [doi: [10.18653/v1/2024.findings-acl.899](#)]
16. Arsenyan V, Bughdaryan S, Shaya F, Small KW, Shahnazaryan D. Large language models for biomedical knowledge graph construction: information extraction from EMR notes. In: Demner-Fushman D, Ananiadou S, Miwa M, Roberts K, Tsujii J, editors. *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing: Association for Computational Linguistics*; 2024:295-317. [doi: [10.18653/v1/2024.bionlp-1.23](#)]

17. Munnangi M, Feldman S, Wallace B, Amir S, Hope T, Naik A. On-the-fly definition augmentation of LLMs for biomedical NER. In: Duh K, Gomez H, Bethard S, editors. Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Association for Computational Linguistics; 2024:3833-3854. [doi: [10.18653/v1/2024.naacl-long.212](https://doi.org/10.18653/v1/2024.naacl-long.212)]
18. Lester B, Al-Rfou R, Constant N. The power of scale for parameter-efficient prompt tuning. In: Moens MF, Huang X, Specia L, Yih SW, editors. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: Association for Computational Linguistics; 2021:3045-3059. [doi: [10.18653/v1/2021.emnlp-main.243](https://doi.org/10.18653/v1/2021.emnlp-main.243)]
19. Meehan TF, Conte N, Goldstein T, et al. PDX-MI: minimal information for patient-derived tumor xenograft models. *Cancer Res* 2017 Nov 1;77(21):e62-e66. [doi: [10.1158/0008-5472.CAN-17-0582](https://doi.org/10.1158/0008-5472.CAN-17-0582)] [Medline: [29092942](https://pubmed.ncbi.nlm.nih.gov/29092942/)]
20. PDCMFinder/MI-standard-in-vitro-models. GitHub. URL: <https://github.com/PDCMFinder/MI-Standard-In-vitro-models> [accessed 2024-12-23]
21. Hripcsak G, Rothschild AS. Agreement, the F-measure, and reliability in information retrieval. *J Am Med Inform Assoc* 2005;12(3):296-298. [doi: [10.1197/jamia.M1733](https://doi.org/10.1197/jamia.M1733)] [Medline: [15684123](https://pubmed.ncbi.nlm.nih.gov/15684123/)]
22. Wei J, Wang X, Schuurmans D, et al. Chain-of-thought prompting elicits reasoning in large language models. arXiv. Preprint posted online on Jan 28, 2022. [doi: [10.48550/arXiv.2201.11903](https://doi.org/10.48550/arXiv.2201.11903)]
23. Wang X, Wei J, Schuurmans D, et al. Self-consistency improves chain of thought reasoning in language models. arXiv. Preprint posted online on Mar 21, 2022. [doi: [10.48550/arXiv.2203.11171](https://doi.org/10.48550/arXiv.2203.11171)]
24. Liu X, Zheng Y, Du Z, et al. GPT understands, too. arXiv. Preprint posted online on Mar 18, 2021. [doi: [10.48550/arXiv.2103.10385](https://doi.org/10.48550/arXiv.2103.10385)]
25. Schulhoff S, Ilie M, Balepur N, et al. The prompt report: a systematic survey of prompting techniques. arXiv. Preprint posted online on Jun 6, 2024. [doi: [10.48550/arXiv.2406.06608](https://doi.org/10.48550/arXiv.2406.06608)]
26. Li XL, Liang P. Prefix-tuning: optimizing continuous prompts for generation. arXiv. Preprint posted online on Jan 1, 2021. [doi: [10.48550/arXiv.2101.00190](https://doi.org/10.48550/arXiv.2101.00190)]
27. Zhou Y, Muresanu A I, Han Z, Paster K, Pitis S, Chan H. Large language models are human-level prompt engineers. arXiv. Preprint posted online on Nov 3, 2022. [doi: [10.48550/arXiv.2211.01910](https://doi.org/10.48550/arXiv.2211.01910)]
28. Mikolov T, Sutskever I, Chen K, Corrado G, Dean J. Distributed representations of words and phrases and their compositionality. arXiv. Preprint posted online on Oct 16, 2013. [doi: [10.48550/arXiv.1310.4546](https://doi.org/10.48550/arXiv.1310.4546)]
29. Rosenblatt F. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychol Rev* 1958 Nov;65(6):386-408. [doi: [10.1037/h0042519](https://doi.org/10.1037/h0042519)] [Medline: [13602029](https://pubmed.ncbi.nlm.nih.gov/13602029/)]
30. Hochreiter S, Schmidhuber J. Long short-term memory. *Neural Comput* 1997 Nov 15;9(8):1735-1780. [doi: [10.1162/neco.1997.9.8.1735](https://doi.org/10.1162/neco.1997.9.8.1735)] [Medline: [9377276](https://pubmed.ncbi.nlm.nih.gov/9377276/)]
31. Tjong Kim Sang EF, De Meulder F. Introduction to the CoNLL-2003 shared task: language-independent named entity recognition. Presented at: Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL; May 31, 2003; Edmonton, Canada. [doi: [10.3115/1119176.1119195](https://doi.org/10.3115/1119176.1119195)]
32. Savova GK, Masanz JJ, Ogren PV, et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *J Am Med Inform Assoc* 2010;17(5):507-513. [doi: [10.1136/jamia.2009.001560](https://doi.org/10.1136/jamia.2009.001560)] [Medline: [20819853](https://pubmed.ncbi.nlm.nih.gov/20819853/)]
33. von Oswald J, Niklasson E, Randazzo E, et al. Transformers learn in-context by gradient descent. arXiv. Preprint posted online on Dec 15, 2022. [doi: [10.48550/arXiv.2212.07677](https://doi.org/10.48550/arXiv.2212.07677)]
34. Hendel R, Geva M, Globerson A. In-context learning creates task vectors. In: Bouamor H, Pino J, Bali K, editors. Findings of the Association for Computational Linguistics: EMNLP 2023: Association for Computational Linguistics; 2023:9318-9333. [doi: [10.18653/v1/2023.findings-emnlp.624](https://doi.org/10.18653/v1/2023.findings-emnlp.624)]
35. OpenAI, Hurst A, Lerer A, et al. GPT-4o system card. arXiv. Preprint posted online on Oct 25, 2024. [doi: [10.48550/arXiv.2410.21276](https://doi.org/10.48550/arXiv.2410.21276)]
36. Grattafiori A, Dubey A, Jauhri A, et al. The Llama 3 herd of models. arXiv. Preprint posted online on Jul 31, 2024. [doi: [10.48550/arXiv.2407.21783](https://doi.org/10.48550/arXiv.2407.21783)]
37. PDCMFinder/prompt-llm. GitHub. URL: <https://github.com/PDCMFinder/prompt-llm> [accessed 2024-12-23]

## Abbreviations

**IAA:** interannotator agreement  
**LLM:** large language model  
**PDCM:** patient-derived cancer model  
**PDX:** patient-derived xenografts  
**SOTA:** state-of-the-art

*Edited by J Finkelstein; submitted 30.12.24; peer-reviewed by P Dadheech, S Eger, Z Chen; revised version received 14.04.25; accepted 27.04.25; published 30.06.25.*

*Please cite as:*

*Yao J, Perova Z, Mandloi T, Lewis E, Parkinson H, Savova G*

*Extracting Knowledge From Scientific Texts on Patient-Derived Cancer Models Using Large Language Models: Algorithm Development and Validation Study*

*JMIR Bioinform Biotech 2025;6:e70706*

URL: <https://bioinform.jmir.org/2025/1/e70706>

doi: [10.2196/70706](https://doi.org/10.2196/70706)

© Jiarui Yao, Zinaida Perova, Tushar Mandloi, Elizabeth Lewis, Helen Parkinson, Guergana Savova. Originally published in JMIR Bioinformatics and Biotechnology (<https://bioinform.jmir.org>), 30.6.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Bioinformatics and Biotechnology, is properly cited. The complete bibliographic information, a link to the original publication on <https://bioinform.jmir.org/>, as well as this copyright and license information must be included.



Original Paper

# A Hybrid Deep Learning–Based Feature Selection Approach for Supporting Early Detection of Long-Term Behavioral Outcomes in Survivors of Cancer: Cross-Sectional Study

Tracy Huang<sup>1\*</sup>, BA; Chun-Kit Ngan<sup>2\*</sup>, BA, PhD; Yin Ting Cheung<sup>3</sup>, PhD; Madelyn Marcotte<sup>2</sup>, BSc; Benjamin Cabrera<sup>4</sup>, BSc

<sup>1</sup>Emory University, Atlanta, GA, United States

<sup>2</sup>Worcester Polytechnic Institute, Worcester, MA, United States

<sup>3</sup>Chinese University of Hong Kong, Hong Kong, China (Hong Kong)

<sup>4</sup>Arizona State University, Tempe, AZ, United States

\*these authors contributed equally

**Corresponding Author:**

Chun-Kit Ngan, BA, PhD

Worcester Polytechnic Institute

100 Institute Rd

Worcester, MA, 01609

United States

Phone: 1 (508) 831 5000

Email: [cngan@wpi.edu](mailto:cngan@wpi.edu)

## Abstract

**Background:** The number of survivors of cancer is growing, and they often experience negative long-term behavioral outcomes due to cancer treatments. There is a need for better computational methods to handle and predict these outcomes so that physicians and health care providers can implement preventive treatments.

**Objective:** This study aimed to create a new feature selection algorithm to improve the performance of machine learning classifiers to predict negative long-term behavioral outcomes in survivors of cancer.

**Methods:** We devised a hybrid deep learning–based feature selection approach to support early detection of negative long-term behavioral outcomes in survivors of cancer. Within a data-driven, clinical domain–guided framework to select the best set of features among cancer treatments, chronic health conditions, and socioenvironmental factors, we developed a 2-stage feature selection algorithm, that is, a multimetric, majority-voting filter and a deep dropout neural network, to dynamically and automatically select the best set of features for each behavioral outcome. We also conducted an experimental case study on existing study data with 102 survivors of acute lymphoblastic leukemia (aged 15-39 years at evaluation and >5 years postcancer diagnosis) who were treated in a public hospital in Hong Kong. Finally, we designed and implemented radial charts to illustrate the significance of the selected features on each behavioral outcome to support clinical professionals' future treatment and diagnoses.

