D1.3

Synthetic data for testing and training patient, cancer, and drug models

Project number	826121
Project acronym	iPC
Project title	individualizedPaediatricCure: Cloud-based virtual-patient models for precision paediatric oncology
Start date of the project	1st January, 2019
Duration	53 months
Programme	H2020-SC1-DTH-2018-1

Deliverable type	Report
Deliverable reference number	SC1-DTH-07-826121 / D1.3 / 1.0
Work package contributing to the deliverable	WP1
Due date	November 2021 – M35
Actual submission date	30th November, 2021

Responsible organisation	BSC
Editor	Davide Cirillo
Dissemination level	PU
Revision	1.0

Abstract	We report on the development of a deep generative model for the generation of synthetic transcriptomic data and its application to paediatric cancer research, specifically a use case in medulloblastoma.
Keywords	Synthetic data, generative model, variational autoencoder, medulloblastoma

The project iPC has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826121.

Editor

Davide Cirillo (BSC)

Contributors (ordered according to beneficiary numbers)

Alejandro Tejada Lapuerta (BSC)

Salvador Capella (BSC)

Alejandro Canosa (BSC)

José María Fernández-González (BSC)

Disclaimer

The information in this document is provided “as is”, and no guarantee or warranty is given that the information is fit for any particular purpose. The content of this document reflects only the author`s view – the European Commission is not responsible for any use that may be made of the information it contains. The users use the information at their sole risk and liability.

Executive Summary

Synthetic data generation is emerging as a dominant solution for precision medicine as it enables to address critical challenges such as yielding the data volumes needed to deliver accurate results and complying with increasingly restrictive privacy regulations, both demanded in paediatric cancer research. In this deliverable, we report on the development of an explainable Variational AutoEnconder (VAE) for synthetic transcriptomics data generation in medulloblastoma, a childhood brain tumour. The model can be used to augment and interpolate available data with synthetic instances, which are automatically annotated with confidence scores to assess the reliability of augmented data points and interpolated trajectories. The model is transparent as it is able to match the learned latent variables with distinct gene expression patterns. We leverage both the synthetic data generation and explainability features of our model to study the gene expression characteristics of the four medulloblastoma subgroups (WNT, SHH, G3, G4) and, in particular, that underlying the unknown relationship between G3 and G4 subgroups. Additionally, we generated a datasets of 4,000 high fidelity synthetic medulloblastoma expression profiles encompassing all subgroups. The draft of the unpublished article resulted from this work and the synthetic dataset that we generated are available at the iPC Nextcloud public repository https://data.ipc-project.bsc.es/s/sqc3WWQTLgC8iAb. The model can be adapted to other paediatric cancers and the resulting synthetic datasets be used for testing and training patient, cancer, and drug models in other workpackages of the iPC project.

Table of Content

List of Figures

Chapter 1 Synthetic data generation in paediatric cancer research

Despite the tremendous medical progress over the last decades, many diseases still lack effective interventions and, in many cases, these are not targeted to the individual, leading to unexpected or unwanted consequences with significant impact on healthcare costs. Paediatric oncology is not exempted from this issue [1], facing challenges such as unfavourable or undesired treatment outcomes (e.g., adverse side effects, unsuccessful results, relapse) among others. Precision medicine, a.k.a. personalised medicine, holds the promise to help address these challenges paving the way to next-generation medicine [2], especially in paediatric oncology. Precision medicine is defined as a patient-centred medical model that uses individuals' phenotypes and genotypes (e.g. molecular profiling, clinical and lifestyle information) to deliver timely and targeted preventive and therapeutic solutions [3]. In paediatric oncology, several approaches to precision medicine have been attempted in the last decades, including specific clinical sequencing studies [4] and clinical trials [5,6] as well as specific strategies for managing medical registries [7] and treatment access [8].

An emerging aspect of precision medicine is the uptake of Artificial Intelligence (AI) solutions able to reveal patterns to predict intervention outcomes for individual patients as well as leverage and enhance the digital transformation of healthcare. Among AI applications for precision medicine, biomedical data synthesis is recognized as a fundamental tool to realize actionable personalisation, although quality standards to guarantee operational validity still need to improve [9]. Typically, synthetic data is produced through computational models generating simulations that approximate real-world data.

