



Privacy compliant health data as a service for AI development

Grant Agreement Number: 101095384

Deliverable D3.3: Technologies for de-identification and synthetic data generation v1

Deliverable Identifier:	D3.3
Deliverable Version:	v1.0
Status	Final (F)
Work Package:	WP3 Data as a Service, DaaS
Task:	Task 3.2 De-identification technologies (including anonymization), Task 3.3. Synthetic data generation Task 3.4. Data quality metrics
Author(s) and Organisation:	Ileana Montoya Perez (UTU), Andrei Kazlouski (UTU), Parisa Movahedi (UTU), Paula Subías-Beltrán (EUT), Rafael Redondo (EUT), Roger Marí (EUT), Carla Pitarch (EUT), Christos Chatzichristos (KU Leuven), CongTing Lai (KU Leuven), Artur Rocha (INESC TEC), Helder F. Oliveira (INESC TEC), Tunc Asuroglu (VTT), Juha Pajula (VTT), Aino Alahäivälä (VTT), Jussi Salmi (Turku UAS)
Peer Reviewer(s):	Turku UAS, Sabanci University
Deliverable Due Date:	2024/10/31
Deliverable Submission Date:	2024/10/23
Dissemination Level:	PU: Public
Funding Authority:	European Commission
Funding Program:	Horizon Europe Health Work Programme 2021 – 2022
Topic:	HORIZON-HLTH-2022-IND-13-02
Rights:	PHASE-IV-AI Consortium

Document Control History

Version	Date	Edited by	Modification reason
v.0.1	2024/05/18	UTU	1 st draft
v.0.2	2024/06/03	EUT	Proposal of contributions on synthetization services
v.0.3	2024/09/09	KUL	Proposal of contributions on quality metrics
v.0.4	2024/09/25	UTU	For internal review
v.0.5	2024/10/08	UTU	After internal review
v.1.0	2024/10/23	UTU	Final version
v.1.0_amended	2025/07/10	UTU	Amended version

Executive Summary

This deliverable provides an initial review and benchmark of state-of-the-art technologies for de-identification, synthetic data generation, and evaluation. It covers available technologies for medical images and tabular data (e.g., electronic health records and longitudinal data). Benefits and limitations of these technologies are also analysed. Synthetic data generators can be categorized in statistical and machine learning. The former usually are applied to tabular data, while the latter can be applied to both tabular and image data. In this document, examples of these methods are explained. In addition, differential privacy is presented as the gold standard for anonymizing data as it provides mathematical guarantees for privacy and it can be ubiquitously applied to different medical data types. The utility-privacy tradeoff should always be considered before applying any data de-identification/synthesis technology in practice. Therefore, to evaluate the quality of synthetic data, the deliverable outlines and explains a set of metrics for privacy, realism (fidelity, and utility).

Disclaimer

The content of the publication herein is the sole responsibility of the publishers and it does not necessarily represent the views expressed by the European Commission or its services.

While the information contained in the documents is believed to be accurate, the authors(s) or any other participant in the PHASE-IV-AI consortium make no warranty of any kind with regard to this material including, but not limited to the implied warranties of merchantability and fitness for a particular purpose.

Neither the PHASE-IV-AI Consortium nor any of its members, their officers, employees or agents shall be responsible or liable in negligence or otherwise howsoever in respect of any inaccuracy or omission herein.

Without derogating from the generality of the foregoing neither the PHASE-IV-AI Consortium nor any of its members, their officers, employees or agents shall be liable for any direct or indirect or consequential loss or damage caused by or arising from any information advice or inaccuracy or omission herein.

Copyright message

©PHASE-IV-AI Consortium, 2023-2026. This deliverable contains original unpublished work except where clearly indicated otherwise. Acknowledgement of previously published material and of the work of others has been made through appropriate citation, quotation or both. Reproduction is authorised provided the source is acknowledged.

Table of Contents

1. Introduction	8
1.1 Purpose of the document.....	8
1.2 Structure of the document.....	8
1.3 List of Acronyms.....	9
2. Pseudonymization and anonymization methods	10
2.1 Basic definitions.....	10
2.1.1 Pseudonymization.....	10
2.1.2 Anonymization	10
2.2 Basic guidelines for pseudonymization	10
2.3 Anonymization tools for medical imaging data	11
2.4 Synthetic data as tool for anonymization.....	13
2.5 Differential Privacy	14
2.6 Differentially Private Synthetic Data.....	15
3. Synthetic data generation methods.....	17
3.1 Methods for tabular data.....	17
3.2 Methods for longitudinal data	21
3.3 Methods for medical imaging data.....	21
3.3.1 Synthetization architectures: GANs, Diffusion Models and Transformers.....	22
3.3.1.1 GAN-based medical image synthesis.....	22
3.3.1.2 Diffusion Models superiority in medical imaging	22
3.3.1.3 Transformers and DMs	23
3.3.2 Conditional Image Synthesis in DMs	24
3.3.2.1 Mask-based conditional synthesis.....	24
3.3.2.2 Context-based conditional synthesis.....	25
3.3.2.3 EHR/Metadata conditional synthesis	26
3.3.2.4 Modality-based translation synthesis	27
3.3.3 Volume Synthesis: from 2D to 3D	28
3.3.3.1 3D-from-2D approaches.....	28
3.3.3.2 3D synthesis	29
3.3.4 Clinical applications of image synthesis	29
3.3.4.1 Synthetic chest CT nodule insertion	29
3.3.4.2 Lung cancer evolution prediction	30
4. Quality metrics for synthetic data	33

4.1	Privacy	34
4.2	Realism - Fidelity	35
4.2.1	Qualitative	36
4.2.2	Quantitative	36
4.2.2.1	Univariate Metrics'	37
4.2.2.2	Multivariate Metrics'	37
4.2.2.3	Space Representations.....	37
4.2.2.4	Knowledge Violation and Association Rule Mining	38
4.2.2.5	Discriminant Models.....	38
4.3	Realism - Utility	38
4.3.1	Cross-Classification Metrics	39
4.3.2	Model Sensitivity.....	39
4.2.3	Downstream Tasks.....	39
4.4	Variety and authenticity	41
5.	DaaS Technologies Integration.....	42
5.1	DaaS Toolbox Integration.....	42
5.2	Data and service operatibility.....	42
5.2.1	Operability for Tabular Data	43
5.2.2	Operability for Image Data	43
5.3	DaaS and MaaS Toolbox Integration	43
5.3.1	Shared Metadata Structures and Card Framework	44
5.3.2	Cross-Linking Between Toolboxes.....	44
5.3.3	Co-Development and Extensibility.....	44
5.3.4	Use Case-Driven Organization and Semantic Alignment	44
5.4	Interfacing via Health Data Hub (HDH).....	44
5.4.1	Model Execution and Access Interfaces	45
5.4.2	Metadata Flow and Alignment with HDH Catalog	45
5.4.3	Deployment and Access Interface via the Portal and Marketplace.....	45
6.	Conclusion.....	46

Table of Figures

Figure 1: Illustration of steps in text removal. (left) Text detection and potentially recognized, text cut-out (middle), image inpainting (right).....	12
Figure 2: Illustration of different de-facing methods (source Schwarz, 2021).....	13
Figure 3: Distribution of the continuous variables in the real dataset compared to their distribution in the synthetic dataset generated by: a) SDV GaussianCopula, b) AIM ($\epsilon = 1$), c) AIM ($\epsilon = 10$).....	19
Figure 4: Distribution of the discrete variables in the real dataset compared to their distribution in the synthetic dataset generated by: a) SDV GaussianCopula, b) AIM ($\epsilon = 1$), c) AIM ($\epsilon = 10$).	20
Figure 5: Spearman correlation coefficients for pairs of continuous and discrete variables in the real dataset, SDV GaussianCopula synthetic dataset, AIM synthetic dataset with a privacy level of $\epsilon = 1$, and AIM synthetic dataset with a privacy level of $\epsilon = 10$	20
Figure 6: Basic pipeline diagram of diffusion models (left) and qualitative comparison of various synthesization methods. Source: (Müller-Franzes et al. 2023).....	23
Figure 7: Illustration of a diffusion model process conditioned on 3D segmentation masks for brain MRI synthesis. Source: (Dorjsembe et al. 2024).....	25
Figure 8: Illustration of a super resolution diffusion workflow conditioned on the metadata variables age, gender, ventricular volume, and brain volume. Source: (Wang et al. 2023).	27
Figure 9: Scheme of a GAN model to insert synthetic nodules in context based on attributes such as size and malignancy Source: (Nishio et al. 2020).....	30
Figure 10: Example of white matter hyperintensities evolution in brain MRI. Source: (Rachmadi et al. 2020).....	32
Figure 11: Synthetic data generation pipeline.....	33
Figure 12: Different types of metrics for quality assessment of synthetic data.....	34

Table of Tables

Table 1: Snippet from an actual publicly available dataset before anonymization. All data, except for the column names, has been synthesized to avoid revealing information about real people	11
Table 2: Pseudonymized version of the data snippet using data masking (Telefono → Telephone), generalization (Nato → Age), and suppression (Cognome, Nome → Pseudonym).....	11
Table 3: The dataset is managed by a trusted entity that provides statistical reports based on the data. If an adversary is aware that a specific individual has been removed from the dataset, it could deduce sensitive attributes (e.g., salary range) from the unaltered data. However, by applying differential privacy, the addition of noise to the outputs makes them indistinguishable, thereby protecting the privacy of the participants.	15
Table 4: List of fidelity and utility metrics.	36

1. Introduction

Technologies for de-identification and synthetic data generation are the enabling components for privacy preservation of medical images and Electronic Health Records (EHR) in PHASE IV AI's Data as a Service (DaaS) solution. Privacy-preserving synthetic data that maintains the structure and statistical properties of the original data allows sharing individual-level data while safeguarding the privacy of data subjects.

Privacy-preserving synthetic data serves several purposes, including ensuring privacy compliance, augmenting existing datasets, enabling more robust machine learning models, and facilitating data sharing for research and development. It can be used to generate additional data for training models, enhancing their performance, and addressing data scarcity.

However, generating synthetic datasets involves a well-known trade-off between utility and privacy. Generic synthetic datasets are only suitable for a limited range of applications. Evaluating the quality of synthetic datasets is challenging, as there is no general consensus on the metrics to be applied for privacy and utility assurance. This challenge is especially pronounced in the case of synthetic medical images, where suitable metrics for evaluation are scarce.

To address the limitations of generic or all-purpose synthetic datasets, DaaS aims to provide synthetic data on demand, optimizing data quality for predetermined purposes. Confidentiality will be ensured using data synthesis methods that offer verifiable privacy guarantees, such as differential privacy. DaaS will also define a set of metrics to evaluate the quality of synthetic data. The utility of synthetic datasets will be validated through three real-world use cases (Objective 5, WP6).

1.1 Purpose of the document

The purpose of this deliverable, as part of WP3 DaaS, is to provide detailed information on state-of-the-art methods and techniques for de-identification, privacy-preservation, synthetic data generation and evaluation. This document explains how these methods generate privacy-compliant individual-level data to support the development of innovative health technologies. It offers background information on the techniques and methods used in DaaS to achieve reliable and privacy-preserving datasets, including methods for generating synthetic data from images and tabular data. Review and initial benchmark of different methods are presented to illustrate their implementation and performance.

1.2 Structure of the document

This document is structured to provide a comprehensive overview of the concepts and methods related to the de-identification of sensitive data. It begins with an introduction to pseudonymization, anonymization, and other related techniques used in the de-identification process.

Following this, the document describes various methods for generating synthetic data, including approaches for both tabular data (such as electronic health records) and medical images. It then covers metrics for assessing the quality of synthetic data, ensuring data realism, privacy, and utility, thereby enhancing the effectiveness and reliability of the synthetic data. The document concludes with final remarks.

1.3 List of Acronyms

List of Acronyms	
AIM	Adaptive and Iterative Mechanism
CNN	Convolutional neural network
DM	Diffusion Model
DP	Differential Privacy
HER	Electronic Health Record
GAN	Generative Adversarial Networks
GDPR	General Data Protection Regulation
PSA	Prostate-Specific Antigen
SDV	Synthetic Data Vault
SVM	Support Vector Machines
VAE	Variational autoencoders

2. Pseudonymization and anonymization methods

Pseudonymization and anonymization are critical techniques in data protection and privacy, used to safeguard personal information while maintaining data utility. These methods are essential for compliance with legal standards like EU's General Data Protection Regulation (GDPR), protecting individuals' identities from unauthorized access. Despite their importance, there's frequent confusion between the two, with many mistakenly believing pseudonymized data is fully anonymous. This misconception, along with the intertwined usage of these terms, highlights the need for clear understanding and distinction to ensure proper data handling and compliance.

2.1 Basic definitions

The sections below offer an overview of both pseudonymization and anonymization, along with a brief description of the methods associated with each. Since these terms can be defined in various ways in the literature, the following section defines how they will be used within the PHASE-IV AI Project and its components. The principle difference between pseudonymization and anonymization is that, in the latter, there is no way to trace the data back to the individuals, not even for the person or entity who performed the procedure.

2.1.1 Pseudonymization

Pseudonymization involves data de-identification techniques where directly identifiable information is replaced with pseudonyms or artificial identifiers. This process allows data to be used without revealing individuals' identities, thereby enhancing privacy. However, pseudonymized data can be re-identified if the pseudonyms are linked back to the original identities through additional information or keys. Therefore, it is crucial to manage quasi-identifiers — pieces of information that are not unique on their own but can be combined with other quasi-identifiers to uniquely identify individuals. Examples of quasi-identifiers include birth dates, postal codes, and demographics. Properly handling quasi-identifiers is essential to prevent re-identification risks. Basic pseudonymization techniques include data masking, generalization, and suppression.

2.1.2 Anonymization

Anonymization is the process of removing or modifying personal data to make re-identifying individuals difficult, either directly or indirectly. Unlike pseudonymization, anonymization provides a higher level of privacy protection because the data cannot be re-linked to the original individuals, even if additional information becomes available. An advanced technique that has become a gold standard in anonymization is differential privacy. Differential privacy provides strong mathematical guarantees that individual data points cannot be re-identified by adding controlled random noise to the data. This ensures that the inclusion or exclusion of a single data point does not significantly affect the outcome of any analysis, thus protecting individual privacy. For more information refer to Section 2.5.

2.2 Basic guidelines for pseudonymization

Before implementing pseudonymization techniques, a data analyst must first assess which features in the dataset may serve as quasi-identifiers—attributes that can potentially be used to re-identify individuals when combined with other data. Identifying these quasi-identifiers is crucial for ensuring data privacy, as they require modification. The extent of these modifications can vary greatly depending on the task at hand, balancing the trade-off between privacy and data utility. While it is very easy to fully anonymize data by substituting all values with random ones or by not releasing any data at all, this approach renders the data completely useless.

Therefore, careful application of pseudonymization techniques is needed. These techniques include data masking, generalization, and suppression. Data masking replaces sensitive information with fictional but realistic data, generalization reduces the granularity of data to make it less specific, and suppression involves omitting sensitive data entirely. Tables 1-2 illustrate these techniques applied to a public dataset for the prostate cancer use case identified within WP6.

Table 1: Snippet from an actual publicly available dataset before pseudonymization. All data, except for the column names, has been synthesized to avoid revealing information about real people

COGNOME	NOME	TELEFONO	NATO	PSA (ng/ml)	GLEASON
Luca	Esposito	3391344993	05/25/1950	4.53	6 (3+3)
Marco	Rossi	3398765432	10/30/1953	13.24	0
Alessandro	Bianchi	3391122334	01/15/1948	19.11	0
Matteo	Romano	3399988776	07/10/1952	4.89	9 (5+4)
Giuseppe	Conti	3394455667	02/20/1957	8.45	0

Table 2: Pseudonymized version of the data snippet using data masking (Telefono → Telephone), generalization (Nato → Age), and suppression (Cognome, Nome → Pseudonym)

PSEUDONYM	TELEPHONE	AGE	PSA (ng/ml)	GLEASON
A8321	339*****	65	4.53	6 (3+3)
B4975	339*****	62	13.24	0
D1098	339*****	67	19.11	0
G5632	339*****	63	4.89	9 (5+4)
K2456	339*****	58	8.45	0

In addition to these basic pseudonymization techniques, more advanced and well-recognized methods can be employed to enhance data privacy. These methods, including k -anonymity¹, l -diversity², and t -closeness³, address the limitations of basic techniques and reduce the probability of successful re-identification.