**Results:** In this pilot study, we demonstrated that our approach outperforms the traditional statistical and computation methods, including linear and nonlinear feature selectors, for the addressed top-priority behavioral outcomes. Our approach holistically has higher  $F_1$ , precision, and recall scores compared to existing feature selection methods. The models in this study select several significant clinical and socioenvironmental variables as risk factors associated with the development of behavioral problems in young survivors of acute lymphoblastic leukemia.

**Conclusions:** Our novel feature selection algorithm has the potential to improve machine learning classifiers' capability to predict adverse long-term behavioral outcomes in survivors of cancer.

(JMIR Bioinform Biotech 2025;6:e65001) doi:[10.2196/65001](https://doi.org/10.2196/65001)

**KEYWORDS**

machine learning; data driven; clinical domain–guided framework; survivors of cancer; cancer; oncology; behavioral outcome predictions; behavioral study; behavioral outcomes; feature selection; deep learning; neural network; hybrid; prediction; predictive modeling; patients with cancer; deep learning models; leukemia; computational study; computational biology

## Introduction

### Background

The number of survivors of cancer is increasing globally. The American Cancer Society recently reported that in 2023, a total of 1,958,310 new cancer cases were projected to occur in the United States [1]. Treatment advances have resulted in a dramatic improvement in the survival rates of most cancers, especially in resource-limited countries and regions. However, this growing population of survivors of cancer may develop a myriad of treatment-related adverse effects that lead to a compromised health status. Studies have also shown that survivors of cancer are more likely than the general population to experience negative long-term behavioral outcomes, such as anxiety, depression, attention problems, and sluggish cognitive tempo, after cancer treatments [2]. Contemporary treatment strategies have led to improved life expectancy after treatment for pediatric cancer, especially in survivors of acute lymphocytic leukemia (ALL) [3]. Given that studies have shown that the promotion of a healthy lifestyle and interventions that reduce physical and mental health burdens can lead to reduction in all-cause and cause-specific mortality, addressing the risk factors of adverse functional outcomes early on is critical [4-6]. Thus, developing an effective approach to identify crucial factors and then detect these negative outcomes in advance is needed so that medical therapists can intervene early and take the appropriate actions and treatments promptly to mitigate adverse effects in survivors of cancer.

### Current Approaches for Detecting Adverse Behavioral Outcomes in Survivors of Cancer

Currently, to support the identification of relevant factors and the early detection of adverse behavioral outcomes for survivors of cancer, clinical scientists use various statistical analyses to understand the relationship among those behavioral outcomes, cancer treatments, chronic health conditions, and socioenvironmental factors [7-9]. Specifically, traditional statistical methods (ie, linear regression analysis) are used to extract predictor variables and then model the relationship between the extracted predictor variables and the behavioral outcomes. This analysis assumes that the behavioral outcomes are, for the most part, linearly correlated with those predictor variables. However, this assumption may not always hold in this complex and dynamic problem. Furthermore, the predictors for those behavioral outcomes extracted by statistical methods may have weak prediction accuracy, as modeling human behavioral outcomes is challenging due to its multifactorial nature (ie, many predictors as well as interactions among the predictors affecting the outcome), heterogeneity (ie, differences across individuals), nonlinearity of data, multicollinearity (ie, highly correlated variables), class imbalance (ie, few observations of the outcome of interest), and missing data [10,11]. As a result, this class of linear regressors can only account for a small proportion of variance, with limited usability in a clinical setting. Thus, developing an effective computational methodology that can maximize the use of those data for prognostic and predictive behavioral outcomes is highly desirable.

To address the abovementioned problems, feature selection techniques in machine learning (ML) play an important role. Feature selection techniques can be broadly divided into 4 categories: filter, wrapper, embedded, and hybrid. Filter methods select features based on their statistical significance to the outcome of interest. Unlike other feature selection methods, such as wrapper and embedded methods, filter methods function independent of any ML classifiers. However, filter methods are less accurate than other methods of feature selection, such as wrapper methods. In addition, there is a risk of selecting redundant features when using filter methods that do not consider the correlation between features. Wrapper methods use a greedy search algorithm (ie, an iterative algorithm that makes the locally optimal choice at each step) with a classifier to sequentially add and remove features from the classifier to maximize the specified scoring metrics, that is, precision, recall, and  $F_1$ -score. The output is the best subset of features that the algorithm found. While wrapper methods are proficient in achieving high classification accuracy, they are not efficient in computation time or complexity. In addition, there is also a risk of overfitting with wrapper methods, where the classifier is highly trained to generate accurate predictions for the training data only and cannot correctly create generalized predictions for testing data or any novel datasets. Embedded methods use qualities from both filter and wrapper methods to perform feature selection during the construction of the ML classifiers. The baseline embedded methods that are commonly used are least absolute shrinkage and selection operator (Lasso), Ridge, and ElasticNet. However, to effectively use embedded methods, prior knowledge of the feature sets is required. In addition, embedded methods could pose problems when identifying small feature sets. Hybrid methods combine filter and wrapper methods to take advantage of the benefits each method provides, while minimizing their limitations [12]. A filter method first selects a subset of features, which are then input into a wrapper method to further select the best subset of features. As hybrid methods are a combination of filter and wrapper methods, they inherit problems from both—filter methods may exclude important features and wrapper methods are inefficient in computation time.

### Goal of This Study

To bridge the abovementioned gaps, we propose a hybrid deep learning-based feature selection approach to support early detection of long-term adverse behavioral outcomes in survivors of cancer. Specifically, our goals are four-fold: (1) devise a data-driven, clinical domain-guided framework to select the best set of features among cancer treatments, chronic health conditions, socioenvironmental factors, and others; (2) develop a 2-stage feature selection algorithm, that is, a multimetric, majority-voting filter and a deep dropout neural network (DDN), to dynamically and automatically select the best set of features for each behavioral outcome; (3) conduct an experimental case study on our existing study data with 102 survivors of ALL (aged 15-39 years at evaluation and >5 years postcancer diagnosis) who were treated in a public hospital in Hong Kong; and (4) design and implement radial charts to illustrate the significance of the selected features on each behavioral outcome to support clinical professionals' future treatment and diagnoses.

In this pilot study, we demonstrate that our approach outperforms the traditional statistical and computation methods, including linear and nonlinear feature selectors, for the addressed top-priority behavioral outcomes.

## Methods

### Review of Baseline Feature Selection Methods

#### Overview

Four baseline feature selection methods were used in the experimental studies as a comparison for our novel feature selection algorithm (Textbox 1).

Textbox 1. Summary of the baseline feature selection methods.

<b>Filter</b> <ul style="list-style-type: none"><li>Correlation-based feature selection (CFS)</li><li>Information gain (IG)</li><li>Maximum relevance minimum redundancy (MRMR)</li></ul>
<b>Wrapper</b> <ul style="list-style-type: none"><li>Sequential forward selection (SFS)</li><li>Sequential backwards selection (SBS)</li><li>Stepwise selection (SS)</li></ul>
<b>Embedded</b> <ul style="list-style-type: none"><li>Least absolute shrinkage and selection operator (Lasso)</li><li>Ridge</li><li>ElasticNet</li></ul>
<b>Hybrid</b> <ul style="list-style-type: none"><li>CFS→SFS</li><li>IG→SFS</li><li>MRMR→SFS</li><li>CFS→SFS</li><li>IG→SFS</li><li>MRMR→SFS</li><li>CFS→SBS</li><li>IG→SBS</li><li>MRMR→SBS</li><li>CFS→SS</li><li>IG→SS</li><li>MRMR→SS</li></ul>

#### Filter Methods

Filter methods select features based on their statistical significance to the outcome of interest, independent of any ML classifiers. To evaluate the performance of existing filter methods, we use information gain (IG), maximum relevance minimum redundancy (MRMR), and correlation-based feature selection (CFS) [13]. IG is calculated by comparing the entropy of the dataset before and after a transformation. When IG is used for feature selection, it is called mutual information and works by evaluating the IG of each variable in the context of the target. The MRMR algorithm selects the best *K* features at

each iteration that have maximum relevance with respect to the target variable and minimum redundancy with respect to the other features. The CFS algorithm involves splitting the features into subsets based on whether their values are continuous or discrete and can be used to measure the correlation between features and the target outcomes. For continuous data, Pearson correlation can be used, and for discrete data, symmetrical uncertainty can be used. Symmetrical uncertainty is a measure of relevance between features and targets that uses mutual information [14]. When evaluating the performance of the existing filter methods, we selected the top 15 features that had the highest scores for each of the 3 approaches.

## Wrapper Methods

For binary classification, wrapper methods use a greedy search algorithm with a classifier to sequentially add and remove features from the classifier to maximize the specified scoring metric, that is, precision, recall, and  $F_1$ -score. The output is the best subset of features that the algorithm found. To evaluate existing wrapper methods' performances, we selected 3 commonly implemented wrapper methods: sequential forward selection (SFS), sequential backward selection (SBS), and stepwise selection (SS). SFS starts with an empty subset of features and iteratively adds features if adding them improves the specified score, according to the ML classifier. The selection terminates when a feature subset of the desired size  $k$ , where  $k$  refers to the number of features expected by the domain experts, is reached. In contrast, SBS starts with a full subset of all the features and iteratively removes features if removing them increases the specified score, according to the classifier. The selection also terminates when a feature subset of the desired size  $k$  is reached. SS, also known as bidirectional selection, alternates between forward and backward selection to select the best subset of features. To implement the wrapper selection approaches, we used the support vector machine classifier and used accuracy as the default scoring metric [15]. We also specified that the selection process should terminate when a feature subset of size 15 is reached. For the purpose of the study, we decided a priori that the feature subset should be limited to 15 because if there are too many exploratory factors in the model, the contribution of each factor to the variance may be too small and its clinical significance may be questionable.

## Embedded Methods

Embedded methods use qualities from both filter and wrapper methods to perform feature selection during the construction of the ML classifier. The embedded classifiers we used were Lasso, Ridge, and ElasticNet. Lasso regression is a form of

linear regression that imposes an L1 regularization penalty to identify the features that minimize the prediction error [16]. Similar to Lasso, Ridge regression is another form of linear regression that uses an L2 penalty instead [17]. ElasticNet regression merges Lasso and Ridge regression using the L1 and L2 regularization penalties [18]. ElasticNet regression can shrink some features to zero, similar to Lasso, while reducing the magnitude of other features, like Ridge. For each evaluated embedded method, we selected the top 15 most relevant features for each behavioral outcome.