Synthetic data generation is emerging as a dominant AI solution for personalised medicine as it enables to address several critical challenges, such as creating the data volumes needed to deliver accurate results, correcting possible biases engrained in real-world data, and complying with increasingly restrictive privacy regulations. These challenges are of utmost importance for precision medicine and tightly intertwined, especially considering the need of data augmentation and privacy preservation for those groups of individuals that are underrepresented in the current landscape of cancer data due to rarity of condition, such as rare paediatric tumours, or discrimination, such as specific demographic strata (age, sex, gender, race) [10].

Among several approaches, deep generative models, such as Variational Autoencoders (VAE) [11], which is an unsupervised representation learning technique, is largely used for synthetic data generation. The idea behind a generative model, such as VAE, is that a latent (unseen) process generates the observed data and a few explanatory factors of its variation, called generative factors, can be identified. The disentangled latent space identified by VAE, which is expected to capture human-understandable dimensions of variation of the data, can be used to generate explainable synthetic instance and even navigated to infer probabilistic trajectories, as recently proposed for single-cell transcriptomics analysis [12].

In this deliverable, we describe a VAE that BSC has developed and applied to the study of medulloblastoma, an aggressive childhood brain tumour. Both the draft of the unpublished article reporting on this work and the medulloblastoma synthetic datasets that have been generated are available at the iPC Nextcloud public repository https://data.ipc-project.bsc.es/s/sqc3WWQTLgC8iAb. Specifically, the code of the VAE is available at https://github.com/bsc-life/Explainable_Synthetic_Data_Generation_Medulloblastoma.

Chapter 2 An explainable deep generative model for paediatric cancer research

2.1 Variational AutoEncoder (VAE)

A generative model aims to model the hidden process that generated the observed data by identifying explanatory factors of its variation, called *generative factors*. The obtained generative process can then be used to generate reliable synthetic instances. The Variational AutoEncoder (VAE) [13] models such generative process in a probabilistic way by relying on a lower-dimensional representation of the data, called *latent space*, whose dimensions, or *latent variables*, are described as known probability distributions and are assumed to map to the generative factors. For ease of reference, we denote the latent factor distributions describing the latent space as *encoding vectors*.

VAE consists of an encoder, which encodes the generative factors of the data into the latent space, and a decoder, which samples the latent space to generate new synthetic instances (Figure 1). Additionally, a latent space is called *disentangled* if each of its latent factors (or sets of covarying latent factors) maps to a specific generative factor. Disentanglement is key to model explainability as it is supposed to capture independent dimensions of variation of the data that are expected to correspond to human-understandable concepts or attributes [14]. However, if the generative factors are unknown, a quantitative evaluation of disentanglement quality is challenging and, indeed, it is an important open research direction [15].

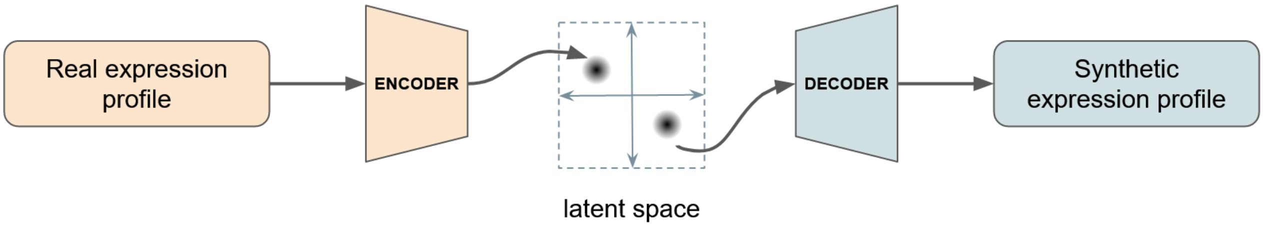


Figure 1: Variational autoencoder for synthetic data generation. The diagram depicts the encoding of a real gene expression profile into the latent space and the decoding of a synthetic gene expression profile from it.