2.3 Anonymization tools for medical imaging data

Under GDPR, anonymization includes removing direct and indirect personal identifiers such as names, addresses, phone numbers, photographs, and other unique characteristics to ensure individuals cannot be identified. This process makes the data no longer subject to GDPR, allowing businesses to use it more freely.

Images are considered direct identifiers under GDPR, meaning they must be completely anonymized to comply with privacy regulations. This means the images should be altered or masked in a way that makes it impossible to identify the individual.

Effective privacy-preserving techniques include data masking, generalization, data swapping, data perturbation, and the creation of synthetic data⁴. For images specifically, this could involve techniques like

¹ Sweeney L. k -anonymity: A model for protecting privacy. International journal of uncertainty, fuzziness and knowledge-based systems. 2002 Oct;10(05):557-70.

² Machanavajjhala A, Kifer D, Gehrke J, Venkatasubramanian M. l -diversity: Privacy beyond k -anonymity. Acm transactions on knowledge discovery from data (tkdd). 2007 Mar 1;1(1):3-es.

³ Li N, Li T, Venkatasubramanian S. t -closeness: Privacy beyond k -anonymity and l -diversity. In 2007 IEEE 23rd international conference on data engineering 2006 Apr 15 (pp. 106-115). IEEE.

⁴ Diaz O, Kushibar K, Osuala R, et al. Data preparation for artificial intelligence in medical imaging: A comprehensive guide to open-access platforms and tools. Physica medica. 2021 Mar 1;83:25-37.

blurring, pixelation, cut-outs or replacing identifiable features with generic substitutes. As for synthetization techniques, it will have a dedicated chapter afterwards, as it is one of the backbone ideas of this project.

To ensure compliance, businesses should conduct motivated intruder tests to simulate potential re-identification attempts and adopt governance structures to oversee and document anonymization processes. This ensures that all personal identifiers in images and other data are effectively anonymized, protecting individual privacy and meeting regulatory requirements.

Medical images often contain sensitive personal information that must be protected to ensure patient privacy. This information can include patient names, dates of birth, medical record numbers, and other identifiers. Although such information is often an overlaid visualization on top of the image, i.e. original values of pixels values can be cleanly recovered, sometimes images are shared with textual annotations embedded in the image in such a way that pixels are altered in an irreversible way. In this case, simply blurring or masking this text is not always sufficient, as advanced techniques can potentially reverse these modifications. Effective de-identification requires robust methods to ensure that all textual information is irreversibly removed or obscured. While doing this, the integrity and diagnostic value of the medical images should be preserved. De-identification processes must ensure that the removal of identifying information does not degrade the quality of the images in medical terms.

Techniques such as edge detection and region-of-interest analysis can help identify and remove sensitive information without affecting the overall image quality. More advanced techniques could not only identify but also recognize text in the image⁵, often referred as OCR (Optical Character Recognition). This text recognition technology is in general mature enough as shown by some solutions close to the market^{6,7}. Although it has some limitations to recognize hand-written scripting.

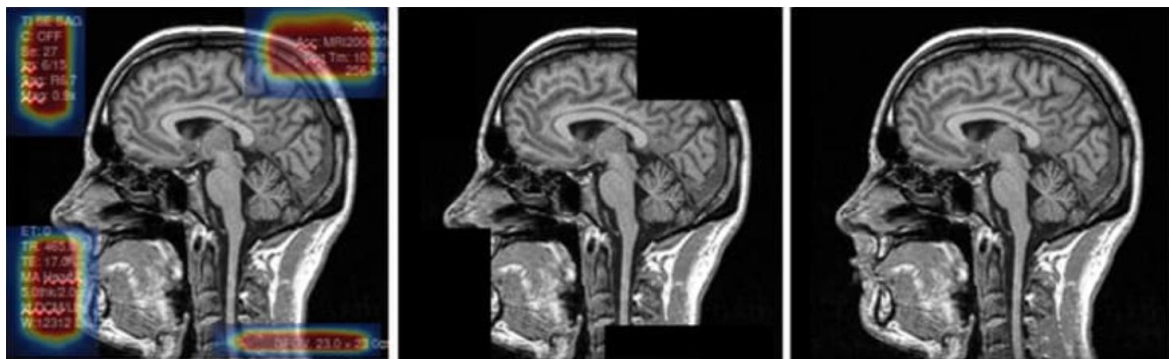


Figure 1: Illustration of steps in text removal. (left) Text detection and potentially recognized, text cut-out (middle), image inpainting (right).

Furthermore, this technique could be cross-validated with personal identifiers included in attached metadata, such as in the DICOM exchange file format. By doing this, advance and more precise methods could be leveraged to just occlude custom identifiers, while leaving others untouched and thus altering the original image as less as possible.

Consistency in de-identification across large datasets is another significant challenge. When dealing with thousands of images, manual de-identification is impractical and prone to errors. Automated de-identification tools must be accurate and reliable, ensuring that all images within a dataset are uniformly processed.

⁵ Gifu, D. (2022). AI-backed OCR in Healthcare. *Procedia Computer Science*, 207, 1134-1143.

⁶ <https://aws.amazon.com/es/blogs/machine-learning/de-identify-medical-images-with-the-help-of-amazon-comprehend-medical-and-amazon-rekognition/>

⁷ <https://cloud.google.com/architecture/de-identification-of-medical-images-through-the-cloud-healthcare-api>

Consistent de-identification helps maintain the utility of the dataset for research while protecting patient privacy.

De-facing technologies in medical imaging are essential tools for anonymizing images and ensuring compliance with privacy regulations. These technologies automatically detect and obscure facial features in medical images such as Magnetic Resonance Imaging (MRI) and Computed Tomography (CT) scans, effectively preventing the re-identification of patients. By using advanced algorithms and machine learning techniques, de-facing tools can accurately identify and mask or replace facial features while preserving the critical diagnostic information in the images. This ensures that the anonymized images remain useful for clinical and research purposes.

Actually, research in de-facing technologies for medical imaging has seen significant advancements, with various scientific publications and software solutions now available⁸.

Pydeface⁹ and *mri_deface*¹⁰, *mask_face*¹¹, *fsl_deface*¹² are tools to remove facial structure from MRI images by altering morphological structures to the point of making the face unrecognizable. A separate mention should be made of the *Reface*¹³ method which adds an additional step in which a synthetic face is attached to the resulting cavity. The growing body of literature demonstrates the efficacy and reliability of these methods, establishing a solid foundation for their widespread adoption in the medical community to protect patient privacy while facilitating data sharing and collaboration.

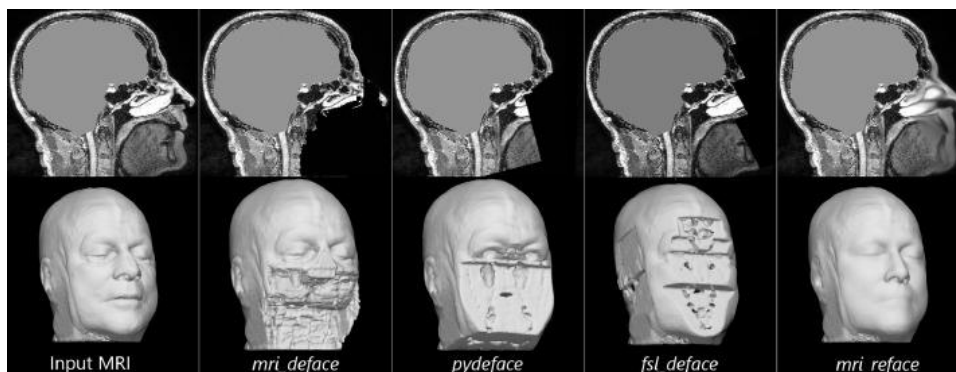


Figure 2: Illustration of different de-facing methods (source Schwarz, 2021).

2.4 Synthetic data as tool for anonymization

Synthetic data is artificially generated to replicate the statistical properties and patterns of real-world data. Unlike real data, which is collected from actual events or transactions, synthetic data is produced by algorithms and models designed to capture the essential characteristics of the original dataset.

⁸ Theyers AE, Zamyadi M, O'Reilly M, Bartha R, Symons S, MacQueen GM, Hassel S, Lerch JP, Anagnostou E, Lam RW, Frey BN. Multisite comparison of MRI defacing software across multiple cohorts. *Frontiers in psychiatry*. 2021 Feb 24;12:617997.

⁹ Gulban OF, Nielson D, Poldrack R, Lee J, Gorgolewski KJ, Vanessasaurus Ghosh S. *poldracklab/pydeface: v2. 0.0*. Zenodo <https://doi.org/10.5281/zenodo.2019;3524401>.

¹⁰ https://surfer.nmr.mgh.harvard.edu/fswiki/mri_deface

¹¹ Milchenko M, Marcus D. Obscuring surface anatomy in volumetric imaging data. *Neuroinformatics*. 2013 Jan;11:65-75.

¹² Alfaro-Almagro F, Jenkinson M, Bangerter NK, Andersson JL, Griffanti L, Douaud G, Sotiropoulos SN, Jbabdi S, Hernandez-Fernandez M, Vallee E, Vidaurre D. Image processing and Quality Control for the first 10,000 brain imaging datasets from UK Biobank. *Neuroimage*. 2018 Feb 1;166:400-24.

¹³ Schwarz CG, Kremers WK, Wiste HJ, Gunter JL, Vemuri P, Spychalla AJ, Kantarci K, Schultz AP, Sperling RA, Knopman DS, Petersen RC. Changing the face of neuroimaging research: Comparing a new MRI de-facing technique with popular alternatives. *NeuroImage*. 2021 May 1;231:117845.

The primary advantage of synthetic data lies in its design to be anonymous. However, the degree to which it achieves true anonymity depends on several factors, including the methods used to generate it and the complexity of the original data. For example, Simple Random Sampling involves creating synthetic data by randomly sampling from the distributions observed in the real dataset. Another example is advanced machine learning models, such as Generative Adversarial Networks (GANs). Simple methods might not capture complex relationships in the data, while complex ones can produce more realistic synthetic data. However, maintaining the statistical properties of the original dataset while ensuring privacy is challenging. Highly accurate synthetic data might inadvertently reveal patterns that could be used to infer sensitive information. Conversely, prioritizing the privacy of individuals can sometimes compromise the dataset's utility for analysis or decision-making. The complexity and diversity of the real data also play a crucial role in achieving anonymity. For less complex datasets, such as simple tabular data with low-dimensional features, achieving anonymity might be relatively simpler. However, complex datasets with many variables or intricate relationships between features present greater challenges for achieving anonymity without losing data utility. The more complex the dataset, the more sophisticated the synthetic data generation methods need to be to ensure privacy.

The risk of re-identification increases if an adversary has access to auxiliary information that can be cross-referenced with the synthetic data. The more information available externally, the harder it is to guarantee anonymity. Even with synthetic data, if attackers have access to multiple datasets, they might perform linkage attacks by finding common patterns or attributes to identify individuals. This highlights the importance of continuously evaluating and improving synthetic data generation methods to protect against such risks.

2.5 Differential Privacy

Differential Privacy (DP) is a mathematical framework that provides privacy guarantees when analyzing and sharing data. In centralized settings, maintaining DP ensures that the inclusion or exclusion of a single user does not significantly affect the analysis outcome beyond a specified ϵ (epsilon) value. DP remains unaffected by quasi-identifiers within the data or any additional information an adversary might possess because the anonymized dataset is not disclosed. Instead, the data controller releases responses to statistical queries about the data. Table 3 illustrates how adding noise to the output of the queries may hide the absence of a user in the dataset.

Formally, a randomized algorithm \mathcal{A} is (ϵ, δ) -differentially private if for all datasets D_1 and D_2 that differ in at most one record (single user), and for all measurable sets S of outputs, the following inequality holds:

$$\Pr[\mathcal{A}(D_1) \in S] \leq e^\epsilon \Pr[\mathcal{A}(D_2) \in S] + \delta$$

Here, parameter ϵ , known as the privacy budget, the user-defined upper bound on the privacy loss which effects the level of noise added to the algorithm, and δ is a small constant that accounts for a very rare event where the privacy guarantee might not hold¹⁴. ϵ -differential privacy (ϵ -DP) is a special case of (ϵ, δ) -DP where δ equals zero. Smaller values of ϵ correspond to stronger privacy guarantees. Selecting an appropriate value for epsilon is context-dependent and remains an open question. However, for instance, $\epsilon \leq 1$ has been suggested

¹⁴ Dwork C, Roth A. The algorithmic foundations of differential privacy. Foundations and Trends® in Theoretical Computer Science. 2014 Aug 10;9(3–4):211-407.

to offer a robust privacy guarantee¹⁵, whereas values between 1 and 10 (i.e., $1 < \epsilon \leq 10$) are still considered to provide reasonable privacy protection¹⁶, depending on the specific task and the nature of the data.

DP is heavily applied in practice, e.g., the US Census Bureau adopted differential privacy for the latest 2020 Census with a privacy budget ϵ of 19.61¹⁷.

Table 3: The dataset is managed by a trusted entity that provides statistical reports based on the data. If an adversary is aware that a specific individual has been removed from the dataset, it could deduce sensitive attributes (e.g., salary range) from the unaltered data. However, by applying differential privacy, the addition of noise to the outputs makes them indistinguishable, thereby protecting the privacy of the participants.

Company Records (May 2023)			Company Records (June 2023)		
Employee	Position	Salary	Employee	Position	Salary
Matti	CEO	3000\$	Matti	CEO	3000\$
Lisa	IT	2500\$	Lisa	IT	2500\$
Ronnie	Cleaning	500\$	Ronnie	Cleaning	500\$
Carla	HR	700\$			

Query: How many people earn <1000\$

Result: 2

Result DP: $2 \pm \text{noise} \approx 1.5$

Attacker knows Carla was fired

Query: How many people earn <1000\$

Result: 1

Result DP: $1 \pm \text{noise} \approx 1.5$

2.6 Differentially Private Synthetic Data

Differentially private synthetic data generators combine the principles of synthetic data and differential privacy to produce datasets that preserve the statistical properties and patterns of the original data while ensuring the privacy of individual data points. The generation process begins with accessing the original dataset containing sensitive information. A probabilistic model (e.g., Bayesian networks, GANs) is then constructed to capture the distributions and relationships between the variables in the original data. Calibrated noise is introduced to the model parameters or the data generation process according to differential privacy mechanisms (e.g., Laplace mechanism, Gaussian mechanism). Finally, this noisy model is used to generate a new synthetic dataset that retains the statistical properties of the original data while ensuring that individual records cannot be reverse-engineered.

Although differentially private synthetic data offers significant benefits—such as reducing the risk of privacy breaches, making the synthetic data useful for analysis and machine learning tasks, and aiding compliance with data regulations (e.g., GDPR) by mitigating the risks associated with handling sensitive data—it also presents several challenges. Finding the right balance between privacy (controlled by epsilon) and data utility is crucial,

¹⁵ Arnold C, Neunhoffer M. Really Useful Synthetic Data--A Framework to Evaluate the Quality of Differentially Private Synthetic Data. arXiv preprint arXiv:2004.07740. 2020 Apr 16.

¹⁶ Abadi M, Chu A, Goodfellow I, McMahan HB, Mironov I, Talwar K, Zhang L. Deep learning with differential privacy. In Proceedings of the 2016 ACM SIGSAC conference on computer and communications security 2016 Oct 24 (pp. 308-318).

¹⁷ U.S. Census Bureau. (2021, June 1). 2020 Census key parameters finalized. U.S. Department of Commerce. <https://www.census.gov/newsroom/press-releases/2021/2020-census-key-parameters.html>

as too much noise can degrade the usefulness of the data¹⁸. Implementing differential privacy algorithms and generating synthetic data can be complex and computationally intensive. Additionally, evaluating the quality and privacy guarantees of the synthetic data requires rigorous testing and validation.

¹⁸ Montoya Perez I, Movahedi P, Nieminen V, Airola A, Pahikkala T. Does Differentially Private Synthetic Data Lead to Synthetic Discoveries?. arXiv e-prints. 2024 Mar:arXiv-2403.

3. Synthetic data generation methods

Synthetic data generation involves creating artificial data that mimics real-world data. By generating synthetic data, researchers and developers can overcome challenges related to privacy, data scarcity, and imbalanced datasets. Synthetic data can be produced through various methods, including statistical models, simulations or generative adversarial networks, among others. Synthetic datasets are valuable for training machine learning models, testing software, and conducting research without compromising sensitive information. Moreover, synthetic data can be tailored to specific scenarios, providing a versatile and ethical solution for data-driven applications.