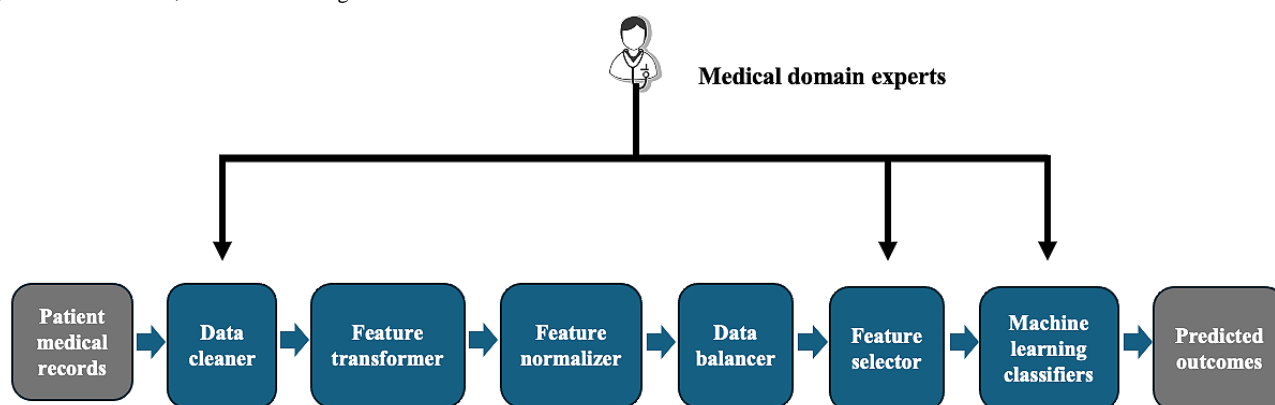
## Hybrid Methods

Hybrid methods combine filter and wrapper methods to take advantage of the benefits each method provides, while minimizing their limitations [12]. We implemented 9 different hybrid methods using the top 30 features selected from the 3 filter methods (ie, CFS, IG, and MRMR) and inputting them each into the 3 wrapper methods, including SFS, SBS, and SS, to subsequently select the top 15 features.

## Data-Driven, Clinical Domain–Guided Framework

In this section, we describe and explain our framework that consisted of 6 main modules (Figure 1). The cancer survivor medical records, including the features, such as biomarkers, chronic health conditions, and socioeconomic factors, were first passed into the data cleaner that “sanitizes” the records with the clinical domain knowledge from our investigators. Note that throughout the framework, our clinical domain experts assisted us with certain processes. In this case study, for example, it consisted of replacing missing values in a patient’s record by averaging the existing values of the corresponding feature among all the other patients’ records grouped by a specific cancer type, age range, and biological sex. Clinical domain experts also helped us interpret and explain what different variable values mean for us to properly transform them into the correct variables.

**Figure 1.** Data-driven, clinical domain–guided framework.



Afterward, the records were passed into the feature transformer, where the one-hot encoding technique was used to transform categorical variables into binary ones [19]. For instance, we transformed the “gender” variable from categorical to binary by replacing “M” and “F” with 1 and 0.

Following feature transformation, the records were normalized by the feature normalizer. The Shapiro-Wilk test, the Kolmogorov-Smirnov test, and the D’Agostino-Pearson test

were used to check whether features follow a normal distribution. If 2 out of the 3 tests conclude that a feature follows a normal distribution, it is standardized by removing the mean and scaling to unit variance [20–22]. Otherwise, features are normalized using the minimum-maximum normalization technique so that all features have values between “0” and “1.” This eliminates any feature bias, where features with high values are given more importance than features with low values [23].



Once the records are cleaned, transformed, and normalized, they are then passed into the data balancer. At this point, the results differ depending on the behavioral outcome being predicted. The synthetic minority oversampling technique for nominal and continuous (SMOTE-NC) is used to artificially balance the instances where the number of patients having a behavioral outcome of “1” is the minority, which is most often the case as cancer survivor datasets are often imbalanced. The SMOTE-NC technique oversamples the minority class in unbalanced datasets by creating synthetic examples instead of oversampling using replacement. The algorithm involves computing the median of the SD of continuous variables for the minority class and using the median to penalize nominal features that differ between the considered feature vector and its potential nearest neighbors, conducting nearest neighbors computation, and populating the synthetic class [24]. The SMOTE-NC technique is also used to artificially oversample the minority gender so the final datasets can have equal instances of “0” and “1” for the behavioral outcome. We specifically chose the SMOTE-NC technique over the regular synthetic minority oversampling technique because our dataset had a mixture of nominal and continuous features. synthetic minority oversampling technique can only handle datasets with continuous features. The data were then split into 69.6% (71/102) training and 30.4% (31/102) testing data.

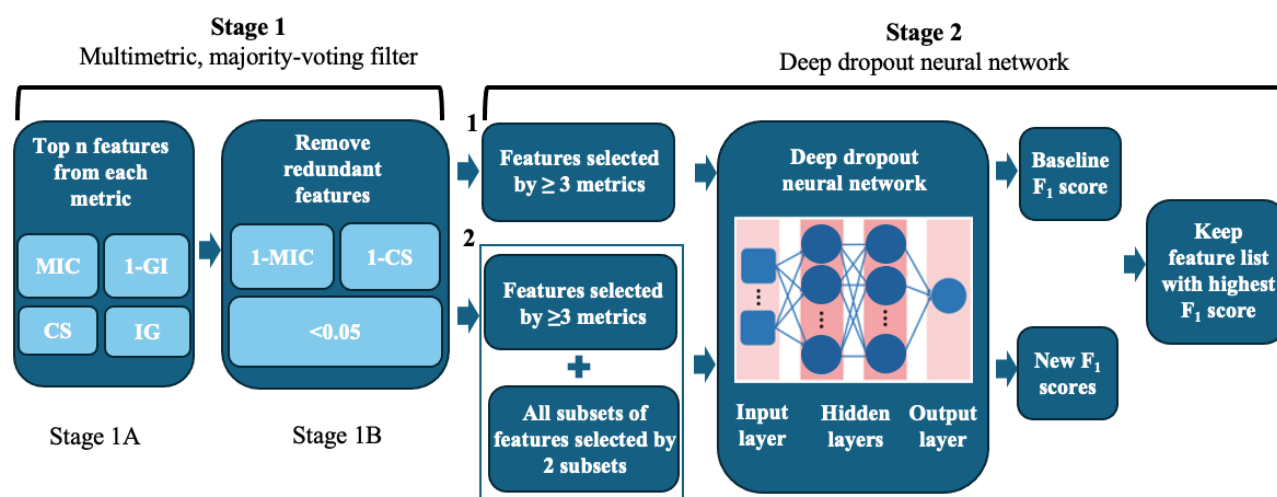
Once the survivors of cancer’ clinical records passed through all the steps of data preprocessing, they were passed into our

hybrid deep learning–based feature selection that was a 2-stage feature selection algorithm, that is, a multimetric, majority-voting filter and a DDN, to dynamically and automatically select the best set of features for each behavioral outcome. Specifically, the first stage was a novel filter method that uses 4 metrics to select the most relevant features for a behavioral outcome and removes any redundant features. The second stage was a DDN that replaces a wrapper method, where it further selects features from the ones selected by the multimetric, majority-voting filter to maximize prediction performance in ML classifiers. Note that our clinical domain experts used their clinical expertise to recommend certain features that should be kept in all the final feature lists due to their clinical importance (ie, gender, current age, and age at diagnosis in our case), if they were not already selected to be in the final feature list by our feature selection approach. Finally, the training data with the final feature list selected from the feature selector with the clinical domain expertise were passed into 3 ML classifiers, including logistic regression, naive Bayes, and k-nearest neighbors, to calculate the precision, recall, and  $F_1$ -score for the performance evaluation on the testing data.

## 2-Stage Feature Selection Algorithm

Our proposed 2-stage feature selection algorithm consisted of 2 sequential stages, including a multimetric, majority-voting filter, and a DDN (Figure 2).

**Figure 2.** Two-stage feature selection algorithm.



### Stage 1: A Multimetric, Majority-Voting Filter

#### Overview

Our hybrid deep learning–based feature selection methodology specifically addressed the limitations of existing feature selection methods. In the first stage, it removed redundant features, which some existing filter methods do not consider. Specifically, our 3 majority-voting (MV) filter had 2 processing steps in stage 1.

In stage 1A, we used 4 different metrics to select the features that are the most relevant to predict a behavioral outcome. Those metrics include maximal information coefficient (MIC), Gini

index (GI), IG, and correlation score (CS) that we calculated between each candidate feature in our preprocessed dataset and the corresponding behavioral outcome of interest. The MIC is a measure of the strength of the linear or nonlinear association between 2 variables  $X$  and  $Y$ , where  $X \in \mathcal{R}$  is the input feature and  $Y \in \mathcal{R}$  is the corresponding behavioral outcome.

The GI represents the amount of probability of a specific feature that is classified incorrectly when selected randomly. Unlike the other 3 metrics, a higher GI score represents lower associations with the behavioral outcome of interest. To make the scale of the correlation strength between  $X$  and  $Y$  consistent among all the metrics, the metric that we used was 1–GI instead.



That is, for all the 4 metrics, a higher value indicated a higher association with the behavioral outcome of interest.

The IG is a measure of the expected reduction in entropy caused by partitioning the samples according to a specific attribute  $X$ .

The CS between  $X$  and  $Y$  is calculated using the Pearson correlation coefficient, point-biserial correlation, and the  $\phi$  coefficient, based upon the data type of  $X$  and  $Y$  [25]. When both  $X$  and  $Y$  are the continuous variables, the Pearson correlation coefficient should be used. When comparing 1 continuous and 1 binary variable, the point-biserial correlation is used [26]. Finally, when comparing 2 binary variables, the  $\phi$  is used. All these measures are values between  $-1$  and  $1$ , with  $-1$  being a perfect negative correlation and  $1$  being a perfect positive correlation, while  $0$  represents no correlation. We take the absolute value of each measure so that the CS is always between  $0$  and  $1$ .

After we calculated the values of all 4 abovementioned metrics between each candidate feature and the behavioral outcome of interest, we ranked the top  $N$  features (ie, the number of features expected by the domain experts) for each of the metrics in descending order and stored them in a master list, without repetition. From this master list, we constructed 3 feature lists. The first list contained the features selected by at least 3 metrics, as they are highly likely relevant to predict the behavioral outcome and are then included in the final feature list. The second one contained the features selected by exactly 2 metrics, as they might have been relevant to predict the behavioral outcome and were then needed for further analysis in stage 1B. The third one combined all the features from the previous 2 lists so that we could evaluate the redundancy between any 2 features from this list.