2.2 Data augmentation and interpolation

We developed a VAE that exhibits two types of data generation processes: the *data augmentation* and the *data interpolation* (Figure 2). Data augmentation creates new data by using randomly chosen encoding vectors allowing increasing the size of the available data. Data interpolation creates new data by traversing the latent space from a source to a destination allowing the study of transitions or intermediate states in the data. Both types of data synthesis are equipped with confidence scores that assess the reliability of the generated synthetic instances based on the proximity to the training data points. Additional details on data pre-processing, model architecture, training scheme, and evaluation metrics are provided in the draft of the unpublished paper that is available at the iPC Nextcloud public repository: https://data.ipc-project.bsc.es/s/sqc3WWQTLgC8iAb.

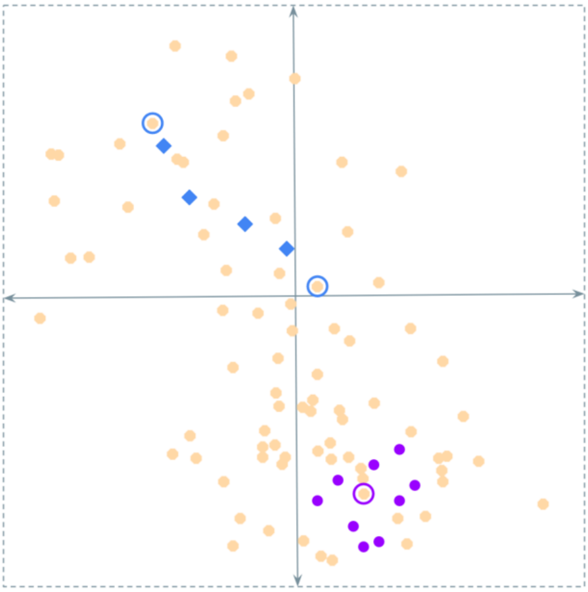


Figure 2: Synthetic data generation by augmentation and interpolation. Real expression profiles are represented in two dimensions as yellow dots. Purple dots represent synthetic expression profiles generated by augmentation targeting a real expression profile (circled in purple). Blue squares represent synthetic expression profiles generated by interpolation between two real expression profiles (circled in blue).

2.3 Model explainability

To explain how and why the VAE is generating synthetic data, it is crucial not only to identify the most relevant latent variables identified by the model but also to puzzle out how it is building them directly from the data. This task represents a depth step into the explainability of VAE. To do so, we propose an explainability approach based on a proxy that leads to the selection of subsets of genes driving the structure of medulloblastoma subgroups (Figure 3). The proxy consists of two steps. In the first step, we classify the four medulloblastoma subgroups on the latent space using XGBoost [16], a gradient boosting algorithm on decision trees. This enables the identification of the most relevant latent variables. In the second step, we project this information onto the original input data so to identify the genes that influence the most those latent variables. This attempt of spotting the most relevant genes by proxy is based on the Shapley Additive Explanations (SHAP) algorithm [17]. In particular, we use two concatenated SHAP-based techniques, namely the SHAP Tree Explainer, used to analyze the XGBoost employed in the latent space, and the DeepLIFT SHAP, used to analyze the encoder network that compresses the input data into that latent space.