The development and use of synthetic data generators in healthcare represent an increasingly significant area of research, especially as it addresses critical challenges related to data availability and privacy¹⁹. The process of generating synthetic data generally involves four key stages:

- i. **Acquisition:** Obtaining and managing the real data that will be used as the basis for the synthetic data.
- ii. **Preparation:** Cleaning, preprocessing, and transforming the real data to ensure it is ready for modeling.
- iii. **Modeling:** Developing and training a model to create synthetic data that replicates the properties and distribution of the real data.
- iv. **Evaluation:** Assessing the synthetic data for its fidelity, privacy, and utility in comparison to the real data.

Next, we will present the methods of interest for WP3, categorized by the types of data to which they are applied.

3.1 Methods for tabular data

Current methods for generating synthetic tabular data fall into three main categories: statistical and probabilistic methods, machine learning techniques, and deep learning models. Statistical and probabilistic methods are often valued for their simplicity and interpretability, making them useful for generating data that approximates the original dataset's statistical properties. However, they may struggle with capturing complex dependencies in high-dimensional data. Machine learning techniques provide a more flexible approach, capable of learning and replicating intricate patterns within the data, but they may require extensive training and can be less transparent in their operations. Deep learning models offer advanced capabilities for modeling complex relationships and generating highly realistic synthetic data, but they often come with increased computational demands and can be prone to overfitting if not carefully managed. Each of these methods contributes to the growing interest in synthetic data as a tool for advancing healthcare research, as they help mitigate some of the risks associated with using real-world data. However, they also come with trade-offs, such as the potential for reduced interpretability, higher computational costs, privacy risk and the need for careful tuning to ensure data quality and utility.

A variety of tools, both commercial and open-source, are available for synthetic tabular data generation (see ²⁰ for a comprehensive list). For example, the Synthetic Data Vault (SDV) is a Python library designed for creating synthetic tabular data. SDV provides a range of models, from classical statistical methods like GaussianCopula to advanced deep learning techniques such as CTGAN, enabling the analysis and replication

¹⁹ Pezoulas VC, Zaridis DI, Mylona E, Androutsos C, Apostolidis K, Tachos NS, Fotiadis DI. Synthetic data generation methods in healthcare: A review on open-source tools and methods. *Computational and Structural Biotechnology Journal*. 2024 Jul 9.

²⁰ <https://github.com/staice/awesome-synthetic-data/tree/master>

of patterns from real datasets. It supports synthetic data generation for single tables, multiple interconnected tables, and sequential data such as timeseries. In addition, it incorporates pre-processing techniques for data de-identification and offering methods for evaluating data quality.

While synthetic data generation can produce datasets that mimic the statistical properties of real data, it does not inherently provide mathematical guarantees for privacy preservation. Synthetic data aims to replicate the patterns and structure of original data without disclosing sensitive information; however, the level of privacy protection it offers is not formally quantified. The risk remains that synthetic data could inadvertently reveal sensitive information if the generation process does not sufficiently obscure individual records or patterns. On the other hand, synthetic data generated with DP provides a mathematically rigorous framework that ensures a defined level of privacy protection, significantly reducing the risk of re-identification or leakage of sensitive details. Over the last decade, several methods for generating DP-synthetic tabular data have been proposed^{21,22,23}.

Studies comparing these DP synthesizers have shown that simple marginal-based methods often outperform more complex methods in terms of preserving privacy while maintaining, to some extent, the utility of the original data^{24,25}. However, marginal methods have the limitation of being mostly suitable for low-dimensional datasets, as preserving all combinations of marginals can be computationally infeasible as the dimensionality grows. Therefore, to address the challenge of high-dimensional datasets, some of these methods have been adapted to automatically select a subset of marginals from the original data to be preserved in the generated DP-synthetic data.

A state-of-the-art example of a marginal-based method for generating DP synthetic data is AIM (Adaptive and Iterative Mechanism²⁶). AIM is a workload-adaptive algorithm that operates within a framework where an initial set of queries is chosen, these queries are then measured privately, and synthetic data is subsequently generated from the noisy measurements. AIM incorporates a range of innovative techniques to iteratively select the most valuable measurements, considering both their relevance to the workload and their effectiveness in approximating the original data. The DP synthetic tabular data generated by AIM has demonstrated superior performance in downstream classification tasks^{27,28}, often outperforming other DP synthesizers. Implementations of AIM are available through the SmartNoise Synthesizers from OpenDP²⁹, with algorithms and further details provided in²⁸.

In the context of the prostate cancer use case (T6.2. UC2), we present an example of synthetic data quality using a real-world public dataset. This dataset, used by (Klingebiel et al., 2022)³⁰, includes data from 785 patients in Germany suspected of having prostate cancer, collected between 2014 and 2019. For this example,

²¹ McKenna R, Miklau G, Sheldon D. Winning the NIST Contest: A scalable and general approach to differentially private synthetic data. arXiv preprint arXiv:2108.04978. 2021 Aug 11.

²² Hardt M, Ligett K, McSherry F. A simple and practical algorithm for differentially private data release. *Advances in neural information processing systems*. 2012;25.

²³ Ping H, Stoyanovich J, Howe B. Datasynthesizer: Privacy-preserving synthetic datasets. In *Proceedings of the 29th International Conference on Scientific and Statistical Database Management 2017 Jun 27* (pp. 1-5).

²⁴ Bowen CM, Snoke J. Comparative study of differentially private synthetic data algorithms from the NIST PSCR differential privacy synthetic data challenge. arXiv preprint arXiv:1911.12704. 2019 Nov 28.

²⁵ Tao Y, McKenna R, Hay M, Machanavajjhala A, Miklau G. Benchmarking differentially private synthetic data generation algorithms. arXiv preprint arXiv:2112.09238. 2021 Dec 16.

²⁶ McKenna R, Mullins B, Sheldon D, Miklau G. Aim: An adaptive and iterative mechanism for differentially private synthetic data. arXiv preprint arXiv:2201.12677. 2022 Jan 29.

²⁷ Movahedi P, Nieminen V, Perez IM, Daafane H, Sukhwai D, Pahikkala T, Airola A. Benchmarking evaluation protocols for classifiers trained on differentially private synthetic data. *IEEE Access*. 2024 Aug 21.

²⁸ Pereira M, Kshirsagar M, Mukherjee S, Dodhia R, Lavista Ferres J, de Sousa R. Assessment of differentially private synthetic data for utility and fairness in end-to-end machine learning pipelines for tabular data. *Plos one*. 2024 Feb 5;19(2):e0297271.

²⁹ <https://opendp.org/>

³⁰ Klingebiel M, Arsov C, Ullrich T et al. Data on the detection of clinically significant prostate cancer by magnetic resonance imaging (MRI)-guided targeted and systematic biopsy. *Data in Brief*. 2022 Dec 1;45:108683.

we selected several key variables related to prostate cancer: age, prostate-specific antigen (PSA) level, prostate volume, PIRADS (Prostate Imaging Reporting and Data System) score, and GGG (Gleason Grade Group).

The synthetic data was generated using the SDV GaussianCopula synthesizer, while DP synthetic data was produced with the AIM generator at two levels of privacy (i.e., $\epsilon = 1$ and $\epsilon = 10$). The goal of this example is to demonstrate how well these two methods preserve the statistical structure of the original variables, both individually and in pairs. For the univariate analysis, we compare the original distribution of the variables with the distribution generated by the synthetic data. For continuous variables, see Figure 3, and for discrete variables, see Figure 4.

From Figures 3 and 4, we observe that the AIM DP synthetic dataset with $\epsilon = 10$ produces univariate distributions closest to those of the real dataset. However, additional analyses, such as multivariate correlation analysis, are necessary to provide a more comprehensive assessment of the synthetic dataset's quality. Therefore, in Figure 5, we present the correlation for paired variables to demonstrate how well these methods preserve the correlations observed in the real dataset. We find that the AIM DP synthetic dataset with $\epsilon = 10$ preserves the bivariate correlations effectively. It is important to note that the correlations in the AIM DP synthetic dataset with $\epsilon = 1$ are lower, as stronger privacy guarantees typically reduce data utility. Additionally, while the correlations in the SDV GaussianCopula synthetic dataset are also relatively close to those of the real dataset, this method does not offer mathematical guarantees of privacy.

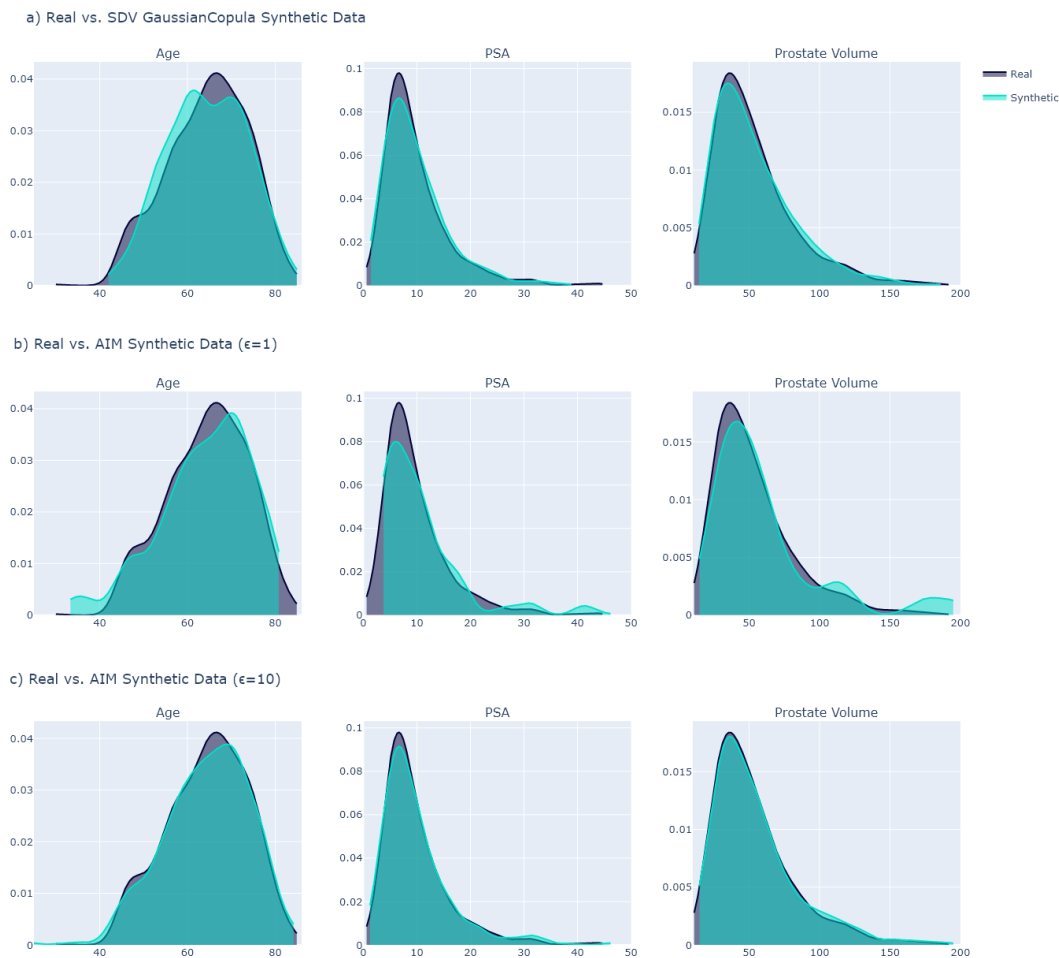


Figure 3: Distribution of the continuous variables in the real dataset compared to their distribution in the synthetic dataset generated by: a) SDV GaussianCopula, b) AIM ($\epsilon = 1$), c) AIM ($\epsilon = 10$).

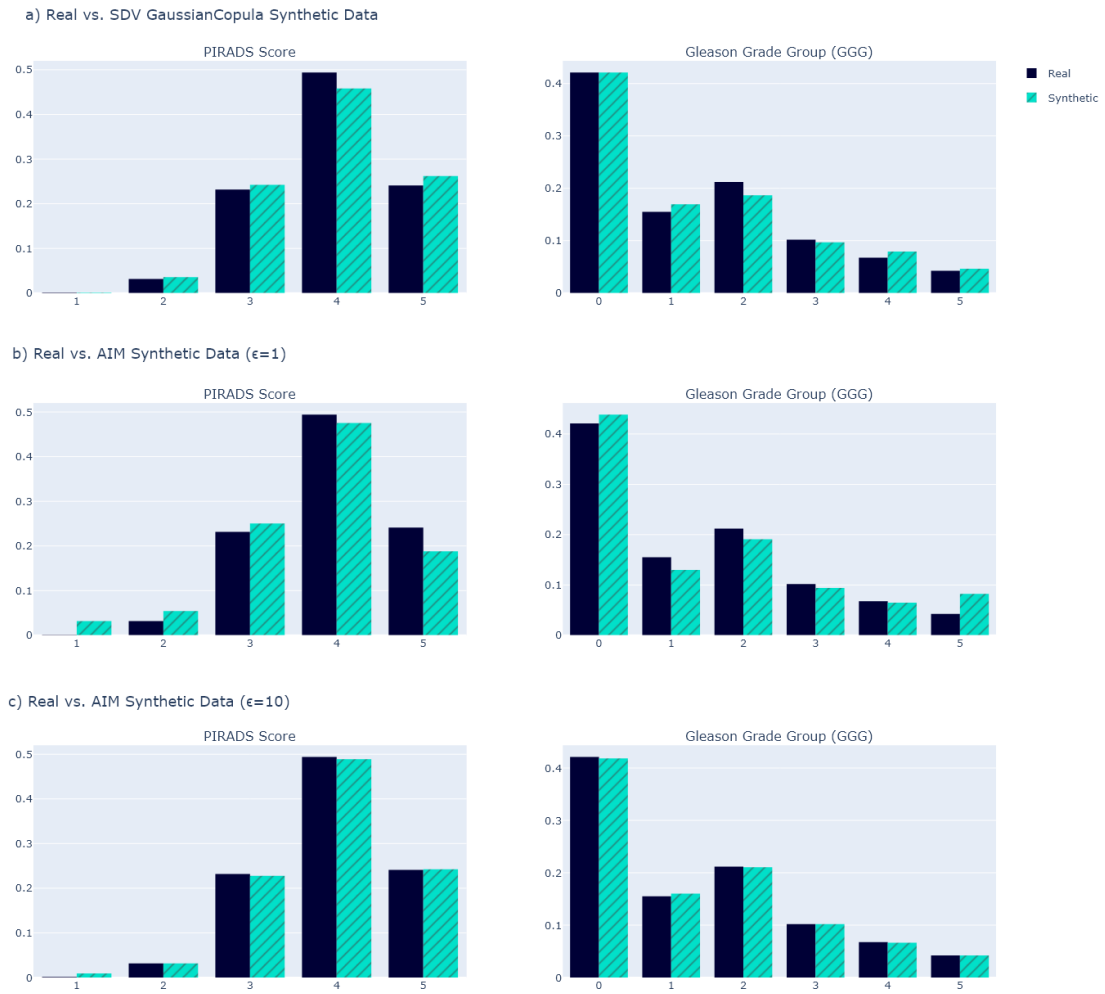


Figure 4: Distribution of the discrete variables in the real dataset compared to their distribution in the synthetic dataset generated by: a) SDV GaussianCopula, b) AIM ($\epsilon = 1$), c) AIM ($\epsilon = 10$).

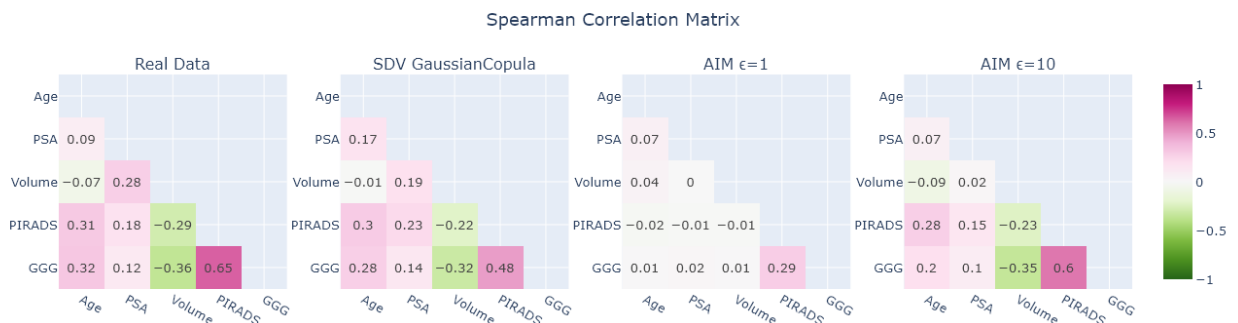


Figure 5: Spearman correlation coefficients for pairs of continuous and discrete variables in the real dataset, SDV GaussianCopula synthetic dataset, AIM synthetic dataset with a privacy level of $\epsilon = 1$, and AIM synthetic dataset with a privacy level of $\epsilon = 10$.