In stage 1B, we removed any redundant features from the third combined list generated from stage 1A. We used the  $MIC$  and the  $CS$  and then calculated these 2 values for all the feature-to-feature combinations in the combined feature list output from stage 1A. We subtracted the  $MIC$  and the  $CS$  values from  $1$  and then used the  $1-MIC$  and  $1-CS$  values to determine if any feature was redundant by other features. The threshold

we set was  $0.05$ , based upon our preliminary experimental analysis, so that any combination of 2 features that resulted in both scores being  $<0.05$  was determined to be redundant. Once it was determined that 2 features were redundant, we looked at the number of metrics that selected the features. If one of the features was selected by fewer metrics, that feature was removed from the third combined list. If both features were selected by the same number of metrics and they were redundant, we then looked at the average rank of each feature across the 4 ranked lists by  $MIC$ ,  $GI$ ,  $IG$ , and  $CS$ . The feature with the lower rank was removed from the third combined list. The pseudocode algorithm is detailed for the multimetric, majority-voting filter in [Multimedia Appendix 1](#).

For illustration, we used our dataset as an example to explain our multimetric, majority-voting filter.

### Stage 1A: Select the Top $N$ Features Per Metric

#### Overview

Suppose we want to select the best features for predicting the behavioral outcome, thought problems. This is our  $B\_Outcome$ .  $F$  is the set of all input candidate features  $F_i$  in the preprocessed clinical records. We then calculate the  $MIC$ ,  $1-GI$ ,  $IG$ , and  $CS$  scores for all the candidate features in the preprocessed clinical records and our  $B\_Outcome$ , thought problems. We store these results in 4 sets,  $MIC$ ,  $1-GI$ ,  $IG$ , and  $CS$ . In this example, our domain experts expected 15 nonredundant input candidate features to be selected; thus,  $N$  was set to 15.

#### Step 1

We first sorted the input features (ie,  $F_i$ s) according to their  $MIC$ ,  $1-GI$ ,  $CS$ , and  $IG$  scores. Since  $N$  was 15, we then took the top 15 features with the highest values from the  $MIC$  set and placed them into a separate set, that is,  $F_{MIC}$ . We repeated this with  $(1-GI)$ ,  $IG$ , and  $CS$  scores and placed the top 15 features into the corresponding sets, that is,  $F_{1-GI}$ ,  $F_{CS}$ , and  $F_{IG}$ . At this point, we had the following features in these sets:  $F_{MIC}$ ,  $F_{1-GI}$ ,  $F_{CS}$ , and  $F_{IG}$ . As there were 15 features in each set, we had 60 features across all the 4 sets ([Textbox 2](#)).

**Textbox 2.** Total input features sorted by maximal information coefficient (MIC), 1- Gini index (GI), correlation score (CS), and information gain (IG) scores in descending order.

#### **$F_{MIC}$**

Physical fatigue>overall fatigue>cognitive fatigue>family communication>family concern>IV high-dose methotrexate (MTX)>sleep fatigue>physical activity>family conflict>parental control>family mutuality>age at cancer diagnosis>intrathecal MTX dose>noncranial radiation>cranial radiation therapy

#### **$F_{1-GI}$**

Years of education>intrathecal chemotherapy>leukemia risk group>intrathecal MTX dose>living space>physical activity>cognitive fatigue>family communication>physical fatigue>family mutuality>IV high-dose MTX>sleep fatigue>family conflict>age at cancer diagnosis>age at evaluation

#### **$F_{CS}$**

Physical fatigue>overall fatigue>cognitive fatigue>family communication>IV high-dose MTX>family concern>sleep fatigue>family conflict>parental control>physical activity>cranial radiation therapy>noncranial radiation>intrathecal MTX dose>years of education>family mutuality

#### **$F_{IG}$**

Impulsivity (on continuous performance test [CPT; Conner continuous performance test to measure a person's performance in attention, particularly in areas of inattentiveness, impulsivity, variation in response speed, sustained attention, and information processing efficiency] attention test)>inattentiveness (on CPT Attention test)>information processing efficiency (on CPT attention test)>hematopoietic stem cell transplant>response speed variability (on CPT Attention Test)>surgery>sustained attention (on CPT attention test)>physical fatigue>overall fatigue>neurological complications>leukemia risk group >living space>inattentiveness (on CPT attention test)>inflammatory interleukin-7

### **Step 2**

We then created a new set  $F_{UNION}$ , the union of sets  $F_{MIC}$ ,  $F_{1-GI}$ ,  $F_{CS}$ , and  $F_{IG}$  in step 1, allowing duplicate values. This set  $F_{UNION}$  represents all the features that have the top 15 MIC, 1-GI, IG, and CS scores. At this point, the set  $F_{UNION}$  contained 60 total features.

### **Step 3**

From the set  $F_{UNION}$ , we created the subset  $3Metrics+$  from the features that were stored in at least 3 of these 4 sets,  $F_{MIC}$ ,  $F_{1-GI}$ ,  $F_{CS}$ , and  $F_{IG}$ . These features were then selected as 1 of the top 15 by at least 3 out of the 4 metrics, so these are likely to be highly relevant to predict our  $B\_Outcome$ , thought problems, and were included in the final feature list. By applying this concept, the subset  $3Metrics+$  contained 10 features.

### **Step 4**

From the set  $F_{UNION}$ , we also created a subset  $2Metrics$  from features that were stored in exactly 2 out of these 4 sets,  $F_{MIC}$ ,  $F_{1-GI}$ ,  $F_{CS}$ , and  $F_{IG}$ . These features were selected as the top 15 by 2 out of the 4 metrics only. Thus, they may be relevant to predict the  $B\_Outcome$ , thought problems, but needed to be further analyzed in stage 2 to determine if they should be kept in the final feature list. By applying this concept, the subset  $2Metrics$  contained 8 features only.

### **Step 5**

We created another set  $3+2Metrics$ , that is, the union of the sets  $3Metrics+$  and  $2Metrics$ , without the duplicate values. At this point, the set  $3+2Metrics$  contained 18 features, including 10 in the  $3Metrics+$  set and 8 in the  $2Metrics$  set (Textbox 3).

**Textbox 3.** Features in the 3Metrics+ and 2Metrics sets.

### 3Metrics+

- Physical fatigue
- Overall fatigue
- Cognitive fatigue
- Family communication
- Sleep fatigue
- Family conflict
- Family mutuality
- Physical activity
- IV high-dose methotrexate (MTX)
- Intrathecal MTX dose

### 2Metrics

- Leukemia risk group
- Living space
- Family concern
- Cranial radiation therapy
- Years of education
- Family control
- Age at cancer diagnosis
- Noncranial radiation

### Step 6

We also created a 1D matrix, *Rank*, which stored the average rank position of each feature in  $3+2Metrics$  from the sets  $F_{MIC}$ ,  $F_{I-Gb}$ ,  $F_{CS}$ , and  $F_{IG}$ . For instance, if we consider the feature “physical fatigue,” as its position was 1, 9, 1, and 9 in the sets  $F_{MIC}$ ,  $F_{I-Gb}$ ,  $F_{CS}$ , and  $F_{IG}$ , respectively, its average position value in *Rank* was equal to 5.

### Step 7

Finally, we evaluated whether there were too many or too few features at this stage. We first evaluated the number of features in  $3Metrics+$ . As  $3Metrics+$  had 10 features, which was less than  $N$ , there was no need to remove any extra features. We then evaluated the number of features in  $3+2Metrics$ . As there were 18 features in  $3+2Metrics$ , which was greater than  $N$ , there was no need to go back to step 1 to find at least 15 features. We now had 3 sets as the outputs:  $3Metrics+$  with 10 features that were selected by at least 3 metrics;  $2Metrics$  with 8 features that were selected by exactly 2 metrics; and  $3+2Metrics$ , with 18 features that included the features from both  $3Metrics+$  and  $2Metrics$ .

### Stage 1B: Remove Redundant Input Features

At this step, we wanted to remove any redundant features from the features that we selected in stage 1A.

### Step 1

We computed  $1-MIC(f_i, f_j)$  values and  $1-CS(f_i, f_j)$  values by the developed *compute\_MIC* and *compute\_CS* functions between any pair of 2 features  $f_i$  and  $f_j$  in  $3+2Metrics$ . We stored the  $1-MIC(f_i, f_j)$  values and  $1-CS(f_i, f_j)$  values in the sets *MIC\_Feature\_Score* and *CS\_Feature\_Score*, respectively.

### Step 2

We iterated each value in *MIC\_Feature\_Score* and *CS\_Feature\_Score* between any pair of 2 features  $f_i$  and  $f_j$  in  $3+2Metrics$  and checked if any values were  $<0.05$ . We then checked if there was any feature pair that had values  $<0.05$  in both *MIC\_Feature\_Score* and *CS\_Feature\_Score*. Suppose we found that the values in *MIC\_Feature\_Score* and *CS\_Feature\_Score* that corresponded to the feature pair, “cranial radiation therapy” and “noncranial radiation,” were indeed both  $<0.05$ , then we select those 2 features as the feature pair that we need to further analyze, as they were categorized as the redundant features at this step. Suppose that “cranial radiation therapy” and “noncranial radiation” were both in the set  $2Metrics$ , meaning that they were both selected by 2 metrics, then according to the algorithm, they were selected by an equal number of metrics and we must compare their rankings in *Rank* to decide which one must be removed. Suppose that “noncranial radiation” had a lower rank, or a higher score, compared to “cranial radiation therapy,” then we remove “noncranial radiation” from the set  $3+2Metrics$ .

**Step 3**

After we removed the redundant features from the set  $3+2Metrics$ , we then split the set  $3+2Metrics$  into 2 new sets:  $F_{3M+}$ , such that its nonredundant features were selected by at least 3 metrics in the set  $F_{UNION}$ , and  $F_{2M}$ , such that its

nonredundant features were selected by exactly 2 metrics in the set  $F_{UNION}$ .

**Step 4**

We now had 2 sets:  $F_{3M+}$  and  $F_{2M}$ . The set  $F_{3M+}$  had 10 features and the set  $F_{2M}$  had 7 features after we removed “noncranial radiation” (Textbox 4).