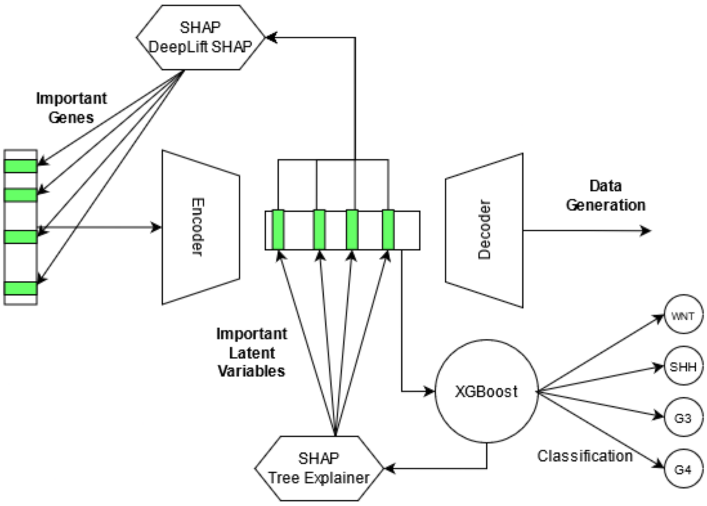


Figure 3: The more relevant genes for accurate synthetic data generation are found by proxy using a SHAP-based approach. The SHAP Tree Explainer technique applied to the XGBoost classifier enables to detect the latent variables that are more relevant to the classification of the four medulloblastoma subgroups (WNT, SHH, G3, G4). Hereafter, the DeepLIFT SHAP technique applied to the encoder networks enables to detect the genes that are more relevant in the definition of those latent variables.

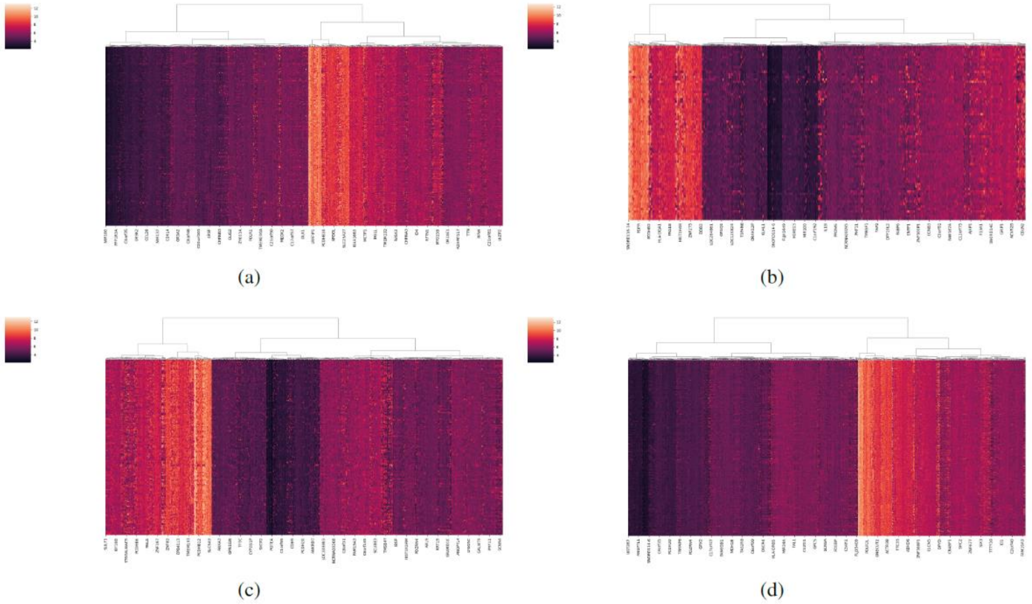
Chapter 3 Application to medulloblastoma

3.1 Expression signatures of synthetic data generation in medulloblastoma

Medulloblastoma is an embryonal tumour that develops in the posterior fossa, is among the most common malignant childhood CNS tumours [18] with an annual incidence of about 5 cases per 1 million individuals [19]. Four distinct medulloblastoma subgroups have been identified (Wingless [WNT], Sonic hedgehog [SHH], Group 3 [G3], and Group 4 [G4]), although the subgroups G3 and G4 are the less characterized molecularly [20].

For this deliverable, BSC produced 4,000 synthetic gene expression profiles of medulloblastoma by training and testing a VAE using the two largest microarray datasets of medulloblastoma stored in the R2 platform [21] with Gene Expression Omnibus (GEO) accession identifiers GSE85217 [22] (18,479 genes x 763 observations) and GSE37382 [23] (18,473 genes x 285 observations). Due to its proportions and independence, the two datasets were used as training and test sets, respectively. The synthetic dataset is included in the iPC Nextcloud public repository: https://data.ipc-project.bsc.es/s/sqc3WWQTLgC8iAb.