3.2 Methods for longitudinal data

The increasing adoption of Electronic Health Record systems has transformed healthcare by providing a comprehensive and continuous digital record of a patient's medical history. EHRs are inherently longitudinal, capturing a patient's health information across multiple visits over time. This includes a wide range of data such as diagnoses, procedures, medications, and lab results. Longitudinal nature of EHRs, allows healthcare providers to track and analyze complex patterns and relationships in patient data supporting various applications, including clinical predictive modeling, health monitoring and treatment recommendations. However, the complexity and sensitivity of EHR data pose significant challenges in terms of privacy, security, and legal constraints, making data sharing difficult. To address these issues, researchers have explored the use of synthetic patient data, generated by advanced models, as a viable alternative to real EHR data. Synthetic datasets can mimic the structure and longitudinal nature of real patient records without compromising patient privacy, thereby providing a safer avenue for sharing and utilizing health data in AI and ML research. One of the most recent and promising synthetic longitudinal data generators is HALO (Hierarchical Autoregressive Language Model)³¹. HALO is a sophisticated generative model designed to synthesize realistic, high-dimensional longitudinal EHRs and can capture the complex relationships and patterns, e.g., correlations, missing data and varying visit sequences, within such data. HALO generates synthetic EHR datasets that closely reflect the statistical properties of real EHRs, making it a powerful tool for enhancing machine learning models and supporting a wide range of healthcare applications.

3.3 Methods for medical imaging data

Deep learning generative models have revolutionized the landscape of image-based medical applications, providing novel methodologies for data synthesis, augmentation, and interpretation³². These models are capable of learning complex data distributions, which allows them to generate high-quality, diverse, and realistic images closely mimicking real data. This capability is crucial in medical fields where data scarcity and privacy concerns limit the availability of large datasets³³.

Furthermore, de-facing technologies facilitate the creation of synthetic datasets that mimic real patient data without compromising privacy. These synthetic datasets, generated through techniques such as GANs or Diffusion Models (DMs), can be used for downstream tasks to train machine learning models, enhancing medical research, and developing new treatments, all while ensuring compliance with stringent data protection laws such as GDPR.

Generative models allow the possibility of generating synthetic data in different spatial domains but also temporal. Thus, GANs have been also utilized for brain tumor segmentation, converting healthy images to diseased ones, or translating between CT and MRI images, often outperforming other AI-based techniques.³⁴ But also, other studies have shown their capability for disease evolution prediction in brain MRI.³⁵ Even such a remarkable performance, further work is required to integrate generative models into clinical practice.

³¹ Theodorou B, Xiao C, Sun J. Synthesize high-dimensional longitudinal electronic health records via hierarchical autoregressive language model. *Nature communications*. 2023 Aug 31;14(1):5305.

³² Croitoru FA, Hondru V, Ionescu RT, Shah M. Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2023 Mar 27;45(9):10850-69.

³³ Kazerouni A, Aghdam EK, Heidari M, Azad R, Fayyaz M, Hacihaliloglu I, Merhof D. Diffusion models in medical imaging: A comprehensive survey. *Medical Image Analysis*. 2023 Aug 1;88:102846.

³⁴ Ali H, Biswas MR, Mohsen F, Shah U, Alamgir A, Mousa O, Shah Z. The role of generative adversarial networks in brain MRI: a scoping review. *Insights Imaging*. 2022;13(1):98.

³⁵ Rachmadi MF, Valdés-Hernández MDC, Makin S, Wardlaw J, Komura T. Automatic spatial estimation of white matter hyperintensities evolution in brain MRI using disease evolution predictor deep neural networks. *Med Image Anal*. 2020;63:101712.

3.3.1 Synthetization architectures: GANs, Diffusion Models and Transformers

3.3.1.1 GAN-based medical image synthesis

Generative Adversarial Networks were introduced in Goodfellow et al.³⁶ and quickly became the state-of-the-art in generative models until the breakthrough of diffusion models. A GAN architecture is characterized by two fundamental components: a generator that creates synthetic data and a discriminator that distinguishes real from fake data. The network is trained through a minimax game where the generator aims to fool the discriminator, while the discriminator aims to correctly identify real and synthetic data. This adversarial training allows to learn accurate data distributions implicitly, synthetizing realistic samples. However, it is sometimes unstable producing what is called model collapse in which the variety of outputs is limited. Variants such as DCGAN, WGAN, and StyleGAN have improved stability and quality, but GANs are still considered one of the most difficult categories of generative networks to train.

GANs have been applied to different medical image problems. Their primary use in the medical field is data augmentation for downstream models, such as classification, detection or segmentation.³⁷ GAN synthesis can also be conditioned as in Mirza et al.³⁸ to enable other tasks such as signal to image reconstruction, image super-resolution, de-noising or registration. GANs have been successfully applied to a wide range of medical imaging techniques³⁹.

3.3.1.2 Diffusion Models superiority in medical imaging

In recent years, DMs have demonstrated significant superiority over GANs in synthetic data generation. The diffusion fundamentals were introduced in Sohl-Dickstein et al.⁴⁰ and they were extended for high-quality image synthesis in Ho et al.⁴¹. After that, DMs became globally known in 2022 after the release of DALL-E 2, a high-resolution text-to-image generation.⁴²

DMs consist of two fundamental processes: a forward diffusion process, which gradually adds noise to an source sample, e.g. an image; and a reverse denoising process, which gradually cleans samples and ultimately generates high-fidelity new samples from pure noise samples.

Compared to GANs, DMs ensure higher stability during training and can capture a wider diversity of image features, leading to more accurate and representative synthetic data. In medical imaging, where fine-grained precision and wide diversity are crucial, DMs have shown improved performance in generating high-fidelity and diverse images, better preserving anatomical structures and avoiding artifacts. For example, Müller-Franzes et al.⁴³ showed superiority of DMs over different GAN variants using funduscopy images, radiographs

³⁶ Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y. Generative adversarial nets. *Advances in neural information processing systems*. 2014;27.

³⁷ Jeong JJ, Tariq A, Adejumo T, Trivedi H, Gichoya JW, Banerjee I. Systematic review of generative adversarial networks (GANs) for medical image classification and segmentation. *J Digit Imaging*. 2022;35(2):137-52.

³⁸ Mirza M, Osindero S. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*. 2014.

³⁹ Kazemina S, Baur C, Kuijper A, van Ginneken B, Navab N, Albarqouni S, et al. GANs for medical image analysis. *Artif Intell Med*. 2020;109:101938.

⁴⁰ Sohl-Dickstein J, Weiss E, Maheswaranathan N, Ganguli S. Deep unsupervised learning using nonequilibrium thermodynamics. In: *International Conference on Machine Learning*. PMLR; 2015. p. 2256-65.

⁴¹ Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models. *Adv Neural Inf Process Syst*. 2020;33:6840-51.

⁴² Ramesh A, Dhariwal P, Nichol A, Chu C, Chen M. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*. 2022;1(2):3.

⁴³ Müller-Franzes G, Niehues JM, Khader F, Arasteh ST, Haarburger C, Kuhl C, et al. A multimodal comparison of latent denoising diffusion probabilistic models and generative adversarial networks for medical image synthesis. *Sci Rep*. 2023;13(1):12098.

and histopathology images. Other comparative studies with similar conclusions for brain MRI images have been performed in Pinaya et al.⁴⁴ or Wolleb et al.⁴⁵, including diversity.⁴⁶

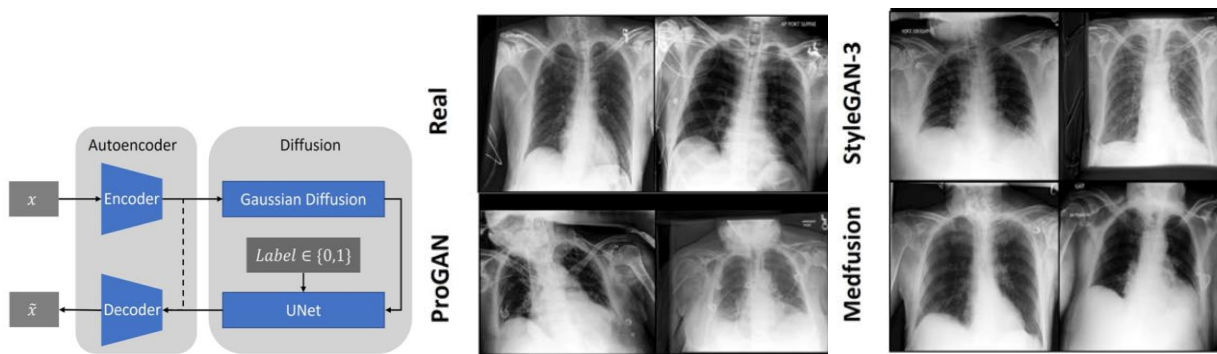


Figure 6: Basic pipeline diagram of diffusion models (left) and qualitative comparison of various synthesization methods. Source: (Müller-Franzes et al. 2023).

The main limitation of DMs is the inference time. The iterative process of generating new samples from the noise space to the image is basically time-consuming as shown in Xiao et al.⁴⁷. However, in medical applications, the synthesis quality is often more important than the sampling speed. A few works also point out other DMs drawbacks, such as memorization issues when working with very small datasets of less than 500 images⁴⁸ or the fact that direct evaluation metrics do not always impact proportionally medical downstream tasks.⁴⁹

Among DMs, latent diffusion is today the most popular approach.⁵⁰ In latent diffusion pipelines, the diffusion process is not performed in pixel domain, but rather in a latent space of reduced dimensionality and much more efficient. The conversion between the image and latent spaces is done by an auto encoder-decoder. Latent space conversion allows to reduce the number of diffusion steps until the noise state is reached. This is a great advantage since it allows to generate high resolution images more efficiently.

3.3.1.3 Transformers and DMs

Transformers were originally introduced in Vaswani et al.⁵¹ as a type of neural network architecture for natural language processing. The transformer architecture relies on self-attention mechanisms to weigh the significance of different parts of the input data, allowing to capture long-range dependencies and contextual

⁴⁴ Pinaya WH, Tudosiu PD, Dafflon J, Da Costa PF, Fernandez V, Nachev P, et al. Brain imaging generation with latent diffusion models. In: MICCAI Workshop on Deep Generative Models. Cham: Springer Nature Switzerland; 2022. p. 117-26.

⁴⁵ Wolleb J, Bieder F, Sandkühler R, Cattin, PC. Diffusion models for medical anomaly detection. In International Conference on Medical image computing and computer-assisted intervention. 2022:35-45.

⁴⁶ Khader F, Müller-Franzes G, Arasteh ST, Han T, Haarbuerger C, Schulze-Hagen M, et al. Medical diffusion: Denoising diffusion probabilistic models for 3D medical image generation. arXiv preprint arXiv:2211.03364. 2022.

⁴⁷ Xiao Z, Kreis K, Vahdat A. Tackling the generative learning trilemma with denoising diffusion GANs. arXiv preprint arXiv:2112.07804. 2021.

⁴⁸ Akbar MU, Wang W, Eklund A. Beware of diffusion models for synthesizing medical images—A comparison with GANs in terms of memorizing brain tumor images. arXiv preprint arXiv:2305.07644. 2023.

⁴⁹ Akbar MU, Larsson M, Blystad I, Eklund A. Brain tumor segmentation using synthetic MR images—A comparison of GANs and diffusion models. Sci Data. 2024;11(1):259.

⁵⁰ Rombach R, Blattmann A, Lorenz D, Esser P, Ommer B. High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022. p. 10684-95.

⁵¹ Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. 2017. p. 6000-10.

relationships with great efficiency. Transformers are very popular for autoregression, a type of algorithms aimed to predict a set of features (or tokens) based on a preceding set.

It is important to note that the transformers architecture is usable in different machine learning problems, and not a category of generative methods *per-se*, such as GANs or DMs. As such, transformers have influenced generative models in recent years as a component that can lead to performance improvements.

The most popular transformer variant for vision problems is the Vision Transformer, that divides images into flattened image patches treated as tokens.⁵² The first DALL-E model employed a Variational Autoencoders (VAE) and a transformer to regress coherent images from text captions.⁵³ Other approached adapted transformers for high-resolution image synthesis by introducing a quantized VQ-VAE encoder-decoder and a sliding-window logic.⁵⁴ After the breakthrough of DMs, transformers have also been plugged into diffusion frameworks. For instance, DiT⁵⁵ and U-ViT⁵⁶ replaced the commonly-used U-Net with a transformer operating on latent patches. As a result, transformed-based DMs have also shown promising results in medical imaging applications.^{57 58 59}

3.3.2 Conditional Image Synthesis in DMs

In DMs, conditional and unconditional synthesis are two distinct approaches to generating data. Unconditional synthesis involves generating data without any specific guidance, and thus observing the training data as a whole. Conversely, conditional synthesis incorporates additional information within the generative process, guiding the model to produce outputs under particular conditions. This is crucial in tasks where specific attributes or a context must be fulfilled, such as generating images based on textual descriptions or modifying existing images to include certain features. Both approaches leverage the iterative refinement process of DMs, but conditional synthesis offers fine-grained control and higher specificity. It is for this reason that conditional models are more often present in medical applications. In fact, medical images are always paired with descriptive metadata, and conditional synthesis is the approach that comes closest to this reality.

The following subsections review some of the most common conditional frameworks in medical image synthesis. Note that employing a specific conditioning methodology does not preclude the use of other conditional data, e.g., localization masks and semantic metadata can be used simultaneously.

3.3.2.1 Mask-based conditional synthesis

Mask-conditioned synthesis is one of the most common and intuitive approach in medical imaging. Masks can be binary —to indicate the location of organs or lesions— or semantic —to indicate anatomical contours.

Mask-conditioned synthesis is usually addressed in DMs by concatenating the mask(s) throughout all the process with the noisy training samples. At each diffusion step, noise is only added to masked area of the

⁵² Dosovitskiy A. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929. 2020.

⁵³ Ramesh A, Pavlov M, Goh G, Gray S, Voss C, Radford A, et al. Zero-shot text-to-image generation. In: International Conference on Machine Learning. PMLR; 2021. p. 8821-31.

⁵⁴ Esser P, Rombach R, Ommer B. Taming transformers for high-resolution image synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021.

⁵⁵ Peebles W, Xie S. Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023.

⁵⁶ Bao F, Nie S, Xue K, Cao Y, Li C, Su H, et al. All are worth words: A ViT backbone for diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023. p. 22669-79.

⁵⁷ Wu J, Ji W, Fu H, Xu M, Jin Y, Xu Y. Medsegdiff-v2: Diffusion-based medical image segmentation with transformer. In: Proceedings of the AAAI Conference on Artificial Intelligence. 2024;38(6):6030-8.

⁵⁸ Chen X, Liu Y, Yang B, Zhu J, Yuan S, Xie X, et al. A more effective CT synthesizer using transformers for cone-beam CT-guided adaptive radiotherapy. *Front Oncol.* 2022;12:988800.

⁵⁹ Zhu J, Zhu H, Jia Z, Ma P. DiffSwinTr: A diffusion model using 3D Swin Transformer for brain tumor segmentation. *Int J Imaging Syst Technol.* 2024;34(3):e23080.

image. Therefore, the reverse process can be initialized from noise and a mask, and the model produces a synthetic output that is consistent with the mask.

In this way, Macháček et al.⁶⁰ use binary mask-conditioned latent diffusion for generating 2D gastrointestinal polyp images with realistic polyp location and sizes. Yu et al.⁶¹ use lymph node masks and anatomical structure masks to control global structure and local node details in 2D synthesis of abdominal CT exhibiting lymph nodes. Similarly, Dorjsembe et al.⁶² use mask-conditioned diffusion models for 3D brain MRI synthesis. The key difference in mask-conditioned 3D synthesis compared to the 2D counterpart lies in the concatenation operation: rather than directly concatenating the entire mask volume behind the image or noisy volume, each spatial layer is concatenated with its corresponding mask layer, alternating the channels.