**Textbox 4.** Nonredundant features in the set  $F_{3M+}$  and set  $F_{2M}$ .

 **$F_{3M+}$** 

- Physical fatigue
- Overall fatigue
- Cognitive fatigue
- Family communication
- Sleep fatigue
- Family conflict
- Family mutuality
- Physical activity
- IV high-dose methotrexate (MTX)
- Intrathecal MTX dose

 **$F_{2M}$** 

- Leukemia risk group
- Living space
- Family concern
- Cranial radiation therapy
- Years of education
- Parental control
- Age at cancer diagnosis

At this step, we checked if the sum of features from  $F_{3M+}$  and  $F_{2M}$  was  $<25$ . After removing redundant features, we still had 17 features, which was greater than  $N=15$ ; thus, we do not need to go back to step 1 in stage 1A to find at least 15 features. We can then proceed to stage 2.

**Stage 2: A DDN****Overview**

In the second stage, the deep neural network had a dropout parameter, where neurons are randomly ignored during construction of the neural network, to avoid model overfitting, which is a problem that the existing wrapper methods have. Thus, our methodology is better suited for finding the best features from the high-dimension, low-sample size dataset. More specifically, after the features were processed by our multimetric majority-voting filter, we passed all the nonredundant features to the deep dropout neural (DDN) network that was designed to determine whether adding any of those features selected by the only 2 metrics to the list of the features selected by at least 3 metrics resulted in a higher  $F_1$ -score. Note that this step was not conducted if the number

of the nonredundant features, that is, those features that were already selected by at least 3 metrics in stage 1, had met the domain experts' expectation. Our designed DDN network was a 2-hidden- and 1-output-layer architecture. Due to the limited number of patients' medical records with many input features, our DDN network was likely to quickly overfit a training dataset. To address this issue, we used the grid search algorithm with the  $K$ -fold cross-validation (CV) to find the best dropout rate for our network. We also dynamically set the network's hidden layer size using the formula  $\lceil \frac{I}{O} \rceil$ , where  $I$  is the number of selected input subset features and  $O$  is the number of labels per behavioral outcome [27]. For the remaining network's initialization parameters, default values were used [28]. The goal was to perform the hyperparameter tuning using the grid search algorithm with the  $K$ -fold CV to obtain the optimal parameters' values, including the dropout rate, all the network's parameters, and the size of each hidden layer [29].

Specifically, the subset of features selected by  $\geq 3$  metrics in stage 1 was used in building the initial network architecture to produce the baseline  $F_1$ -score. This baseline  $F_1$ -score tells us how well the network predicts that a cancer survivor will


develop the behavioral outcome of interest, using only the features selected by at least 3 metrics. Afterward, we wanted to see whether adding any subset of features selected by 2 metrics would improve the baseline  $F_1$ -score. To achieve this, we tried different combinations among the features selected by 2 metrics; added them on top of the features selected by at least 3 metrics; used all those features to build, train, and optimize our network using the grid search algorithm with the  $K$ -fold CV to obtain the optimal parameters' values; and then recorded each new  $F_1$ -score. This allowed us to compare  $F_1$ -scores between the baseline and the baseline plus additional subsets of features. If any of the new  $F_1$ -scores were higher than the baseline, then our final feature list was the one that produced the highest  $F_1$ -score. If none of the new  $F_1$ -scores were higher than the baseline, then our final feature list was simply the baseline features, that is, the features selected by at least 3 metrics. A step-by-step pseudocode algorithm for our DDN network is detailed in [Multimedia Appendix 2](#).

Let us use our dataset as an example to explain our DDN network. At this stage, we wanted to determine whether any features selected by 2 metrics should be kept in the final feature list on top of the features selected by at least 3 metrics. Our input included the following:

1.  $F_{3M+}$  and  $F_{2M}$ , which were our outputs from stage 1B.
2.  $Drop\_Out\_Rate$ , a set of fine-tuning dropout rates for building a DDN network.
3.  $D\_Train$ , which was the training dataset that only included features in  $F_{3M+}$ .
4.  $Z$ , the set that included all possible subsets from  $F_{2M}$ , excluding the null set, where the size of subsets was less than or equal to  $N$  minus the size of  $F_{3M+}$  so that the total number of features does not exceed  $N$ . In our example, the set  $Z$  only included all the possible subsets of size  $\leq 5$  because we already had 10 features in  $non\_redundant\_three\_more$  and  $N$  minus 10 was 5. Given that there were 7 features in  $non\_redundant\_two$ , there were 128 possible subsets. However, because we only needed the subsets with size  $\leq 5$  and we also excluded the null set, we ended up with a total of 119 different subsets in the set  $Z$ .
5.  $M$ , a set of lists that add all the possible subsets in the set  $Z$  to the set  $F_{3M+}$ ; thus, there were 119 different lists.
6.  $E\_Train$ , which is the set of training datasets that includes features in each list in  $M$ .
7.  $K$ , the number of training partitions on  $D\_Train$  and  $E\_Train$  for performing CV.

### Step 1

We wanted to find the best dropout rate for the neural network, using the *grid-search* technique,  $F_1$ -score, and  $K$ -fold CV, on  $D\_Train$ ,  $F_{3M+}$ ,  $B\_Outcome$ , and  $Drop\_Out\_Rate$  of a DDN network.  $K$  was set to 5. We thus first constructed a neural network using the *create\_DD*N function to perform the *grid-search* technique. The neural network was initialized to have a learning rate of 0.001, 500 epochs, used the “Adam” optimizer, used the “Binary Cross Entropy” loss function, had

2 hidden layers with  number of neurons and the “Relu” activation function, and 1 output layer with 1 neuron and the activation function “Sigmoid.” Suppose using the *grid-search* technique with the  $D\_Train$  training dataset, the  $F_{3M+}$  feature set, the  $B\_Outcome$  thought problems, the set of fine-tuning dropout rates  $Drop\_Out\_Rate$ , and using 5-fold CV, we found that the best dropout rate was 0.1 (*bestDropOutRate* was set to 0.1).

### Step 2

We constructed a deep neural network with the initialized attributes in the *create\_DD*N function, *bestDropOutRate*,  $D\_Train$ ,  $F_{3M+}$ , and  $B\_Outcome$ , and then performed 5-fold CV to obtain the baseline  $F_1$ -score,  $F1_{Baseline}$ .

### Step 3

We then iterated through each feature set (ie,  $F_{3M+}+Z_r$ ) in  $M$  and constructed a deep neural network with the same initialized attributes in the *create\_DD*N function, *bestDropOutRate*,  $E\_Train$ ,  $F_{3M+}+Z_r$ , and  $B\_Outcome$ , and then performed 5-fold CV to obtain the  $F_1$ -score,  $F1$ , for each training dataset in  $E\_Train$ . The hidden layer size of each neural network was calculated using the number of features in  $M+1$ , divided by 2. If any  $F_1$ -score was greater than  $F1_{Baseline}$ , the final feature list (ie, *Final\_Features*) was set to the feature set (ie,  $F_{3M+}+Z_r$ ) in  $M$  in which the  $F_1$ -score was obtained.

### Step 4

We had the feature list with the best  $F_1$ -score (ie, *Final\_Features*), which was passed into 3 ML classifiers: logistic regression, naive Bayes, and k-nearest neighbors.

### Pilot Experimental Study

In our experimental study, we used a 2018 to 2020 dataset that contained 102 ALL survivors' clinical records collected from a public hospital in Hong Kong. The survivors were aged between 15 and 39 years, had completed treatment, and were >5 years postcancer diagnosis at the time of recruitment. In each patient record, there were >50 features, including demographic factors (eg, age, gender, and education level), cancer treatments received (eg, radiation, chemoradiotherapy, and surgery), inflammatory biomarkers (eg, interleukin-7, monocyte chemoattractant protein-1, and tumor necrosis factor alpha- $\alpha$ ), physical health conditions (eg, BMI, sleep fatigue, and cognitive fatigue), family life and socioeconomic descriptors (eg, family conflict, family communication and living space), attention-related outcomes (eg, measures of inattentiveness, impulsivity, and sustained attention), and lifestyle habits (eg, drinking, smoking, and physical activity). The features were obtained from a behavioral assessment that included the traditional Chinese version of the Achenbach System of Empirically Based Assessment youth self-report checklist. It consisted of syndrome scales measuring attention problems, thought problems, internalizing problems (eg, somatic complaints, anxiety and depressive symptoms, and withdrawn behavior), externalizing problems (eg, aggressive behavior, intrusive behavior, and rule-breaking behavior), and sluggish cognitive tempo. The Achenbach System of Empirically Based



Assessment measures were previously validated and used in the local young adult cancer population [9,30]. The inclusion of these features specifically in patient records was based on existing evidence in the literature and data from the local study cohort. The features predicting behavioral outcomes included clinical factors (eg, leukemia risk group, age at cancer diagnosis, and neurological complications), treatment factors (eg, cranial radiation therapy, intrathecal methotrexate dose, intravenous high-dose methotrexate, and hematopoietic stem cell transplant), socioenvironmental factors (eg, living space and family functioning), and lifestyle factors (eg, physical activity and sleep fatigue) [9,30-34].

After preprocessing the data and using our 2-stage feature selection algorithm, we selected 15 input features, expected by our medical investigators, to train and test our 3 ML classifiers, that is, logistic regression, naive Bayes, and k-nearest neighbors, to predict 6 behavioral outcomes (ie, anxiety and depression, thought problems, attention problems, internalizing problems, externalizing problems, and sluggish cognitive tempo) that our medical investigators would like to focus on. Due to their

clinical importance recommended by our medical investigators, we also added 3 more clinically relevant features (ie, gender, current age, and age at diagnosis) to the final feature list if those features had not been already selected by our 2-stage feature selection approach.

### Ethical Considerations

Approval of this study was obtained from the Joint Chinese University of Hong Kong – New Territories East Cluster Clinical Research Ethics Committee (2017.701). Written informed consent was obtained from all participants.

## Results

### Overview

The experimental results included the  $F_1$ -score, precision, and recall on the testing data (Table 1). Note that for each feature selection method category, those scores are the average values of prediction performance among all the 3 ML classifiers for every behavioral outcome.