By applying our explainability pipeline (see section 2.3 “Model explainability”), we identified 881 genes that contribute the most to the latent variables that best classify among all the subgroups of medulloblastoma. In all of them, a structure consisting of two clusters of up-regulated and down-regulated genes can be appreciated, representing a distinct gene expression signature of each medulloblastoma subgroup (Figure 4).



(a)

(b)

(c)

(d)

Figure 4: Clustermap of SHH (a), WNT (b), G3 (c) and G4 (d) subgroups based on the 881 genes that contribute the most to the latent variables that better classify between all the subgroups of the medulloblastoma.

3.2 The G3-G4 subgroup of medulloblastoma

A number of computational [24,25] and experimental [26,27] studies have pointed out the possible existence of an unknown relationship between G3 and G4, two of the most frequent medulloblastoma subgroups, with the poorest prognosis and less characterized molecular alterations. In particular, G4 has been reported to be the more heterogeneous among the four subgroups with molecular markers being shared with G3 and SHH [28]. A marked overlapping area between G3 and G4 can be appreciated in the visualization of the distribution of the medulloblastoma expression profiles (Figure 5). Studying this overlapping area could shed light on actionable insights to improve therapeutic strategies and survival rates of both G3 and G4, for which targeted therapies have not been identified yet or only suggested. For this reason, we propose to leverage the generative power of the VAE and our explainability pipeline to study this G3-G4 relationship.

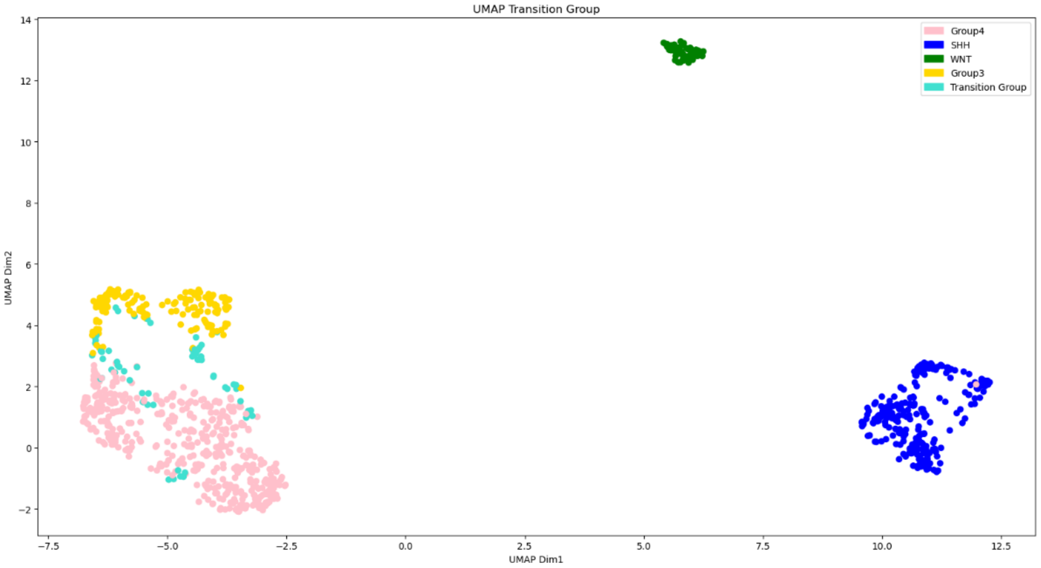


Figure 5: UMAP visualization of the data highlighting the boundary points (labelled as "Transition group") between G3 and G4 in the training dataset (GSE85217).

To characterize the overlap between G3 and G4 subgroups, we build a k-nearest neighbours graph (k-NNG) based on the G3 and G4 observations. Boundary points are defined as the points of a subgroup that have points of the other subgroup among their nearest neighbours. By varying the k value of the k-NNG, the amount of boundary points varies. The lower the k value, the more restricted is the set of points considered as boundary points.