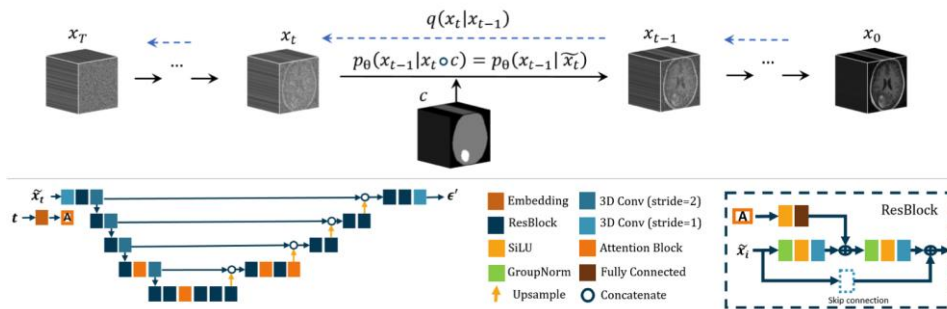


Figure 7: Illustration of a diffusion model process conditioned on 3D segmentation masks for brain MRI synthesis. Source: (Dorjsembe et al. 2024).

Another important aspect of using conditional masks is the origin of the masks at the time of synthetic data generation. There are various possibilities. One option is to reuse the training masks, either directly or by transforming or manipulating them to account for realistic variations.^{63 64} However, reusing the training masks does not ensure complete coverage of the entire mask space. Another approach is to use genuinely new synthetic masks. This can be generated heuristically, adhering to a series of priors (e.g., minimum and maximum organ sizes, more likely locations, etc.), or by training a secondary diffusion model on real masks.⁶⁵

3.3.2.2 Context-based conditional synthesis

In certain applications, it may be desirable to preserve part of an existing image (the context) while modifying only certain parts. This approach can be beneficial for partial data augmentation, where instead of synthesizing entirely new images, specific features are removed or added to the real data—for instance, anonymization purposes. It is common that the modifications target one or more subregions (inpainting). However, other

⁶⁰ Macháček R, Mozaffari L, Sepasdar Z, Parasa S, Halvorsen P, Riegler MA, Thambawita V. Mask-conditioned latent diffusion for generating gastrointestinal polyp images. In: Proceedings of the 4th ACM Workshop on Intelligent Cross-Data Analysis and Retrieval. 2023. p. 1-9.

⁶¹ Yu Y, Chen H, Zhang Z, Xiao Q, Lei W, Dai L, et al. CT synthesis with conditional diffusion models for abdominal lymph node segmentation. arXiv preprint arXiv:2403.17770. 2024

⁶² Dorjsembe Z, Pao HK, Odonchimed S, Xiao F. Conditional diffusion models for semantic 3D brain MRI synthesis. IEEE J Biomed Health Inform. 2024.

⁶³ Dorjsembe Z, Pao HK, Odonchimed S, Xiao F. Conditional diffusion models for semantic 3D brain MRI synthesis. IEEE J Biomed Health Inform. 2024.

⁶⁴ Yu Y, Chen H, Zhang Z, Xiao Q, Lei W, Dai L, et al. CT synthesis with conditional diffusion models for abdominal lymph node segmentation. arXiv preprint arXiv:2403.17770. 2024.

⁶⁵ Macháček R, Mozaffari L, Sepasdar Z, Parasa S, Halvorsen P, Riegler MA, Thambawita V. Mask-conditioned latent diffusion for generating gastrointestinal polyp images. In: Proceedings of the 4th ACM Workshop on Intelligent Cross-Data Analysis and Retrieval. 2023. p. 1-9.

forms of context-conditional synthesis are possible as well. As an example of the latter, in the DM SDEdit local details are collectively altered without affecting a given overall structure.⁶⁶

Diffusion models are currently the state of the art in image inpainting. At training time, DMs can perform inpainting by gradually adding noise to the image and exploiting contextual information in the form of conditional masks delimiting inpainting regions.⁶⁷ For greater efficiency, it is possible to avoid the use of masks by adding noise only to pixels in inpainting regions instead of the entire image.⁶⁸ Noise can be injected uniformly within the inpainting boundaries for sharper and stronger modifications; alternatively, the addition of noise can be weighted based on the distance from a given centroid or stroke, for smoother transitions between the inpainting areas and the context. As an example of the latter, Hansen et al.⁶⁹ use a latent DM to inpaint pathologies in lumbar spine MRI. They work with 3D data and use a spherical noise weighting scheme based on the distance with respect to some landmark points annotated in the spine.

3.3.2.3 EHR/Metadata conditional synthesis

Metadata such as age, gender, weight, descriptive attributes of organs or the patient's medical history are essential for interpreting medical images. These data can be extracted from EHRs or text labels in medical image files (e.g. DICOM, TIFF, NIFTI).

Image synthesis conditioned by text or numerical variables is one of the main topics in diffusion modeling. Pinaya et al.⁷⁰ introduced a DM for brain MRI synthesis conditioned on age, gender, and brain structure attributes. Instead of directly feeding scalar values to the model, the conditional variables are mapped to a higher-dimensional feature space using an auxiliary learnable encoder. The conditional feature vectors are concatenated with the input data and mapped to the intermediate layers of the DM via cross-attention mechanisms. Wang et al.⁷¹ use the same conditional variables and mechanism in their super-resolution DM for brain MRI synthesis. Chambon et al.⁷² also use an auxiliary pre-trained text encoder to train a text-to-image latent DM to generate chest X-Rays conditioned on short radiology reports. It is important to note that pre-trained text encoders may be a useful initialization, but medical image synthesis requires fine-tuning to adapt the weights to the medical language, which has specific peculiarities such as brief but semantically rich. Other studies Montoya-del-Angel et al.⁷³ adopt an analogous framework for synthesizing mammographic images. In their case, the text describes the type of view (CC or MLO), breast density, breast area, and the mammographic unit vendor. Also, Osorio et al.⁷⁴ used latent DM for histopathology image synthesis conditioned to a text prompt indicating health condition and morphology.

⁶⁶ Meng X, Kabashima Y. Diffusion model based posterior sampling for noisy linear inverse problems. arXiv preprint arXiv:2211.12343. 2022.

⁶⁷ Durrer A, Cattin PC, Wolleb J. Denoising diffusion models for inpainting of healthy brain tissue. arXiv preprint arXiv:2402.17307. 2024.

⁶⁸ Saharia C, Chan W, Chang H, Lee C, Ho J, Salimans T, et al. Palette: Image-to-image diffusion models. In: ACM SIGGRAPH 2022 Conference Proceedings. 2022. p. 1-10.

⁶⁹ Hansen C, Glinskis S, Raju A, Kornreich M, Park J, Pawar J, et al. Inpainting pathology in lumbar spine MRI with latent diffusion. arXiv preprint arXiv:2406.02477. 2024.

⁷⁰ Pinaya WH, Tudosi PD, Dafflon J, Da Costa PF, Fernandez V, Nachev P, et al. Brain imaging generation with latent diffusion models. In: MICCAI Workshop on Deep Generative Models. Cham: Springer Nature Switzerland; 2022. p. 117-26.

⁷¹ Wang J, Levman J, Pinaya WHL, Tudosi PD, Cardoso MJ, Marinescu R. Inversesr: 3D brain MRI super-resolution using a latent diffusion model. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Cham: Springer Nature Switzerland; 2023.

⁷² Chambon P, Bluethgen C, Delbrouck JB, Van der Sluijs R, Polacin M, Chaves JMZ, et al. Roentgen: vision-language foundation model for chest x-ray generation. arXiv preprint arXiv:2211.12737. 2022.

⁷³ Montoya-del-Angel R, Sam-Millan K, Vilanova JC, Martí R. MAM-E: Mammographic synthetic image generation with diffusion models. *Sensors*. 2024;24(7):2076.

⁷⁴ Osorio P, Jimenez-Perez G, Montalt-Tordera J, Hooge J, Duran-Ballester G, Singh S, et al. Latent diffusion models with image-derived annotations for enhanced AI-assisted cancer diagnosis in histopathology. arXiv preprint arXiv:2312.09792. 2023.

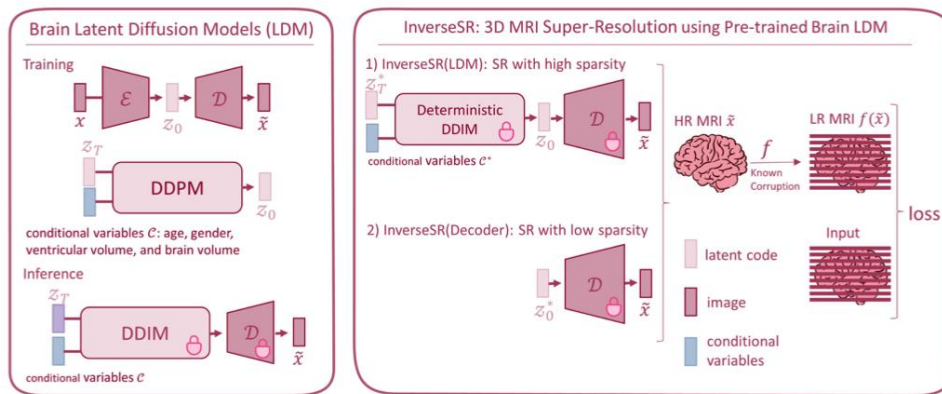


Figure 8: Illustration of a super resolution diffusion workflow conditioned on the metadata variables age, gender, ventricular volume, and brain volume. Source: (Wang et al. 2023).

3.3.2.4 Modality-based translation synthesis

The main medical image modalities include X-ray, which uses ionizing radiation for skeletal and chest imaging; MRI (Magnetic Resonance Imaging), which uses magnetic fields and radio waves for detailed soft tissue visualization; and CT (Computed Tomography), combining X-rays and computer processing to produce cross-sectional images for both bone and soft tissue analysis. Ultrasound uses high-frequency sound waves to capture soft tissues and organs. Nuclear medicine techniques, like PET and SPECT, involve radioactive tracers to assess physiological functions. Medical images of different modalities can provide complementary and valuable information. For example, a chest X-ray can be followed by a CT or PET-CT scan for better assessment.

However, acquiring multiple medical image modalities of a patient can be challenging due to factors such as cost, limited scan time, and safety considerations. Modality-based translation synthesis offers a solution by generating missing or complementary imaging modalities, thus enhancing diagnostic capabilities without requiring additional scans. These methods can be categorized into single or multi-modality translation models. Multi-modality translation models are particularly advantageous as they eliminate the need for training separate models for each pair of modalities. However, the task is highly ambitious and requires extensive data from multiple potentially non-aligned datasets.

Single modality-translation diffusion models:

Pan et al.⁷⁵ propose a MRI-to-CT transformer-based DM to synthesize CT scans matching the anatomy of a conditional MRI. Lyu et al.⁷⁶ address the same problem using a UNet DM. Gong et al.⁷⁷ use a DM to synthesize high-resolution PET images from low-resolution PET (less invasive for patients) and MRI priors. Ablation studies show that combining different modalities in the input improves the performance of the DM. PET synthesis using MRI priors is also addressed in Xie et al.⁷⁸. These works pay careful attention to how the acquisition volumes are co-registered and preprocessed to train the DM. Another possibility, less common, is

⁷⁵ Pan S, Abouei E, Wynne J, Wang T, Qiu RL, Li Y, et al. Synthetic CT generation from MRI using 3D diffusion model. In: Medical Imaging 2024: Image Processing. Vol. 12926. SPIE; 2024. p. 636-43.

⁷⁶ Lyu Q, Wang G. Conversion between CT and MRI images using diffusion and score-matching models. arXiv preprint arXiv:2209.12104. 2022.

⁷⁷ Gong K, Johnson K, El Fakhri G, Li Q, Pan T. PET image denoising based on denoising diffusion probabilistic model. Eur J Nucl Med Mol Imaging. 2024;51(2):358-68.

⁷⁸ Xie T, Cao C, Cui ZX, Guo Y, Wu C, Wang X, et al. Synthesizing PET images from high-field and ultra-high-field MR images using joint diffusion attention model. Med Phys. 2023.

to synthesize a new modality that merges the characteristics of different inputs. Zhao et al.⁷⁹ explore this problem for merging MRI-CT, MRI-PET and MRI-SPECT image pairs into a single image. The fusion problem is split into a DM to leverage image priors and a maximum likelihood sub-problem to preserve cross-modality information.

Multiple modality-translation diffusion models:

Kim et al.⁸⁰ employ a U-Net DM architecture but incorporate a MS-SPADE block of multiple switchable spatially adaptive normalization layers for style transfer. The MS-SPADE block allows the same DM to tackle different translation tasks of one source modality to various targets, removing the need for training a separate DM for each combination of modalities. The method is tested on different MRI modalities (T1, T1ce, T2, FLAIR). Zhan et al.⁸¹ recently introduced another state-of-the-art DM to unify the generation and translation between different modalities (CT, MRI, X-ray, text). In this framework, data from different modalities are encoded into a unified latent space by utilizing the common presence of text in most medical cross-modal paired data. Encoders for CT, MRI, and X-ray images are then trained to project their resulting feature vectors into the same latent space as the paired text descriptions.

3.3.3 Volume Synthesis: from 2D to 3D

Medical images can be 2D or 3D depending on the modality. CT, MRI, PET or 3D Ultrasounds are widely used 3D data modalities. 3D information aids in understanding spatial relationships and morphologies, essential for accurate diagnostic and treatment planning. Volume or 3D synthesis is therefore vital to accurately mimic these modalities and make an impact on downstream applications.

Different ways of approaching 3D data synthesis using DMs have been presented in the literature. We mainly distinguish between models that generate 3D indirectly, from 2D data, and models that generate 3D directly, from 3D inputs.

3.3.3.1 3D-from-2D approaches

Handling 3D information with neural networks is a long standing challenge due to the extremely high memory and computational cost. To avoid working directly with 3D data, some works in the literature propose to divide the volume synthesis into 2D slices processed independently by the DM. However, if the inter-dependency between slices is not considered, the output volumes will exhibit morphological incoherence. Additional constraints are then imposed to encourage spatial coherence across slices from the same volume. Chung et al.⁸² propose an alternating approach where the diffusion-based denoising step is applied slice-by-slice, and a total variation (TV) regularization loss is applied along the depth axis. Their method achieves state-of-the-art performance in 3D synthesis from sparse-view CT, limited-angle CT and compressed MRI. Zhu et al.⁸³

⁷⁹ Zhao Z, Bai H, Zhu Y, Zhang J, Xu S, Zhang Y, et al. DDFM: denoising diffusion model for multi-modality image fusion. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023. p. 8082-93.

⁸⁰ Kim J, Park H. Adaptive latent diffusion model for 3D medical image-to-image translation: multi-modal magnetic resonance imaging study. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2024. p. 7604-13.

⁸¹ Zhan C, Lin Y, Wang G, Wang H, Wu J. MedM2G: Unifying medical multi-modal generation via cross-guided diffusion with visual invariant. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024. p. 11502-12.

⁸² Chung H, Ryu D, McCann MT, Klasky ML, Ye JC. Solving 3D inverse problems using pre-trained 2D diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023. p. 22542-51.

⁸³ Zhu L, Xue Z, Jin Z, Liu X, He J, Liu Z, Yu L. Make-a-volume: leveraging latent diffusion models for cross-modality 3D brain MRI synthesis. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Cham: Springer Nature Switzerland; 2023. p. 592-601.

proposed a two-stage framework instead by first train a DM using independent 2D slices, then insert volumetric layers and fine-tune using paired 3D inputs.

3.3.3.2 3D synthesis

The alternative to 3D-from-2D methods is to design DM architectures that can directly ingest and generate 3D volumes. This can be done by replacing 2D convolutions by 3D convolutions in the DM as well as in the auto-encoder layers, as in the Medical Diffusion state-of-the-art latent diffusion pipeline by Khader et al.⁸⁴ Medical Diffusion uses a VQ-GAN encoder-decoder and a U-Net shaped DM. The authors also incorporate a slice-wise discriminator and a 3D discriminator to strengthen morphological intra and inter-slice coherence. All these architectural changes are computationally costly and require small batch sizes to process only a few volumes at each forward pass. The computational cost of 3D latent diffusion pipelines can be also reduced by increasing the compression factor between the image and the latent space, although Khader et al.⁸⁵ indicate that excessive compression can lead to loss of detail or distorted details.

Other works on 3D latent diffusion of medical images (MRI and CT) are Pinaya et al.⁸⁶, very similar to Medical Diffusion from a technical point of view, or PatchDDM, which improves memory-efficiency with a patch-based logic.⁸⁷ To this end, PatchDDM performs training on 3D coordinate-encoded patches instead of the entire volume, and only synthesizes all patches of a volume at inference time. Durrer et al.⁸⁸ also investigate 3D brain MRI synthesis and claim that pseudo-3D convolutions are more efficient and outperform actual 3D convolutions. Pseudo-3D convolutions result from 2D convolutional layers followed by 1D convolutions in the z-axis.