**Table 1.** Average  $F_1$ -scores.

Behavioral outcome	Filter	Wrapper	Embedded	Hybrid	Our method	Percentage change (our method vs highest baseline)
<b>Anxiety and depression</b>						
$F_1$ -score	0.624	0.437	0.585	0.449	<i>0.738<sup>a</sup></i>	+18.27
Precision score	0.562	0.407	0.563	0.424	<i>0.708</i>	+25.75
Recall score	0.813	0.519	0.630	0.580	<i>0.778</i>	−4.31
<b>Thought problems</b>						
$F_1$ -score	0.490	0.438	0.477	0.394	<i>0.511</i>	+4.29
Precision score	0.522	0.385	0.590	0.496	0.448	−24.07
Recall score	0.537	0.556	0.463	0.383	<i>0.611</i>	+9.89
<b>Attention problems</b>						
$F_1$ -score	0.348	0.417	0.440	0.350	<i>0.568</i>	+29.10
Precision score	0.290	0.360	0.350	0.329	<i>0.515</i>	+43.10
Recall score	0.463	0.519	0.630	0.424	<i>0.667</i>	+5.87
<b>Internalizing problems</b>						
$F_1$ -score	0.533	0.706	0.619	0.637	0.700	−0.85
Precision score	0.583	0.665	0.665	0.668	0.618	−7.49
Recall score	0.587	0.762	0.651	0.651	<i>0.857</i>	+12.47
<b>Externalizing problems</b>						
$F_1$ -score	0.219	0.459	0.267	0.265	0.278	−39.43
Precision score	0.230	0.417	0.278	0.297	<i>0.444</i>	+6.47
Recall score	0.222	0.556	0.259	0.259	0.222	−60.07
<b>Sluggish cognitive tempo</b>						
$F_1$ -score	0.560	0.463	0.582	0.489	<i>0.639</i>	+9.79
Precision score	0.542	0.409	0.577	0.494	0.570	−1.21
Recall score	0.654	0.568	0.617	0.568	<i>0.741</i>	+13.30

<sup>a</sup>Italicized values indicate that our score was higher than the other 4 methods.

Our 2-stage feature selection approach outperformed or leveled the existing feature selection methods to support the prediction of 5 out of 6 behavioral outcomes (ie, anxiety and depression, thought problems, attention problems, internalizing problems, and sluggish cognitive tempo) in terms of the average  $F_1$ -scores (Table 1). Although the wrapper method outperformed our feature selection approach to support the prediction of externalizing problems, our approach's performance was more stable, as the  $F_1$ -score variance was smaller. Thus, our feature selection approach still outperforms the other 3 existing feature selection methods.

In addition, our feature selection approach outperformed or leveled the existing feature selection methods to support the prediction of 5 out of 6 behavioral outcomes (ie, anxiety and depression, attention problems, internalizing problems, externalizing problems, and sluggish cognitive tempo) in terms of precision scores (Table 1). Although the embedded method outperformed our feature selection approach to support the prediction of thought problems, our approach's performance variance was much smaller, which implies our approach was more stable.

Finally, our feature selection approach outperformed the existing feature selection methods to support the prediction of 4 out of 6 behavioral outcomes (ie, thought problems, attention problems, internalizing problems, and sluggish cognitive tempo) in terms of recall scores (Table 1). Although the filter and wrapper method outperformed our feature selection approach to support the prediction of anxiety and depression and externalizing problems, our approach's performance variance was much smaller as well.

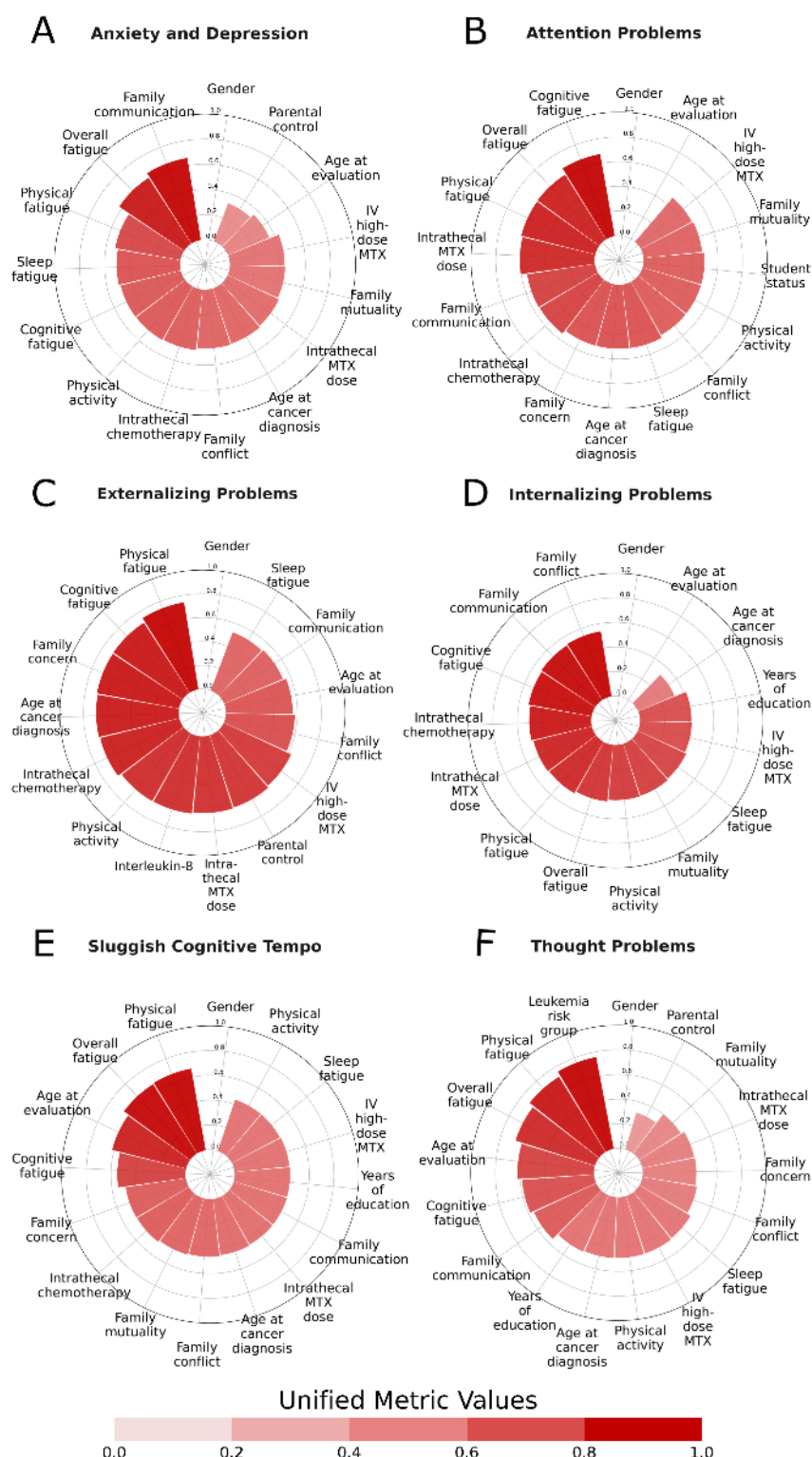
As the  $F_1$ -scores were calculated from both precision and recall scores, we can infer that our feature selection approach improves the  $F_1$ -scores largely because it increases the recall scores as opposed to the precision scores (Table 1). Overall, the experimental results show promising evidence that our method improves the ML classifiers' prediction performance to support better early detection of long-term behavioral outcomes in survivors of cancer.

### Radial Feature Charts

Radial feature charts were generated for each of the 6 behavioral outcomes analyzed, including anxiety and depression, thought problems, attention problems, internalizing problems, externalizing problems, and sluggish cognitive tempo (Figure 3). Each chart includes the top 15-plus features selected by our proposed methodology. The size and the color of each red slice is measured by the unified metric value of each feature, which is calculated by averaging the scores of the metrics that select each feature during stage 1A of our proposed method.

The variables represent the documented risk factors associated with the development of behavioral problems in the literature. They include (1) sociodemographic variables (ie, age at evaluation and gender), (2) clinical variables (ie, age at cancer diagnosis, intrathecal chemotherapy, intrathecal methotrexate dose, IV high-dose methotrexate, and inflammatory interleukin-8 levels), and (3) socioenvironmental and lifestyle variables (ie, sleep, fatigue, physical activity, and family functioning). Physicians can interpret the charts by seeing which features have the darkest color and largest size, indicating higher unified metric values and thus greater associations with the behavioral problem of interest. Those features can then be further used to devise customized prevention plans and advice.

Figure 3. Radial feature charts.



## Discussion

### Principal Findings

In this work, we sought to develop a prognostic ML framework and feature selection approach to predict the trajectory of functional outcomes in a specific population: survivors of ALL. Our hybrid deep learning-based feature selection approach outperforms or equals the existing feature selection methods

assessed (ie, filter, wrapper, embedded, and hybrid) for 5 out of 6 long-term behavioral outcomes. Even in cases where our feature selection method did not outperform existing methods, our approach's performance variance was much smaller and thus more stable. We observed that the performance of the model was significantly weaker in predicting externalizing problems than internalizing problems. This may be attributed to the complex phenotypic nature of externalizing behaviors, such as antisocial or aggressive behaviors and conduct problems. In

addition, there are other factors that may predict externalizing problems that were not considered in this study. For example, our previous work showed that increased screen time during the COVID-19 pandemic was associated with inattentiveness and impulsivity in pediatric survivors of cancer in China, but screen time was not included in the data [35]. Social support and rehabilitation, which are important interventions addressing behavioral functioning and mental health in young Chinese survivors of cancer, were also not assessed in this study [36]. From the data, we infer that our feature selection approach improves  $F_1$ -scores from ML classifiers compared to existing feature selection methods largely because it increases the recall scores as opposed to the precision scores. We also developed radial feature charts that can quickly and effectively help clinicians understand which predictor variables were most important in predicting long-term behavioral outcomes. Overall, the experimental results show promising evidence that our method improves ML classifiers' prediction performance on high-dimension low-sample size data, which can support better early detection of long-term behavioral outcomes in survivors of cancer.