Boundary points, henceforth called Transition group, correspond to observations located in the overlapping region between G3 and G4 subgroups. As those observations lay in a peculiar patch of the latent space, we expect them to have intermediate gene expression levels compared to the remaining G3 and G4 observations. To test this hypothesis, we leverage the data augmentation feature of our VAE to generate synthetic instances uniformly across the G3 and G4 area and vary the k value to identify genes that are differential expressed in G3, G4 and the Transition group. The optimized condition corresponds to parameter $k = 3$.

We perform the non-parametric Kruskal-Wallis and Dunn's statistical tests with Bonferroni correction for multiple comparison to detect those genes of the 881 previously selected that show a statistically relevant difference in expression among G3, G4 and the Transition group (Figure 6). We obtain a list of 26 genes, namely ALX1, BCAR3, C21orf90, CALB1, CCDC141, CCNA1, CDH4, CRABP1, DLX5, DUSP5P, DYRK4, EDA2R, FAM183A, FCN3, FLJ37786, GFRA1, HSPA1L, IFI44, IRX3, KLF8, LOC391169, LOC440173, LRRC4C, MTUS2, NEUROG2, PARP8, SEPT12, SKAP2, SLC13A4, SLC30A3, SLFN13, SREBF1, SULT1A2, TMEM220, TTN and VWA5A.

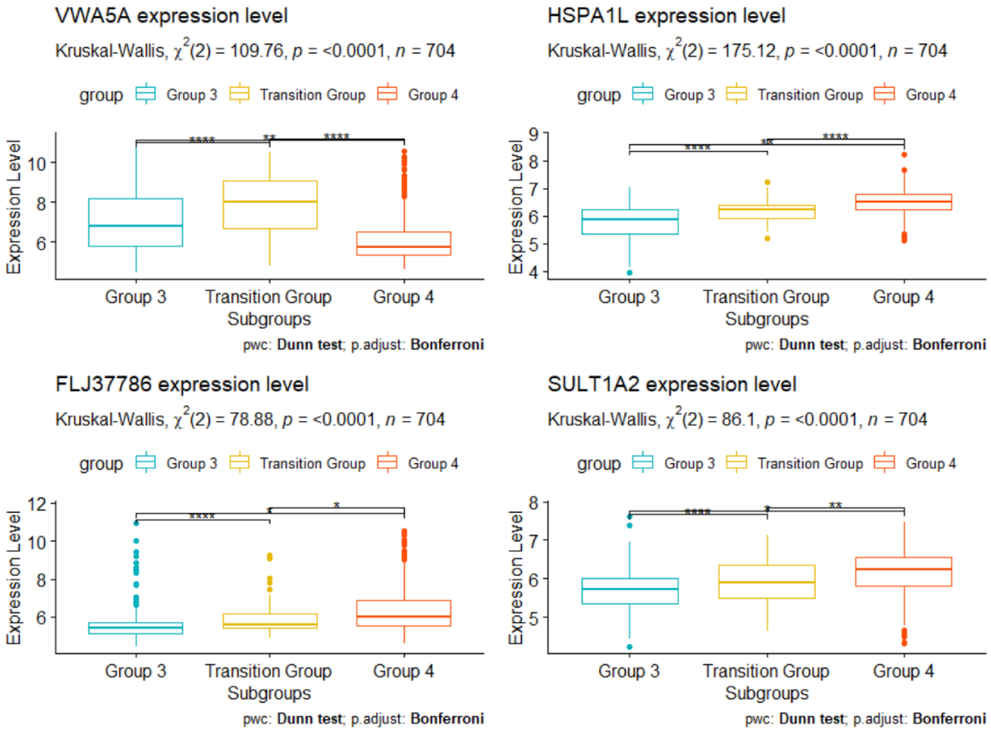


Figure 6: VWA5A, HSPA1L, FLJ37786 and SULT1A2 are illustrative genes with statistically significant differences in their expression levels among G3, G4 and the Transition Group.