3.3.4 Clinical applications of image synthesis

3.3.4.1 Synthetic chest CT nodule insertion

Nodule insertion techniques in pulmonary CT scans are used primarily for the creation of realistic simulated nodules within lung images, which can have significant medical applications.

AI and machine learning models for nodule detection and classification can be trained on datasets that include both real and synthetic nodules. The use of synthetic nodules ensures that the model encounters a broad range of nodule types, sizes, and densities, which can improve its accuracy and robustness. Moreover, by inserting known nodules into CT scans, researchers can create benchmark datasets to evaluate and compare the performance of different CAD systems under controlled conditions.

Nodule insertion can be also used to create patient-specific models that simulate the growth or change of nodules over time, helping clinicians to predict disease progression and tailor treatment plans accordingly. Additionally, some nodule types or patterns are rare in clinical practice, making it difficult to study them in large numbers. Insertion techniques can simulate these rare conditions, providing more opportunities for research.

⁸⁴ Khader F, Müller-Franzes G, Tayebi Arasteh S, Han T, Haarbuerger C, Schulze-Hagen M, et al. Denoising diffusion probabilistic models for 3D medical image generation. *Sci Rep.* 2023;13(1):7303.

⁸⁵ Khader F, Müller-Franzes G, Tayebi Arasteh S, Han T, Haarbuerger C, Schulze-Hagen M, et al. Denoising diffusion probabilistic models for 3D medical image generation. *Sci Rep.* 2023;13(1):7303.

⁸⁶ Pinaya WH, Tudosiu PD, Dafflon J, Da Costa PF, Fernandez V, Nachev P, et al. Brain imaging generation with latent diffusion models. In: *MICCAI Workshop on Deep Generative Models*. Cham: Springer Nature Switzerland; 2022. p. 117-26.

⁸⁷ Bieder F, Wolleb J, Durrer A, Sandkuehler R, Cattin PC. Memory-efficient 3D denoising diffusion models for medical image processing. In: *Medical Imaging with Deep Learning*. 2023.

⁸⁸ Durrer A, Wolleb J, Bieder F, Friedrich P, Melie-Garcia L, Ocampo-Pineda M, et al. Denoising diffusion models for 3D healthy brain tissue inpainting. *arXiv preprint arXiv:2403.14499*. 2024.

Not least, simulated nodules allow radiologists to practice detecting, characterizing, and diagnosing pulmonary nodules. This is especially important in teaching hospitals or training programs where experience with real cases might be limited.

In this context, Schultheiss et al.⁸⁹ proposed a simple forward projection to merge nodules in healthy areas of chest radiographies. Beyond that, several in-context synthesis methods, i.e. methods guided by the surroundings of the target area, have been proposed. Han et al.⁹⁰ discusses a method for enhancing pulmonary nodule detection in CT scans. It introduces a 3D Multi-Conditional GAN that can generate realistic and diverse lung nodules in CT images. This approach helps in creating extensive training datasets for object detection models, improving their robustness and accuracy in detecting nodules, especially in scenarios where annotated data is limited. The GAN approach of Nishio et al.⁹¹ allows additional control over attributes such as nodule size and malignancy, enabling the creation of varied and realistic nodules. The work of Jin et al.⁹² includes innovative convolutional features to handle the complex and uncertain boundaries between tumors and surrounding healthy tissue. The work of Chung et al.⁹³ used a GAN approach with contextual attention modules to capture long-range spatial dependencies dedicated to refine coarse outputs. All of these methods showed real improvements of downstream Computer-Aided Detection methods.

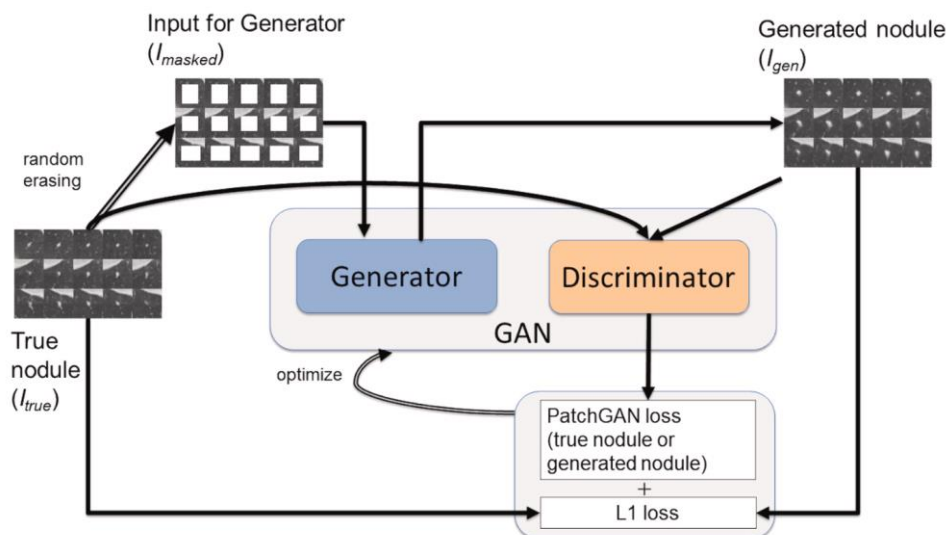


Figure 9: Scheme of a GAN model to insert synthetic nodules in context based on attributes such as size and malignancy Source: (Nishio et al. 2020).

3.3.4.2 Lung cancer evolution prediction

Tumors have their origin in the abnormal growth of cells, a process that is frequently difficult to detect at an early stage. As they progress, tumor cells invade nearby tissues irreversibly, significantly impacting the body's

⁸⁹ Schultheiss M, Schmette P, Bodden J, Aichele J, Müller-Leisse C, Gassert FG, et al. Lung nodule detection in chest X-rays using synthetic ground-truth data comparing CNN-based diagnosis to human performance. *Sci Rep.* 2021;11(1):15857.

⁹⁰ Han C, Kitamura Y, Kudo A, Ichinose A, Rundo L, Furukawa Y, et al. Synthesizing diverse lung nodules wherever massively: 3D multi-conditional GAN-based CT image augmentation for object detection. In: 2019 International Conference on 3D Vision (3DV). IEEE; 2019. p. 729-37.

⁹¹ Nishio M, Muramatsu C, Noguchi S, Nakai H, Fujimoto K, Sakamoto R, et al. Attribute-guided image generation of three-dimensional computed tomography images of lung nodules using a generative adversarial network. *Comput Biol Med.* 2020;126:104032.

⁹² Jin Q, Cui H, Sun C, Meng Z, Su R. Free-form tumor synthesis in computed tomography images via richer generative adversarial network. *Knowl Based Syst.* 2021;218:106753.

⁹³ Chung M, Kong ST, Park B, Chung Y, Jung KH, Seo JB. Utilizing synthetic nodules for improving nodule detection in chest radiographs. *J Digit Imaging.* 2022;35(4):1061-8.

resources and contributing to high cancer mortality rates. It is therefore imperative to develop longitudinal tumor prediction models in order to anticipate future developments. Based on factors such as tumor type and growth rate, doctors can implement personalized interventions, including monitoring, surgery, or drug therapy, in order to effectively manage the disease. Medical imaging offers essential insights into tumor morphology and physiological parameters, while EHRs provide contextual information about risk factors influencing disease progression.

One approach to understanding cancer progression is through the analysis of EHR data. Griffith et al.⁹⁴ analyzed cancer progression by examining an EHR at multiple time points, offering more detailed insights than overall survival metrics alone. They correlated their findings with overall survival, demonstrating the reliability, clinical relevance, and scalability of their approach on a large scale. However, focusing solely on clinical history may overlook valuable information extractable from medical imaging.

Conventional longitudinal prediction techniques reliant on image exploitation have historically relied on intricate mathematical models and differential equations to estimate tumor growth over time. These methods frequently failed to capture detailed growth information or model the entire tumor development process effectively. This difficulty led to the adoption of deep learning as the reference technique to address this problem. Zhang et al.⁹⁵ pioneered a deep learning study for pancreatic cancer, using dual-stream convolutional neural networks (CNNs) to predict future tumor segmentations and growth rates based on 2D patch images.

This approach was subsequently enhanced through the incorporation of clinical data, thereby facilitating the identification of potential risk factors that are not readily quantifiable in images. Li et al.⁹⁶ demonstrated this with a proposed multimodal method based on a 3D U-Net for predicting nodule growth. This strategy was further developed by Wang et al.⁹⁷, who introduced the Static-Dynamic Coordinated Transformer for Tumor Longitudinal Growth Prediction. This approach aims to analyze tumor growth trends over extended time sequences. Their method extracts static features from tumors at each time point, employs enhanced deformable convolutions for better spatial sampling of tumor features, and incorporates cascade self-attention operations to predict future trends accurately.

An additional innovative approach, proposed by Rachmadi et al.⁹⁸, integrates clinical and image data using a GAN-based model, designated as the Disease Evolution Predictor. This model uses Feature-wise Linear Modulation layers to simulate non-image factors influencing disease evolution, thereby enabling end-to-end prediction and spatial estimation of pathology progression. Similarly, Xiao et al.⁹⁹ utilized conditional recurrent VAE to predict lung cancer tumor growth based on longitudinal imaging data, further demonstrating the potential of integrating image features with clinical context for accurate predictions.

⁹⁴ Griffith SD, Miksad RA, Calkins G, You P, Lipitz NG, Bourla AB, et al. Characterizing the feasibility and performance of real-world tumor progression end points and their association with overall survival in a large advanced non-small-cell lung cancer data set. *JCO Clin Cancer Inform.* 2019;3:1-13.

⁹⁵ Zhang L, Lu L, Summers RM, Kebebew E, Yao J. Convolutional invasion and expansion networks for tumor growth prediction. *IEEE Trans Med Imaging.* 2017;37(2):638-48.

⁹⁶ Li Y, Yang J, Xu Y, Xu J, Ye X, Tao G, et al. Learning tumor growth via follow-up volume prediction for lung nodules. In: *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part VI 23.* Springer International Publishing; 2020. p. 508-17.

⁹⁷ Wang H, Xiao N, Zhang J, Yang W, Ma Y, Suo Y, et al. Static-dynamic coordinated transformer for tumor longitudinal growth prediction. *Comput Biol Med.* 2022;148:105922.

⁹⁸ Rachmadi MF, Valdés-Hernández MDC, Makin S, Wardlaw J, Komura T. Automatic spatial estimation of white matter hyperintensities evolution in brain MRI using disease evolution predictor deep neural networks. *Med Image Anal.* 2020;63:101712.

⁹⁹ Xiao N, Qiang Y, Zhao Z, Zhao J, Lian J. Tumour growth prediction of follow-up lung cancer via conditional recurrent variational autoencoder. *IET Image Process.* 2020;14(15):3975-81.

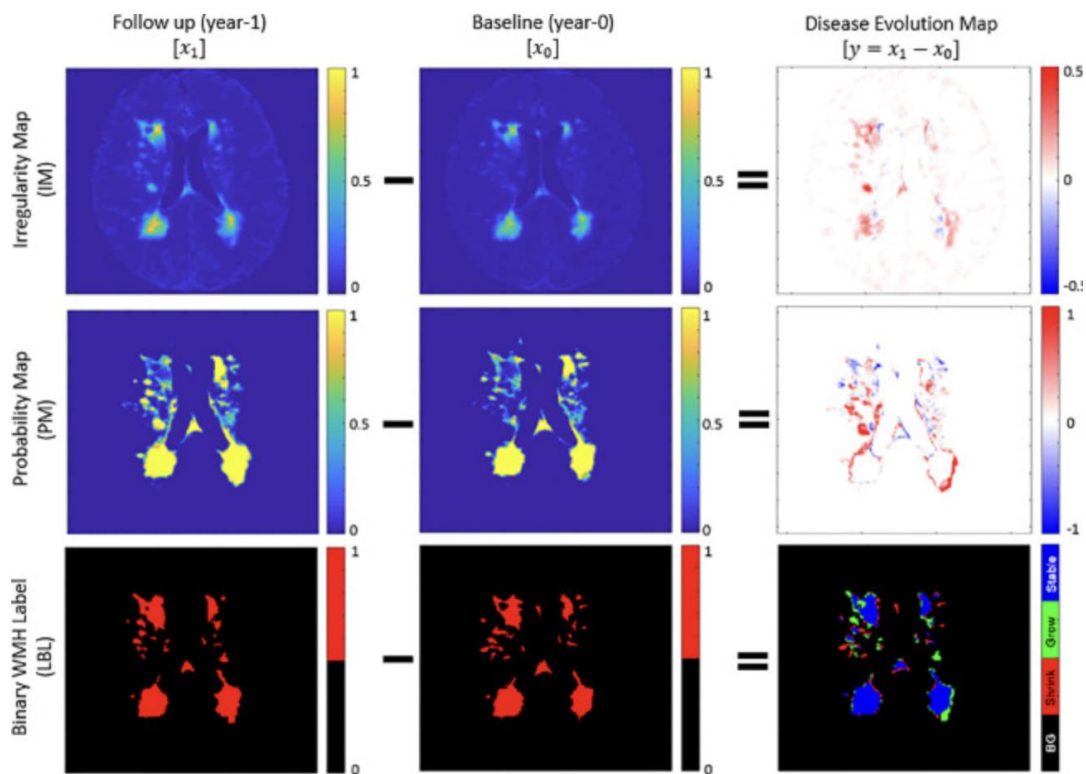


Figure 10: Example of white matter hyperintensities evolution in brain MRI. Source: (Rachmadi et al. 2020)

Notable recent advancements include Palou et al.¹⁰⁰, a comprehensive pipeline for lung cancer nodule analysis, which integrates nodule detection¹⁰¹, re-identification¹⁰², growth estimation¹⁰³, and malignancy classification¹⁰⁴. Their approach employs a range of deep learning techniques to provide a comprehensive understanding of nodule evolution, demonstrating superior performance in predicting tumor growth and malignancy compared to traditional methods such as probabilistic U-Net and Pix2Pix GAN.

An accurate characterization of the evolution of nodules provides medical professionals with robust tools for diagnosis' support and treatment decisions. In particular, image-based deep learning techniques for tumor progression prediction can leverage diagnosis by providing additional and valuable visual clues—unavailable in other statistical methods—for the clinicians to assist and reinforce their decisions. Hence, the integration of clinical and imaging data through advanced computational models offers the potential for significant advancements in personalized medicine and disease management.

¹⁰⁰ Rafael-Palou X, Aubanell A, Ceresa M, Ribas V, Piella G, González Ballester MA. Detection, growth quantification, and malignancy prediction of pulmonary nodules using deep convolutional networks in follow-up CT scans. In: Artificial Intelligence in Cancer Diagnosis and Prognosis, Volume 1: Lung and kidney cancer. Bristol, UK: IOP Publishing; 2022. p. 7-1.

¹⁰¹ Rafael-Palou X, Aubanell A, Bonavita I, Ceresa M, Piella G, Ribas V, et al. Re-identification and growth detection of pulmonary nodules without image registration using 3D Siamese neural networks. Med Image Anal. 2021;67:101823.

¹⁰² Rafael-Palou X, Aubanell A, Bonavita I, Ceresa M, Piella G, Ribas V, et al. Re-identification and growth detection of pulmonary nodules without image registration using 3D Siamese neural networks. Med Image Anal. 2021;67:101823.

¹⁰³ Rafael-Palou X, Aubanell A, Ceresa M, Ribas V, Piella G, González Ballester MA. Detection, growth quantification, and malignancy prediction of pulmonary nodules using deep convolutional networks in follow-up CT scans. In: Artificial Intelligence in Cancer Diagnosis and Prognosis, Volume 1: Lung and kidney cancer. Bristol, UK: IOP Publishing; 2022. p. 7-1.

¹⁰⁴ Rafael-Palou X, Aubanell A, Bonavita I, Ceresa M, Piella G, Ribas V, et al. Pulmonary nodule malignancy classification using its temporal evolution with two-stream 3D convolutional neural networks. arXiv preprint arXiv:2005.11341. 2020.