### Limitations

Our study was limited to a pilot study with young Chinese survivors of leukemia. As one's neurodevelopment and social skills are often dependent on cultural norms, our findings may not be extrapolated or applicable to other populations. However, the contemporary treatment for childhood ALL is similar in most countries or regions, consisting of high-dose methotrexate, intrathecal chemotherapy, and a standard set of intravenous and oral chemotherapy drugs as the backbone. Therefore, we reasoned that our findings may still be generalizable to the existing population of individuals in the health care system of Hong Kong who have survived leukemia over the past decade. In addition, although clinical domain experts assisted with additional input for the features that were kept in ML classifiers, there remains room for human error, and domain experts' opinions may occasionally differ from what features would optimize ML classifiers' performance. Furthermore, as this is a cross-sectional study, it was not possible to delineate the causal relationship between the risk factors and behavioral outcomes. The model developed through this study should be validated in a larger cohort with prospective collection of outcome data to better reflect the trajectories of functional outcomes in these young survivors as they advance from young to middle adulthood. Finally, additional biases may have influenced the data, such as those related to patients who had access to hospital care and were willing to share their data with our clinical investigators.

### Comparison With Prior Work

Our findings reinforce existing evidence that adverse behavioral outcomes in survivors of cancer are a complex and multifactorial phenotype. Most preexisting research is focused on either disease- or treatment-related factors as predictors of cognitive dysfunction. However, socioenvironmental factors play an important role in the neurodevelopment of these young survivors. Our findings showed the interaction and unique contribution of the socioenvironmental factors, such as family

dynamics and lifestyle factors, on anxiety, depression, and sluggish cognitive symptoms in survivors. Studies have found associations of parents' psychological distress on the child's cognitive and behavioral outcomes [8,37]. Environmental events can elicit a biological stress response that results in neurological reactions to that stress. This is especially relevant in the context of Hong Kong and Mainland China, where much emphasis is now placed on ameliorating the adverse health effects of the urban environment in children and adolescents. The findings provide directions for the development of multidisciplinary services and interventions. For example, social workers can pay more attention to the occupational or employment challenges of young survivors who experience fatigue symptoms from treatment and manifest adverse behavioral outcomes. The study findings can help us identify high-risk subgroups from dysfunctional families or households struggling with financial problems and conflicts. Interventions that promote self-confidence and positive peer interaction can be implemented during the early survivorship phase when young survivors transit back to their full-time school or work.

Our results also build upon existing computational methods and feature selection approaches for predicting behavioral outcomes in survivors of cancer. Traditional computational methods in the clinical and social sciences typically use regression analysis to model the relationship between  $\geq 2$  variables for prediction. However, modeling human behavioral data is challenging due to its multifactorial nature, heterogeneity, nonlinearity of data, and class imbalance [10,11]. As a result, the model can only account for a small proportion of variance, with limited utility in clinical settings. For example, we have reported that cranial radiation, chronic health conditions, and poor physical activity are associated with worse cognitive and behavioral outcomes in Chinese survivors of childhood leukemia [9]. However, these factors only accounted for 22.9% to 35.8% of the variance in the traditional regression models. Identifying an effective computational method that minimizes algorithmic bias, such as the 2-stage feature selection algorithm within the clinical domain-guided framework outlined in this study, can maximize the use of clinical and behavioral data for predictive purposes. Such prognostic models will aid in informing strategies aimed at changing behavior and designing social and clinical interventions.

### Conclusions

Future studies can validate our prediction model in other Chinese populations of survivors of cancer sharing similar cultural norms in mainland China and Taiwan, as well as validate the model in larger samples with a longitudinal prospective cohort study design. In addition, studies can further investigate the real-world feasibility of incorporating such algorithms into health care systems as risk stratification tools to assist clinicians and psychologists in identifying patients at risk of adverse behavioral outcomes. Incorporation of diverse populations, larger sample sizes, and similar prediction models in future studies may provide deeper insights into the interaction among clinical, treatment, socioeconomic, and lifestyle factors and their impact on functional outcomes, ultimately enabling the incorporation of such multifactorial insights to improve strategies for the personalized care of patients with cancer.

Given that we are working with such small cancer survivor datasets, even a slight improvement in prediction performance from ML classifiers can make a substantial difference in helping survivors of cancer. Our data-driven, clinical domain-guided approach can potentially address the problem of “high dimension low sample size.” The pilot analysis shows that this approach has allowed us to identify a set of interacting clinical and socioenvironmental characteristics that predicted behavioral outcomes in survivors.

In late 2019, the American Cancer Society had a special call for attention to financial, social, and emotional concerns that uniquely affect young survivors of cancer [38]. Currently, in Hong Kong, there are no centralized cancer programs for adolescent and young adult patients. From a clinical perspective, identifying the unique factors associated with interindividual

differences in functional outcomes will help clinicians to identify individualized modifiable risk factors. This will contribute to the development of a personalized, patient-centered cancer care program for local patients with cancer. From a research perspective, this project serves as a pilot study to apply ML-based prognostic technology, guided by clinical knowledge, on a combination of objective data (ie, clinical and demographics variables) and subjective data (ie, behavioral and patient-reported variables). The framework and algorithms developed through this analysis can be applied to address clinically relevant research questions in patients with other chronic diseases. The aim of this application is in line with the recent call by the government of the Hong Kong Special Administrative Region to harness data-driven analytics to formulate health care policies [39].

## Acknowledgments

This study is supported by the US National Science Foundation (1852498), awarded to CKN, and partially funded by the Hong Kong Research Grant Council's Early Career Scheme (24614818) and General Research Fund (14604022), awarded to YTC.

## Conflicts of Interest

None declared.

### Multimedia Appendix 1

Step-by-step pseudocode algorithm for the multimetric, majority-voting filter.

[DOCX File, 22 KB - [bioinform\\_v6i1e65001\\_app1.docx](#)]

### Multimedia Appendix 2

Step-by-step pseudocode algorithm for the DDN network.

[DOCX File, 18 KB - [bioinform\\_v6i1e65001\\_app2.docx](#)]

## References

1. Siegel RL, Miller KD, Wagle NS, Jemal A. Cancer statistics, 2023. *CA Cancer J Clin* 2023 Jan;73(1):17-48 [FREE Full text] [doi: [10.3322/caac.21763](#)] [Medline: [36633525](#)]
2. Brinkman TM, Recklitis CJ, Michel G, Grootenhuis MA, Klosky JL. Psychological symptoms, social outcomes, socioeconomic attainment, and health behaviors among survivors of childhood cancer: current state of the literature. *J Clin Oncol* 2018 Jul 20;36(21):2190-2197 [FREE Full text] [doi: [10.1200/JCO.2017.76.5552](#)] [Medline: [29874134](#)]
3. Yeh JM, Ward ZJ, Chaudhry A, Liu Q, Yasui Y, Armstrong GT, et al. Life expectancy of adult survivors of childhood cancer over 3 decades. *JAMA Oncol* 2020 Mar 01;6(3):350-357 [FREE Full text] [doi: [10.1001/jamaoncol.2019.5582](#)] [Medline: [31895405](#)]
4. Dixon SB, Liu Q, Chow EJ, Oeffinger KC, Nathan PC, Howell RM, et al. Specific causes of excess late mortality and association with modifiable risk factors among survivors of childhood cancer: a report from the Childhood Cancer Survivor Study cohort. *Lancet* 2023 Apr 29;401(10386):1447-1457 [FREE Full text] [doi: [10.1016/S0140-6736\(22\)02471-0](#)] [Medline: [37030315](#)]
5. Lam CS, Lee CP, Chan JW, Cheung YT. Prescription of psychotropic medications after diagnosis of cancer and the associations with risk of mortality in Chinese patients: a population-based cohort study. *Asian J Psychiatr* 2022 Dec;78:103290. [doi: [10.1016/j.ajp.2022.103290](#)] [Medline: [36209707](#)]
6. Suh E, Stratton KL, Leisenring WM, Nathan PC, Ford JS, Freyer DR, et al. Late mortality and chronic health conditions in long-term survivors of early-adolescent and young adult cancers: a retrospective cohort analysis from the Childhood Cancer Survivor Study. *Lancet Oncol* 2020 Mar;21(3):421-435 [FREE Full text] [doi: [10.1016/S1470-2045\(19\)30800-9](#)] [Medline: [32066543](#)]
7. Alias H, Morthy SK, Zakaria SZ, Muda Z, Tamil AM. Behavioral outcome among survivors of childhood brain tumor: a case control study. *BMC Pediatr* 2020 Feb 05;20(1):53 [FREE Full text] [doi: [10.1186/s12887-020-1951-3](#)] [Medline: [32020861](#)]
8. Patel SK, Wong AL, Cuevas M, Van Horn H. Parenting stress and neurocognitive late effects in childhood cancer survivors. *Psychooncology* 2013 Aug 25;22(8):1774-1782 [FREE Full text] [doi: [10.1002/pon.3213](#)] [Medline: [23097416](#)]