Chapter 4 Summary and Conclusion

An emerging aspect of precision medicine, especially in the domain of paediatric cancer research, is the uptake of AI solutions able to reveal patterns to predict personalised intervention outcomes for subgroups of patients. Among the AI applications for paediatric oncology, data synthesis is recognized as a fundamental approach to realize actionable personalisation, especially deep generative models, such as Variational AutoEncoder (VAE), which is an unsupervised representation learning technique that is largely used for synthetic data generation in many areas.

In this deliverable, we developed an explainable VAE for synthetic transcriptomics data generation in medulloblastoma, a childhood brain tumour, trained and tested using the largest microarray datasets of medulloblastoma stored in the R2 platform. We used the model to generate a datasets of 4,000 high fidelity synthetic medulloblastoma expression profiles encompassing all four medulloblastoma subgroups (WNT, SHH, G3, G4), identifying 881 genes that contribute the most to the latent variables that better classify them. We also leveraged the model to study the unknown relationship between G3 and G4 subgroups, identifying 26 genes that show a statistically relevant difference in expression among G3, G4, and a potential intermediate subgroup.

Our VAE can be adapted to other paediatric cancers and the resulting synthetic datasets be used for testing and training patient, cancer, and drug models in other workpackages of the iPC project.

List of Abbreviations

Abbreviation	Translation
VAE	Variational AutoEncoder

Bibliography

[1] DuBois SG, Corson LB, Stegmaier K, Janeway KA. Ushering in the next generation of precision trials for pediatric cancer. Science. 2019;363(6432):1175-1181. doi:10.1126/science.aaw4153

[2] Nimmesgern E, Norstedt I, Draghia-Akli R. Enabling personalized medicine in Europe by the European Commission’s funding activities. Personalized Medicine 2017;14:355–65. https://doi.org/10.2217/pme-2017-0003.

[3] Council conclusions on personalised medicine for patients. EUR-Lex - 52015XG1217(01) - EN - EUR-Lex 2015. https://eur-lex.europa.eu/

[4] Mody RJ, Prensner JR, Everett J, Parsons DW, Chinnaiyan AM. Precision medicine in pediatric oncology: Lessons learned and next steps. Pediatr Blood Cancer. 2017;64(3):10.1002/pbc.26288. doi:10.1002/pbc.26288

[5] Renfro LA, An MW, Mandrekar SJ. Precision oncology: A new era of cancer clinical trials. Cancer Lett. 2017;387:121-126. doi:10.1016/j.canlet.2016.03.015

[6] Cohen JW, Akshintala S, Kane E, et al. A Systematic Review of Pediatric Phase I Trials in Oncology: Toxicity and Outcomes in the Era of Targeted Therapies. Oncologist. 2020;25(6):532-540. doi:10.1634/theoncologist.2019-0615

[7] van Tilburg, Cornelis M; Pfaff, Elke; Pajtler, Kristian W; et al; Gerber, Nicolas U (2021). The Pediatric Precision Oncology INFORM Registry: Clinical Outcome and Benefit for Patients with Very High-Evidence Targets. Cancer Discovery, 11(11):2764-2779.

[8] Shulman DS, Kiwinda LV, Edwards S, et al. Retrospective evaluation of single patient investigational new drug (IND) requests in pediatric oncology [published online ahead of print, 2021 Mar 9]. Cancer Med. 2021;10(7):2310-2318. doi:10.1002/cam4.3791

[9] Chen J, Chun D, Patel M, Chiang E, James J. The validity of synthetic clinical data: a validation study of a leading synthetic data generator (Synthea) using clinical quality measures. BMC Med Inform Decis Mak. 2019;19(1):44. Published 2019 Mar 14. doi:10.1186/s12911-019-0793-0

[10] Cirillo D, Núñez-Carpintero I, Valencia A. Artificial intelligence in cancer research: learning at different levels of data granularity. Molecular Oncology 2021;15:817–29. https://doi.org/10.1002/1878-0261.12920.

[11] Kingma DP, Welling M. Auto-Encoding Variational Bayes. ArXiv:13126114 [Cs, Stat] 2014.