4. Quality metrics for synthetic data

The quality assessment of synthetic data is closely tied to the application for which it will be used. Ideally, a synthetic dataset would serve as a perfect substitute for real data across all applications. This would require identical distributions in all dimensions and retention of correlations between variables across multiple dimensions. Achieving first-order data patterns, such as distribution moments, is relatively straightforward. For instance, capturing the first and second moments (mean and variance) in a dataset with limited features is less computationally intensive compared to analyses requiring higher-order moments. Higher-order data structures present greater challenges in synthetic data generation, as the complexity increases and may even become intractable. Consequently, generating realistic higher-order data is difficult, necessitating the use of several assessment metrics. Furthermore, a continuous assessment of the dataset within different lines must be performed (an indicative pipeline can be viewed in Figure 11). The subsequent section will delve deeper into higher quality metrics and their relevance to accurate synthetic data generation.

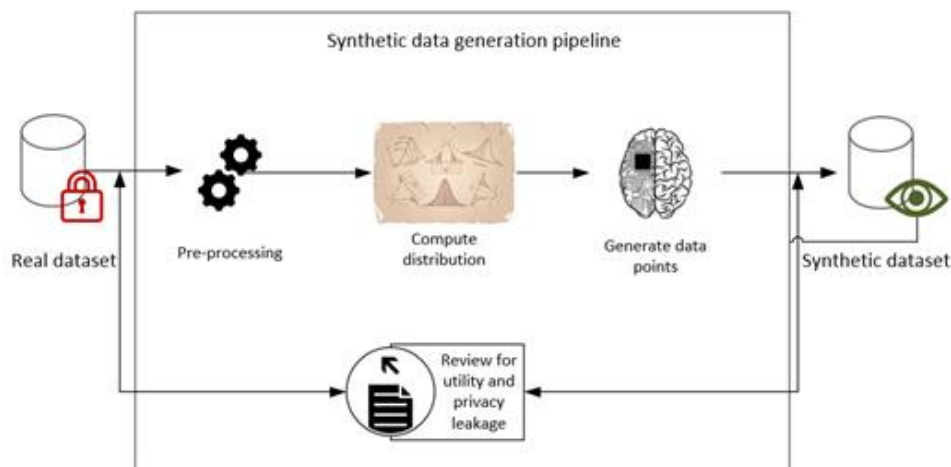


Figure 11: Synthetic data generation pipeline

Each new dataset requires a tailored quality check procedure. Approaches that rely solely on synthetic data must make assumptions about the original dataset before generating synthetic data. This means the synthetic data is designed to mirror the anticipated behaviors and relationships in the original data. However, this process introduces bias, aligning the synthetic data with the initial assumptions. Such bias can be problematic in downstream analyses or modeling, as any inherent biases in the assumptions will be reflected and amplified in the results. Inaccurate or incomplete assumptions can lead to misleading conclusions, highlighting the importance of validating these assumptions and considering potential biases. Combining synthetic and real data or using diverse assumptions can help mitigate bias.

A paradox exists between patient privacy and data utility in synthetic data mapping. Reducing data to lower dimensions through resampling protects patient privacy but diminishes the data's utility for analyses requiring high-dimensional clinical correlations. To advance diagnostics, therapeutics, and treatments that depend on these intricate correlations, substantial investments are necessary. In healthcare, leveraging synthetic data involves navigating challenges related to privacy, data utility, and innovation. Balancing these elements is crucial for harnessing synthetic data's transformative potential in healthcare research and practice.

Robust and reliable quality metrics are urgently needed to assess synthetic data generation approaches. Ideally, these metrics should be generic, but they must also be adapted for specific contexts, such as real-world clinical data. Generally, quality metrics fall into two categories: data realism and privacy (or information disclosure) as can be seen in Figure 12. Data realism assesses how well the synthetic dataset captures the properties of the

real dataset. It is important to note that a universal quality metric is unattainable; metrics must consider the scientific question at hand. Privacy metrics ensure that no private information from the real data is revealed by the synthetic data, which is crucial in low-data regimes. The misconception that synthetic data is inherently private can lead to privacy violations.

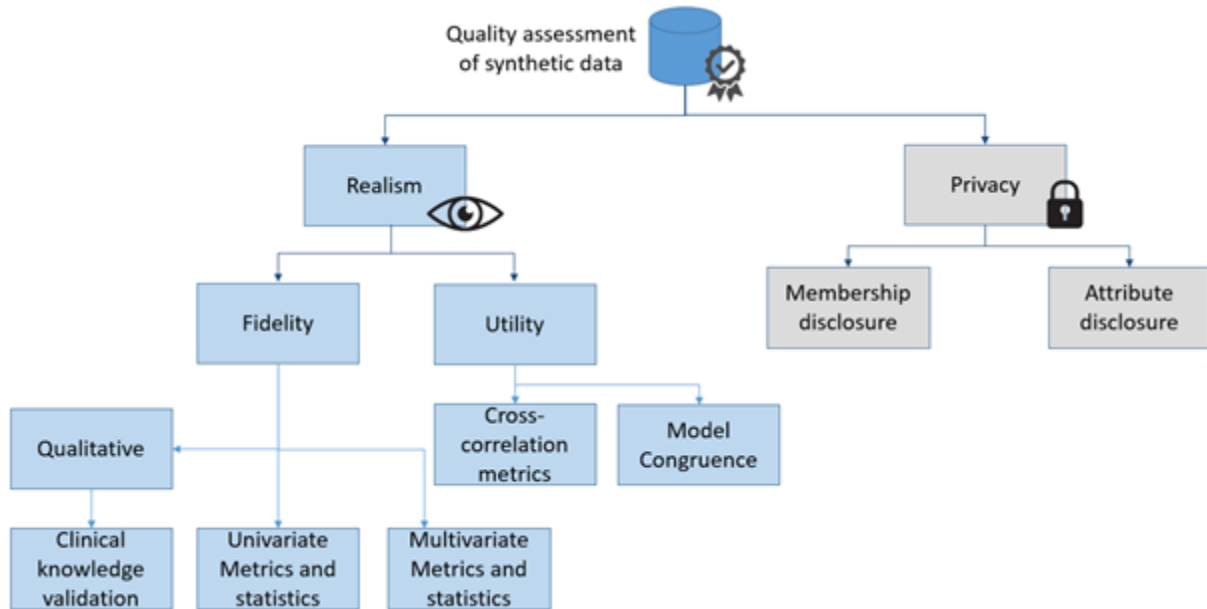


Figure 12: Different types of metrics for quality assessment of synthetic data

Information disclosure metrics are broadly classified into identity disclosure and feature disclosure. Synthetic data should not be linked to the original data, but improper creation can lead to privacy attacks. Identity disclosure involves tracing data points back to the original dataset and identifying individuals. Although technically challenging, such attacks are possible with sufficient resources, posing significant risks. Attribute disclosure involves inferring sensitive features by guessing the original values of synthesized attributes based on known attributes. For instance, attackers can use nearest neighbor methods to infer unknown attributes in fully synthetic data, though the risk is low if the synthetic generation method does not memorize the private dataset.

Data realism metrics are divided into qualitative and quantitative metrics, with the latter being more common. Qualitative metrics involve visual inspection by experts, usually as a final check after privacy and quantitative assessments. However, visual inspection is time-consuming and prone to human error. Quantitative metrics are further divided into fidelity (or resemblance) and utility measures. Each type of those metrics their importance and various examples will be introduced in the following subsections.

In summary, developing effective quality assessment metrics for synthetic data is crucial for ensuring data realism, privacy, and utility. Addressing these challenges will enhance the reliability and applicability of synthetic data in various domains, especially in healthcare.

4.1 Privacy

Privacy plays a critical role in ensuring that synthetic data does not lead to the re-identification of individuals or the exposure of sensitive attributes. There are two primary types of information disclosure metrics: identity (or membership) disclosure and feature (or attribute) disclosure.

Identity (or membership) disclosure refers to the risk of re-identifying individuals in a dataset, which can happen when the similarity between a real record and a synthetic record falls below a certain

threshold, leading the attacker to infer that the real data was used to generate the synthetic data. Techniques like k -anonymity and k -map help reduce this risk by ensuring that an individual's data cannot be distinguished from at least k other individuals in the dataset¹⁰⁵. For example, in k -anonymity, the records in a dataset are generalized so that each record is identical to at least $k-1$ other records, making it difficult for attackers to pinpoint specific individuals. k -map is another approach where the risk is reduced by mapping individuals in the synthetic dataset to a group of at least k individuals in the real dataset, making it harder to trace back to a specific person. This method is particularly important when combining datasets or using external data sources. Similarly, δ -presence defines a bound on how much the presence or absence of a particular record affects the likelihood that a specific individual can be distinguished from others in the dataset¹⁰⁶. A smaller δ value indicates stronger privacy protection, as it implies that the influence of any individual record on the overall dataset is minimal, thus reducing the risk of re-identification.

Feature(or attribute) disclosure refers to the risk that attackers with access to a subset of known real data attributes, can infer remaining unknown sensitive attributes by identifying and analyzing similar records in the synthetic dataset. This becomes particularly dangerous if the synthetic data generation process inadvertently memorizes the original dataset. It is especially true for cases where conditions of l -diversity, which forces group of indistinguishable individuals to have enough diversity, are not satisfied¹⁰⁷. This technique helps protect against attackers who might use majority voting rules or nearest-neighbor methods to infer sensitive attributes.

In conclusion, while synthetic data generation offers a way to protect privacy, it requires careful design and robust privacy metrics to prevent both identity and feature disclosure. Missteps in synthetic data creation could lead to privacy violations, underscoring the need for rigorous methods to protect sensitive information in low-data environments.

4.2 Realism - Fidelity

The fidelity of synthetic data plays a critical role in ensuring that it accurately reflects the statistical properties, relationships, and context of the real data while maintaining privacy. Fidelity metrics are essential for evaluating how well synthetic data replicates the complexity of the original data while offering protection against privacy risks. These metrics can be broadly categorized into two classes: qualitative and quantitative. Qualitative metrics assess the overall structure, visual representation, and expert interpretation of the synthetic data, ensuring that it closely resembles real-world datasets in a meaningful way. Quantitative metrics, on the other hand, involve rigorous statistical tests that evaluate the alignment of distributions, feature relationships, and correlations between synthetic and real data. Additionally, quantitative metrics assess the predictive performance of machine learning models trained on synthetic data, particularly focusing on their effectiveness across diverse subgroups, including minority populations. Ensuring fidelity is crucial for generating synthetic data that is not only statistically valid but also useful in downstream tasks, such as predictive modeling and decision-making, while maintaining fairness across all demographic groups. Moreover, high-fidelity synthetic

¹⁰⁵ Pezoulas VC, Zaridis DI, Mylona E, Androutsos C, Apostolidis K, Tachos NS, Fotiadis DI. Synthetic data generation methods in healthcare: A review on open-source tools and methods. *Computational and Structural Biotechnology Journal*. 2024 Jul 9.

¹⁰⁶ Qian Z, Cebere BC, van der Schaar M. Synthcity: facilitating innovative use cases of synthetic data in different data modalities. arXiv preprint arXiv:2301.07573. 2023 Jan 18.

¹⁰⁷ Murtaza H, Ahmed M, Khan NF, Murtaza G, Zafar S, Bano A. Synthetic data generation: State of the art in health care domain. *Computer Science Review*. 2023 May 1;48:100546.

data can enable better model generalization and provide robustness against biases, offering a reliable alternative for use in research, development, and real-world applications without compromising data privacy.

Utility metrics are used to evaluate the effectiveness and practical value of synthetic data in real-world applications, particularly in tasks like predictive modeling, classification, or decision-making. These metrics assess how well synthetic data can be used as a substitute for real data when training machine learning models or conducting analyses. Utility is often measured by comparing the performance of models trained on synthetic data with those trained on original data, using metrics like accuracy, precision, recall, and F1 score. Additionally, utility metrics examine the consistency of predictions across different subgroups and the ability of synthetic data to generalize across various use cases. A key aspect of utility is ensuring that synthetic data can support both exploratory data analysis and model development without introducing significant biases or compromising model performance, even when dealing with minority or underrepresented groups. High utility ensures that synthetic data provides meaningful insights and reliable outcomes, making it a valuable tool in research and production environments.

4.2.1 Qualitative

Qualitative metrics often involve visual inspection by domain experts. These checks are usually implemented as a final evaluation after the data has been assessed using privacy and quantitative metrics. The primary purpose of qualitative metrics is to examine whether the synthetic data "looks" realistic and aligns with the expected structure of real-world data.

Subjective metrics involve the use of human judgment to assess the realism and plausibility of synthetic data. One prominent example is the **Visual Turing test**, where experts attempt to distinguish between real and synthetic data. In the medical field, this might involve tasks such as the detection of fake nodules inserted into images of healthy lungs. Confusion matrices can be used to categorize the assessment of these images into three main areas: image quality, slice consistency, and anatomic correctness. These assessments were then scored subjectively on a scale of 1-4, as outlined by Khader et al.¹⁰⁸, where higher scores indicate higher perceived accuracy of the synthetic data in mimicking real data.

In addition to the Visual Turing Test, Expert Inspection plays a crucial role in evaluating the quality of synthetic data. This method involves a qualitative assessment by domain experts, such as medical practitioners in the context of healthcare data. The experts manually review each synthetic record, evaluating its clinical plausibility. This includes checking whether the data adheres to known constraints, falls within expected value ranges, maintains realistic feature correlations, and mirrors the correct frequencies of clinical conditions or attributes. A domain expert might inspect whether a synthetic patient diagnosed with a particular disease shows the expected corresponding symptoms, treatment plans, and outcomes. The expert's role is to ensure that the synthetic data adheres to real-world medical knowledge and practices, thus providing a level of assurance that automated methods alone may not achieve.

But due to time constraints, visual inspection is generally performed on a subset of the synthetic data rather than the entire dataset. This type of assessment, although subjective, provides valuable insights into the data's plausibility, especially in complex datasets where statistical metrics alone may not reveal subtle inconsistencies.

4.2.2 Quantitative

Quantitative metrics are primarily used for statistical comparisons between real and synthetic data, ensuring that synthetic data preserves the key characteristics and relationships of the original dataset. These metrics can

¹⁰⁸ Khader F, Müller-Franzes G, Tayebi Arasteh S, Han T, Haarbuerger C, Schulze-Hagen M, Schad P, Engelhardt S, Baeßler B, Foersch S, Stegmaier J. Denoising diffusion probabilistic models for 3D medical image generation. *Scientific Reports*. 2023 May 5;13(1):7303.

be divided into several categories, including univariate and multivariate methods, as well as clustering, latent space representations, and knowledge violation metrics.

4.2.2.1 Univariate Metrics^{109, 110}

- *Kullback-Leibler (KL) Divergence*: Measures the difference between two probability distributions over the same variable. It is used to assess the divergence between real and synthetic data distributions for individual variables.
- *Wasserstein Distance*: A metric for comparing the distributions of two probability distributions, useful for distributions that may not overlap fully. It captures the minimum effort required to transform one distribution into the other.
- *Jensen-Shannon Distance*: A symmetrized and smoothed version of KL divergence that measures the similarity between two probability distributions.
- *Chi-Squared Tests*: Used for categorical variables, this test checks whether the distributions of categorical variables in the synthetic dataset follow the same patterns as those in the real dataset.
- *Kolmogorov-Smirnov Test*: A non-parametric test that compares the cumulative distribution functions of continuous variables in the real and synthetic datasets, ensuring that their distributions are aligned.

4.2.2.2 Multivariate Metrics^{111, 112}

- All above distance metrics if whole dataset distribution is considered.
- *Pairwise Correlation Differences*: These metrics assess whether the correlation structure between variables is preserved in the synthetic data. For example, Pearson and Spearman correlation coefficients can be used to evaluate the strength of the relationships between variables.
- *Covariate Matrix Comparison*: Assesses whether the covariance among variables in the synthetic data matches that in the real data, ensuring that inter-variable relationships are preserved.
- *Maximum Mean Discrepancy*: This statistical test compares the means of two distributions in a feature space to measure how well the synthetic data approximates the real data distribution across multiple variables.
- *Propensity Score Matching*: Evaluates the covariate balance by comparing the propensity scores of the real and synthetic datasets. This metric helps in assessing whether synthetic data provides a balanced representation of the original dataset.
- *Multivariate Probability Density Function Comparisons*: These methods compare higher-order dependencies in both real and synthetic datasets, ensuring that complex interactions between variables are preserved.
- *Support Coverage*: This metric assesses how much of the variable space covered by the real data is also represented in the synthetic dataset. It provides insight into whether the synthetic data sufficiently spans the feature space, reducing the likelihood of missing important information.

4.2.2.3 Space Representations

- *Log-Cluster*: This method evaluates the similarity of clustering structures between real and synthetic datasets. It checks whether the observations in both datasets form similar clusters, preserving the underlying distribution patterns¹¹³.