9. Peng L, Yang LS, Yam P, Lam CS, Chan AS, Li CK, et al. Neurocognitive and behavioral outcomes of Chinese survivors of childhood lymphoblastic leukemia. *Front Oncol* 2021;11:655669 [FREE Full text] [doi: [10.3389/fonc.2021.655669](https://doi.org/10.3389/fonc.2021.655669)] [Medline: [33959507](https://pubmed.ncbi.nlm.nih.gov/33959507/)]
10. Kliegr T, Bahník Š, Fürnkranz J. Advances in machine learning for the behavioral sciences. *Am Behav Sci* 2019 Jul 24;64(2):145-175. [doi: [10.1177/0002764219859639](https://doi.org/10.1177/0002764219859639)]
11. Turgeon S, Lanovaz MJ. Tutorial: applying machine learning in behavioral research. *Perspect Behav Sci* 2020 Dec 10;43(4):697-723 [FREE Full text] [doi: [10.1007/s40614-020-00270-y](https://doi.org/10.1007/s40614-020-00270-y)] [Medline: [33381685](https://pubmed.ncbi.nlm.nih.gov/33381685/)]
12. Thejas GS, Garg R, Iyengar SS, Sunitha NR, Badrinath P, Chennupati S. Metric and accuracy ranked feature inclusion: hybrids of filter and wrapper feature selection approaches. *IEEE Access* 2021;9:128687-128701. [doi: [10.1109/access.2021.3112169](https://doi.org/10.1109/access.2021.3112169)]
13. Cherrington M, Thabtah F, Lu J, Xu Q. Feature selection: filter methods performance challenges. In: *Proceedings of the 2019 International Conference on Computer and Information Sciences*. 2019 Presented at: ICCIS '19; April 3-4, 2019; Sakaka, Saudi Arabia p. 1-4 URL: <https://ieeexplore.ieee.org/document/8716478> [doi: [10.1109/iccisci.2019.8716478](https://doi.org/10.1109/iccisci.2019.8716478)]
14. Lin X, Li C, Ren W, Luo X, Qi Y. A new feature selection method based on symmetrical uncertainty and interaction gain. *Comput Biol Chem* 2019 Dec;83:107149. [doi: [10.1016/j.compbiolchem.2019.107149](https://doi.org/10.1016/j.compbiolchem.2019.107149)] [Medline: [31751882](https://pubmed.ncbi.nlm.nih.gov/31751882/)]
15. Huang S, Cai N, Pacheco PP, Narrandes S, Wang Y, Xu W. Applications of support vector machine (SVM) learning in cancer genomics. *Cancer Genomics Proteomics* 2018 Jan 2;15(1):41-51 [FREE Full text] [doi: [10.21873/cgp.20063](https://doi.org/10.21873/cgp.20063)] [Medline: [29275361](https://pubmed.ncbi.nlm.nih.gov/29275361/)]
16. Jonas R, Cook J. LASSO regression. *Br J Surg* 2018;105(10):1348. [doi: [10.1002/bjs.10895](https://doi.org/10.1002/bjs.10895)]
17. Hoerl RW. Ridge regression: a historical context. *Technometrics* 2020 Oct 23;62(4):420-425. [doi: [10.1080/00401706.2020.1742207](https://doi.org/10.1080/00401706.2020.1742207)]
18. Alhamzawi R, Ali HT. The Bayesian elastic net regression. *Commun Stat Simul Comput* 2017 Jun 20;47(4):1168-1178. [doi: [10.1080/03610918.2017.1307399](https://doi.org/10.1080/03610918.2017.1307399)]
19. Usman AU, Hassan S, Tukur K. Application of dummy variables in multiple regression analysis. *Int J Recent Sci Res* 2015;7(11):7440-7442 [FREE Full text] [doi: [10.4324/9781315748788-15](https://doi.org/10.4324/9781315748788-15)]
20. Berger VW, Zhou Y. Kolmogorov–Smirnov test: overview. *Wiley StatsRef* 2014;63 [FREE Full text] [doi: [10.1002/9781118445112.stat06558](https://doi.org/10.1002/9781118445112.stat06558)]
21. González-Estrada E, Cosmes W. Shapiro–Wilk test for skew normal distributions based on data transformations. *J Stat Comput Simu* 2019 Aug 27;89(17):3258-3272. [doi: [10.1080/00949655.2019.1658763](https://doi.org/10.1080/00949655.2019.1658763)]
22. Saculinggan M, Balase EA. Empirical power comparison of goodness of fit tests for normality in the presence of outliers. *J Phys Conf Ser* 2013 Apr 26;435:012041. [doi: [10.1088/1742-6596/435/1/012041](https://doi.org/10.1088/1742-6596/435/1/012041)]
23. Patro S. Normalization: a preprocessing stage. *arXiv Preprint* posted online March 19, 2015 [FREE Full text]
24. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. *J Artif Intell* 2002 Jun 01;16:321-357. [doi: [10.1613/jair.953](https://doi.org/10.1613/jair.953)]
25. Akoglu H. User's guide to correlation coefficients. *Turk J Emerg Med* 2018 Sep;18(3):91-93 [FREE Full text] [doi: [10.1016/j.tjem.2018.08.001](https://doi.org/10.1016/j.tjem.2018.08.001)] [Medline: [30191186](https://pubmed.ncbi.nlm.nih.gov/30191186/)]
26. Kornbrot D. Point biserial correlation. *Wiley StatsRef* 2014;22. [doi: [10.1002/9781118445112.stat06227](https://doi.org/10.1002/9781118445112.stat06227)]
27. Lawrence S, Giles CL, Tsoi AC. What size neural network gives optimal generalization? Convergence properties of backpropagation. *Institute for Advanced Computer Studies, University of Maryland*. 1998. URL: <https://api.drum.lib.umd.edu/server/api/core/bitstreams/bf781aeb-eb41-4803-a2ac-7d915d0ac791/content> [accessed 2024-04-29]
28. Zollanvari A. Deep learning with Keras-TensorFlow. In: *Zollanvari A, editor. Machine Learning With Python: Theory and Implementation*. Cham, Switzerland: Springer; 2023:351-391.
29. Belete DM, Huchaiah MD. Grid search in hyperparameter optimization of machine learning models for prediction of HIV/AIDS test results. *Int J Comput Appl* 2021 Sep 12;44(9):875-886. [doi: [10.1080/1206212X.2021.1974663](https://doi.org/10.1080/1206212X.2021.1974663)]
30. Cheung YT, Ma CT, Li MC, Zhou KR, Loong HH, Chan AS, et al. Associations between lifestyle factors and neurocognitive impairment among Chinese adolescent and young adult (AYA) survivors of sarcoma. *Cancers (Basel)* 2023 Jan 28;15(3):799 [FREE Full text] [doi: [10.3390/cancers15030799](https://doi.org/10.3390/cancers15030799)] [Medline: [36765757](https://pubmed.ncbi.nlm.nih.gov/36765757/)]
31. Cheung YT, To KK, Hua R, Lee CP, Chan AS, Li CK. Association of markers of inflammation on attention and neurobehavioral outcomes in survivors of childhood acute lymphoblastic leukemia. *Front Oncol* 2023 Jun 21;13:1117096 [FREE Full text] [doi: [10.3389/fonc.2023.1117096](https://doi.org/10.3389/fonc.2023.1117096)] [Medline: [37416531](https://pubmed.ncbi.nlm.nih.gov/37416531/)]
32. Krull KR, Hardy KK, Kahalley LS, Schuitema I, Kesler SR. Neurocognitive outcomes and interventions in long-term survivors of childhood cancer. *J Clin Oncol* 2018 Jul 20;36(21):2181-2189 [FREE Full text] [doi: [10.1200/JCO.2017.76.4696](https://doi.org/10.1200/JCO.2017.76.4696)] [Medline: [29874137](https://pubmed.ncbi.nlm.nih.gov/29874137/)]
33. Mavrea K, Efthymiou V, Katsibardi K, Tsarouhas K, Kanaka-Gantenbein C, Spandidos D, et al. Cognitive function of children and adolescent survivors of acute lymphoblastic leukemia: a meta-analysis. *Oncol Lett* 2021 Apr 05;21(4):262 [FREE Full text] [doi: [10.3892/ol.2021.12523](https://doi.org/10.3892/ol.2021.12523)] [Medline: [33664825](https://pubmed.ncbi.nlm.nih.gov/33664825/)]
34. van der Plas E, Modi AJ, Li CK, Krull KR, Cheung YT. Cognitive impairment in survivors of pediatric acute lymphoblastic leukemia treated with chemotherapy only. *J Clin Oncol* 2021 Jun 01;39(16):1705-1717. [doi: [10.1200/JCO.20.02322](https://doi.org/10.1200/JCO.20.02322)] [Medline: [33886368](https://pubmed.ncbi.nlm.nih.gov/33886368/)]

35. Cai J, Cheung YT, Au-Doung PL, Hu W, Gao Y, Zhang H, et al. Psychosocial outcomes in Chinese survivors of pediatric cancers or bone marrow failure disorders: a single-center study. PLoS One 2022 Dec 13;17(12):e0279112 [[FREE Full text](#)] [doi: [10.1371/journal.pone.0279112](#)] [Medline: [36512620](#)]
36. Zhu W. The impact of social support on the mental health of cancer patients: evidence from China. Psycho Oncol 2024;18(1):69-77. [doi: [10.32604/po.2023.046593](#)]
37. Hile S, Erickson SJ, Agee B, Annett RD. Parental stress predicts functional outcome in pediatric cancer survivors. Psychooncology 2014 Oct 10;23(10):1157-1164. [doi: [10.1002/pon.3543](#)] [Medline: [24817624](#)]
38. Bhatia S, Pappo AS, Acquazzino M, Allen-Rhoades WA, Barnett M, Borinstein S, et al. Adolescent and Young Adult (AYA) oncology, version 2.2024, NCCN clinical practice guidelines in oncology. J Natl Compr Canc Netw 2023 Aug;21(8):851-880. [doi: [10.6004/jnccn.2023.0040](#)] [Medline: [37549914](#)]
39. Leung KY, Lee HY. Implementing the smart city: who has a say? Some insights from Hong Kong. Int J Urban Sci 2021 Nov 08;27(sup1):124-148. [doi: [10.1080/12265934.2021.1997634](#)]

## Abbreviations

**ALL:** acute lymphocytic leukemia

**CFS:** correlation-based feature selection

**CS:** correlation score

**CV:** cross-validation

**DDN:** deep dropout neural network

**GI:** Gini index

**IG:** information gain

**Lasso:** least absolute shrinkage and selection operator

**MIC:** maximal information coefficient

**ML:** machine learning

**MRMR:** maximum relevance minimum redundancy

**SBS:** sequential backwards selection

**SFS:** sequential forward selection

**SMOTE-NC:** synthetic minority oversampling technique for nominal and continuous

**SS:** stepwise selection

*Edited by Z Yue; submitted 01.08.24; peer-reviewed by W Fu, D Bracken-Clarke, SS Kollala; comments to author 27.10.24; revised version received 16.12.24; accepted 06.01.25; published 13.03.25.*

*Please cite as:*

Huang T, Ngan CK, Cheung YT, Marcotte M, Cabrera B

A Hybrid Deep Learning-Based Feature Selection Approach for Supporting Early Detection of Long-Term Behavioral Outcomes in Survivors of Cancer: Cross-Sectional Study

JMIR Bioinform Biotech 2025;6:e65001

URL: <https://bioinform.jmir.org/2025/1/e65001>

doi: [10.2196/65001](#)

PMID: [40080820](#)

©Tracy Huang, Chun-Kit Ngan, Yin Ting Cheung, Madelyn Marcotte, Benjamin Cabrera. Originally published in JMIR Bioinformatics and Biotechnology (<https://bioinform.jmir.org>), 13.03.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Bioinformatics and Biotechnology, is properly cited. The complete bibliographic information, a link to the original publication on <https://bioinform.jmir.org/>, as well as this copyright and license information must be included.

---

Publisher:  
JMIR Publications  
130 Queens Quay East.  
Toronto, ON, M5A 3Y5  
Phone: (+1) 416-583-2040  
Email: [support@jmir.org](mailto:support@jmir.org)

---

<https://www.jmirpublications.com/>