[12] Yang KD, Damodaran K, Venkatachalapathy S, Soylemezoglu AC, Shivashankar GV, Uhler C. Predicting cell lineages using autoencoders and optimal transport. PLOS Computational Biology 2020;16:e1007828. https://doi.org/10.1371/journal.pcbi.1007828.

[13] Kingma DP, Welling M. Auto-Encoding Variational Bayes. ArXiv:13126114 [Cs, Stat] 2014.

[14] Y. Bengio, A. Courville and P. Vincent, "Representation Learning: A Review and New Perspectives," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 35, no. 8, pp. 1798-1828, Aug. 2013, doi: 10.1109/TPAMI.2013.50.

[15] Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, Olivier Bachem. Challenging Common Assumptions in the Unsupervised Learning of

Disentangled Representations. Proceedings of the 36th International Conference on Machine Learning, PMLR 97:4114-4124, 2019.

[16] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16). Association for Computing Machinery, New York, NY, USA, 785–794. DOI:https://doi.org/10.1145/2939672.2939785

[17] Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17). Curran Associates Inc., Red Hook, NY, USA, 4768–4777.

[18] Northcott PA, Robinson GW, Kratz CP, et al. Medulloblastoma. Nat Rev Dis Primers. 2019;5(1):11. Published 2019 Feb 14. doi:10.1038/s41572-019-0063-6

[19] Ostrom QT, Gittleman H, Truitt G, Boscia A, Kruchko C, Barnholtz-Sloan JS. CBTRUS Statistical Report: Primary Brain and Other Central Nervous System Tumors Diagnosed in the United States in 2011-2015 [published correction appears in Neuro Oncol. 2018 Nov 17;:null]. Neuro Oncol. 2018;20(suppl_4):iv1-iv86. doi:10.1093/neuonc/noy131

[20] Juraschka K, Taylor MD. Medulloblastoma in the age of molecular subgroups: a review. J Neurosurg Pediatr. 2019;24(4):353-363. doi:10.3171/2019.5.PEDS18381

[21] R2: Genomics Analysis and Visualization Platform (http://r2.amc.nl)

[22] Ramaswamy V, Taylor MD. Bioinformatic Strategies for the Genomic and Epigenomic Characterization of Brain Tumors. Methods Mol Biol. 2019;1869:37-56. doi:10.1007/978-1-4939-8805-1_4

[23] Northcott PA, Shih DJ, Peacock J, et al. Subgroup-specific structural variation across 1,000 medulloblastoma genomes. Nature. 2012;488(7409):49-56. doi:10.1038/nature11327

[24] Castillo-Rodríguez RA, Dávila-Borja VM, Juárez-Méndez S. Data mining of pediatric medulloblastoma microarray expression reveals a novel potential subdivision of the Group 4 molecular subgroup. Oncol Lett. 2018;15(5):6241-6250. doi:10.3892/ol.2018.8094

[25] Núñez-Carpintero I, Petrizzelli M, Zinovyev A, Cirillo D, Valencia A. The multilayer community structure of medulloblastoma. iScience. 2021;24(4):102365. Published 2021 Mar 26. doi:10.1016/j.isci.2021.102365

[26] Gajjar AJ, Robinson GW. Medulloblastoma-translating discoveries from the bench to the bedside. Nat Rev Clin Oncol. 2014;11(12):714-722. doi:10.1038/nrclinonc.2014.181

[27] Gajjar A, Bowers DC, Karajannis MA, Leary S, Witt H, Gottardo NG. Pediatric Brain Tumors: Innovative Genomic Information Is Transforming the Diagnostic and Clinical Landscape. J Clin Oncol. 2015;33(27):2986-2998. doi:10.1200/JCO.2014.59.9217

[28] Kool M, Korshunov A, Remke M, et al. Molecular subgroups of medulloblastoma: an international meta-analysis of transcriptome, genetic aberrations, and clinical data of WNT, SHH, Group 3, and Group 4 medulloblastomas. Acta Neuropathol. 2012;123(4):473-484. doi:10.1007/s00401-012-0958-8