- *Latent Space Representations*: Metrics such as alpha-precision and β -recall are used to evaluate how well the synthetic data captures the latent (underlying) features of the original data. Latent features often represent abstract relationships within the data that may not be captured by surface-level statistics¹¹⁴.

4.2.2.4 Knowledge Violation and Association Rule Mining

- *Knowledge Violation*: Knowledge violation metrics assess inappropriate disclosures within synthetic data, particularly in specialized domains like clinical data. For example, if synthetic data contradicts established medical knowledge (such as assigning pregnancy to male patients), it indicates a potential failure in the generation process. Knowledge violation metrics ensure that synthetic data aligns with real-world knowledge and context¹¹⁵.
- *Association Rule Mining*: This method assesses how well synthetic data preserves relationships between frequently co-occurring items or features. For example, in medical datasets, association rules might represent common comorbidities. By preserving these relationships, synthetic data can better replicate the contextual structure of the real data.

4.2.2.5 Discriminant Models

Discriminant models are used to assess whether synthetic data can be distinguished from real data. A discriminant model's ability to classify a dataset as real or synthetic serves as an indicator of how well the synthetic data mimics the original data. The less distinguishable the two datasets are, the higher the fidelity of the synthetic data.

4.3 Realism - Utility

The utility of synthetic data is essential for ensuring that models trained on it can generalize well and perform effectively in real-world scenarios. Utility metrics, particularly cross-classification metrics, are vital in evaluating whether models trained on synthetic data perform comparably to those trained on real data. Other metrics such as decision boundary preservation (a post-model verification process), feature ranking agreement,

¹⁰⁹ Hernadez M, Epelde G, Alberdi A, Cilla R, Rankin D. Synthetic tabular data evaluation in the health domain covering resemblance, utility, and privacy dimensions. *Methods of information in medicine*. 2023 Jun;62(S 01):e19-38.

¹¹⁰ Foraker RE, Yu SC, Gupta A, Michelson AP, Pineda Soto JA, Colvin R, Loh F, Kollef MH, Maddox T, Evanoff B, Dror H. Spot the difference: comparing results of analyses from real patient data and synthetic derivatives. *JAMIA open*. 2020 Dec 1;3(4):557-66.

¹¹¹ Woo MJ, Reiter JP, Oganian A, Karr AF. Global measures of data utility for microdata masked for disclosure limitation. *Journal of Privacy and Confidentiality*. 2009 Apr 1;1(1).

¹¹² Heine J, Fowler EE, Berglund A, Schell MJ, Eschrich S. Techniques to produce and evaluate realistic multivariate synthetic data. *Scientific Reports*. 2023 Jul 28;13(1):12266, <https://doi.org/10.1038/s41598-023-38832-0>.

¹¹³ Goncalves A, Ray P, Soper B, Stevens J, Coyle L, Sales AP. Generation and evaluation of synthetic patient data. *BMC medical research methodology*. 2020 Dec;20:1-40.

¹¹⁴ Alaa A, Van Breugel B, Saveliev ES, van der Schaar M. How faithful is your synthetic data? sample-level metrics for evaluating and auditing generative models. In *International Conference on Machine Learning* 2022 Jun 28 (pp. 290-306). PMLR..

¹¹⁵ Yan C, Yan Y, Wan Z, Zhang Z, Omberg L, Guinney J, Mooney SD, Malin BA. A multifaceted benchmarking of synthetic electronic health record generation models. *Nature communications*. 2022 Dec 9;13(1):7609.

and model congruence collectively ensure that synthetic data behaves similarly to real data in predictive modeling tasks.

4.3.1 Cross-Classification Metrics

Cross-classification metrics are fundamental for assessing the performance of machine learning models trained on synthetic data. These metrics involve training models on both synthetic and real datasets, followed by testing on hold-out samples from each. Several machine learning models are commonly employed, including Support Vector Machines (SVM), Logistic Regression (or Ridge Regression), Decision Trees, Random Forests (RF), and Multilayer Perceptrons (MLP). These models help assess whether training on synthetic data leads to comparable results when applied to real-world data¹¹⁶. Model performance is typically measured using standard evaluation metrics such as sensitivity, specificity, and F1-score, making these metrics particularly relevant for classification or prediction tasks.

In addition to accuracy, it's crucial to consider that while improving accuracy, the model's ability to generalize may decrease if the synthetic data lacks sufficient variety. This highlights the need for synthetic data to strike a balance between accuracy and diversity to ensure models trained on it remain robust and generalize effectively across various real-world scenarios¹¹⁷.

4.3.2 Model Sensitivity

The accuracy of models on synthetic data can be model-dependent. For example, models like Random Forests may perform similarly on both synthetic and real data, whereas models like SVM might experience notable drops in performance when trained on synthetic data. To address the variability in model performance on synthetic versus real data, **the Synthetic Ranking Agreement** metric has been introduced. This approach involves training multiple models on both synthetic and real datasets and then comparing their rankings based on feature importance or hyperparameter selection. The agreement in rankings between models trained on real and synthetic data provides insight into the fidelity of the synthetic data for downstream tasks¹¹⁸.

Similarly, **model congruence** evaluates whether models trained on synthetic data produce outputs similar to those trained on real data under comparable conditions. This comparison is particularly important when assessing the transferability of models trained on synthetic data to real-world applications. For example, in healthcare settings, this could mean the difference between accurate and inaccurate patient diagnoses.

4.2.3 Downstream Tasks

In healthcare, where data precision is paramount, the stakes are even higher. Patient outcomes and treatment efficacy are directly influenced by the quality of the data used for decision-making. Populations in healthcare are often heterogeneous, with specific subsets of patients exhibiting unique characteristics or responses to treatments. In these cases, synthetic data that fails to model these subsets accurately can lead to significant disparities in healthcare delivery, potentially compromising patient safety. For instance, synthetic data might be used in lung cancer detection tasks, where accuracy in detection and prediction is critical. The metrics used in these applications go beyond standard detection metrics (e.g., precision, recall) and include geometry metrics like the Jaccard Index, Dice coefficient, Volumetric Similarity, Relative Volume Difference and Mean Surface Distance. Mahalanobis distance is also important when estimating nodule growth in lung cancer

¹¹⁶Du Y, Li N. Towards principled assessment of tabular data synthesis algorithms. arXiv preprint arXiv:2402.06806. 2024 Feb 9.

¹¹⁷Hansen L, Seedat N, van der Schaar M, Petrovic A. Reimagining synthetic tabular data generation through data-centric AI: A comprehensive benchmark. *Advances in Neural Information Processing Systems*. 2023 Dec 15;36:33781-823.

¹¹⁸Giuffrè M, Shung DL. Harnessing the power of synthetic data in healthcare: innovation, application, and privacy. *NPJ digital medicine*. 2023 Oct 9;6(1):186.

predictions. In these scenarios, the use of synthetic data must be rigorously evaluated to ensure it captures the intricate patterns necessary for reliable medical diagnosis and treatment planning¹¹⁹.

You can view a descriptive table with different utility and fidelity metrics in the Table 4.

Table 4: List of fidelity and utility metrics.

Type	Metric	Objective Description	
Fidelity	Univariate	Kullback-Leibler (KL) Divergence *	Measures the difference between two probability distributions over the same variable, assessing divergence from the original data.
		Wasserstein Distance [25]*	A measure of the distance between two probability distributions, useful for comparing distributions that do not overlap fully.
		Jensen-Shannon Distance *	A symmetric version of the KL divergence, providing a smoothed and bounded measure of distributional difference.
		Chi-Squared Tests	Tests the hypothesis that categorical variables in the synthetic data follow the same distributions as those in the original data.
		Support Coverage*	Measures how much of the variable space covered by the real data is also covered by the synthetic data.
		Average of the Absolute Prevalence Differences	Assesses the average of absolute differences in prevalence of features or outcomes between synthetic and original data.
		Kolmogorov-Smirnov Test	A non-parametric test that measures the differences between the cumulative distributions of continuous variables.
	Multivariate	Pairwise Correlation Difference	Compares overall correlation structures between the entire datasets, checking for preservation of inter-variable relationships.
		Covariate Matrix Comparison	Assesses how well the covariance among variables in the synthetic data matches that in the original data.
		Multivariate Probability Density Functions Comparisons	Compares higher-order dependencies and distributions involving multiple variables.
		Propensity Score Matching	Assesses covariate balance by comparing the distribution of propensity scores across the synthetic and original datasets.
		Latent Space Representations	Evaluates how well the synthetic data captures the latent space (underlying features) of the original data.
		Maximum Mean Discrepancy	A statistical test that measures the difference between the means of two distributions in a feature space.
		Association Rule Mining	Analyzes how well synthetic data preserves the relationships between items or features that frequently occur together in the original data.
		Inception Score	Used primarily in evaluating synthetic images, this score assesses the diversity and quality of generated samples.
		Log-cluster	Measures the similarity of the underlying latent structure of the real and synthetic datasets in terms of clustering. Deviations from expected balanced cluster distributions (50% of observations expected to belong to each group for random partitioning) indicate discrepancies.
		* Can be used also as multivariate if the whole dataset distribution is considered	
	Clinical knowledge violation		Evaluates the proportion of records that contradict established medical knowledge, checks for illogical data, like synthetic records assigning pregnancy diagnosis codes to male patients. This metric highlights the synthetic data's realism and guides improvements to ensure it accurately mirrors real-world medical logic.
	Utility	Model Parameter Agreement	Compares model parameters (like weights of neural networks) to determine if models trained on synthetic data converge similarly to those on original data.
		Decision Boundary Preservation	Investigates whether synthetic data maintains the decision boundaries set by models trained on original data.
Feature Ranking Agreement		Measures the consistency in feature importance rankings derived from models trained on synthetic versus original data.	
Cross-classification metrics		Uses traditional metrics like accuracy and F1 score to evaluate the performance of models trained on synthetic (real) data and tested on real (synthetic) data.	
Model Congruency		Evaluates whether models trained on synthetic data produce similar outputs to those trained on original data under comparable conditions.	

¹¹⁹ Usman Akbar M, Larsson M, Blystad I, Eklund A. Brain tumor segmentation using synthetic MR images-A comparison of GANs and diffusion models. Scientific Data. 2024 Feb 29;11(1):259.

4.4 Variety and authenticity

Variety in synthetic data generation can be assessed through metrics that measure the diversity of samples generated across different dimensions, such as class distributions, feature ranges, and subgroup representations. For example, one way to evaluate variety is by examining the coverage of the synthetic data over the feature space compared to the real data, ensuring that all meaningful combinations of features are represented. Metrics like entropy or distributional overlap can be used to quantify how well the synthetic data captures the full breadth of variation in the real dataset. A lack of variety might indicate that the model is only producing a narrow range of outputs, which could lead to poor generalization when the synthetic data is used to train machine learning models.

Authenticity, on the other hand, is often measured using metrics like uniqueness or distance-based approaches. One common method is to compute the similarity between the synthetic data points and the real data, using measures such as nearest-neighbor distance. If synthetic data points are too close to the real data, it could indicate that the model is overfitting, effectively memorizing the training data instead of generating new, original samples. Another approach involves evaluating the proportion of synthetic data points that have no real counterpart in the training set, which can be assessed using novelty detection algorithms. In contexts where privacy is paramount, metrics like membership inference or reconstruction attacks may also be used to evaluate the risk that specific real data points could be re-identified in the synthetic data, ensuring that high authenticity is maintained without compromising privacy.

For instance, in healthcare applications, synthetic patient records should exhibit a variety of demographic characteristics, medical conditions, and treatment outcomes, while authenticity ensures that none of the generated records are identical to real patients' confidential information. Similarly, in financial data, the synthetic dataset should cover a wide range of transaction types and amounts, while maintaining authenticity to avoid generating records that replicate actual transactions. Balancing variety and authenticity is crucial for producing high-quality synthetic data that is both representative of the real world and safe for use in sensitive domains.

5. DaaS Technologies Integration

This section illustrates how de-identification, synthetic data generation, and evaluation technologies are currently integrated and are planned to be integrated with other work packages. In particular, it outlines their integration with Data as a Service (DaaS), as well as their interconnection with Model as a Service (MaaS) and Health Data Hub (HDH). Since technology integration is not the main focus of this deliverable, only high level schemas are presented. For more specific information on the MaaS, DaaS, and HDH technologies, please refer to the deliverables 3.6, 4.6, and 2.7 respectively.

5.1 DaaS Toolbox Integration

The DaaS Toolbox is a curated and searchable catalogue of data services, tools, and resources developed within WP3 of the PHASE IV project. It provides an intuitive entry point for researchers, developers, and healthcare professionals to discover, evaluate, and utilize AI-related data tools and components essential for PHASE IV use cases, particularly in the post-market (Phase IV) setting.

At its core, the DaaS Toolbox builds upon open-source Landscape2¹²⁰, a modular, extensible platform co-developed with WP4 (Model as a Service – MaaS) and implemented as a tailored fork of the Landscape2 tool. The Landscape serves both toolboxes (DaaS and MaaS), ensuring a consistent user experience and seamless navigation across both domains.

The DaaS Toolbox offers search and filtering functionalities, enabling users to locate the most relevant tools and use case-oriented services for tasks such as data harmonization, anonymization, synthesis, and data quality assessment—key components required in AI model training, validation, and regulatory analysis in healthcare.

Each tool or service in the DaaS Toolbox is represented as a technology card, which encapsulates detailed metadata including tool description and functionality, associated use case(s), maturity level and documentation, technical interfaces and dependencies, and licenses and access modalities.

The structure and metadata schema of these cards are documented in Deliverable D3.6, while the final operational implementation will be detailed in Deliverable D3.7.

The current version of the DaaS Toolbox is currently hosted on the project GitLab instance maintained by INESC TEC, with a fully automated CI/CD deployment pipeline. This pipeline ensures that any updates to the metadata source files—submitted by project partners—are immediately reflected in the active toolbox instance.

5.2 Data and service operability

Synthetic datasets generated from services in DaaS and curated within the PHASE IV AI project will be stored, annotated, and made discoverable by using open-source frameworks for comprehensive healthcare data management and analysis, e.g. OBiBa and/or OpenMetadata. Post-processing, metadata records should be ingested into data cards within data catalogues in a compliant way with open-source frameworks, which includes fields such as dataset ID, title, description, owner, license, data type(s), data origin, or tags. Synthetic data will additionally consider specific fields such as a reference to source data and generator models and data quality store(s) to ensure traceability and auditability compliance. A preliminary list of UC-oriented metadata fields is described in D2.1 Data Specifications.

¹²⁰ <https://github.com/cncf/landscape2>

5.2.1 Operability for Tabular Data

Technologies for generating tabular synthetic data will be provided through a software package. This package will be based on the synthetic data generation process described in Section 3 and will include the following core modules:

- **Data Acquisition:** Uploading real, pseudonymized tabular datasets.
- **Data Preparation:** Preprocessing the uploaded data to ensure it is ready for synthesis.
- **Modeling:** Selecting and training synthesizers to generate synthetic data.
- **Evaluation:** Assessing the synthetic data in terms of fidelity, privacy, and utility.

The software package will be referenced in the DaaS toolbox landscape item cards. Instructions and examples will be provided to help interested parties download, install, and use the tool. This package is intended for users who have access to pseudonymized real datasets and aim to further anonymize them by generating synthetic equivalents.

5.2.2 Operability for Image Data

Services related with the synthetic imaging data generation will enable users to generate privacy-preserving, high-fidelity synthetic medical images. This service is offered through the HDH and exposed via the DaaS Toolbox as a searchable card in the offering catalogue. Imaging synthesis service will implement the following core stages:

- **Data acquisition:** users should be able to upload medical imaging datasets (e.g. DICOM, NifTI) into a secure, encrypted storage vault within the HDH. Imaging-specific metadata should be collected (e.g., modality, resolution, institution, scan region) ingested via an specific data ingestion module.
- **Data preparation:** prepares the images for the synthesis through standardization of image dimensions and voxel spacing, image intensity processing, optionally alignment of masks and labels. Transformations steps must be stored as part of the data trace.
- **Model training:** image synthesis models are trained on the uploaded dataset using generative models, e.g. GANs or Diffusion Models. The user could chose a model and select default presets, including training epochs or image resolution. Training occurs on HDH federated learning node infrastructure with containerized environments.
- **Model inference:** user can simply perform inference on pre-trained models to generate new synthetic samples on demand.
- **Evaluation:** this allows to assess generated synthetic images across several dimensions, i.e. image fidelity quality —typically 2D and 3D Fréchet Inception Distance— data utility/privacy metrics, e.g. cross-MS-SSIM variability and other inter-sample similarity distances, useful for downstream tasks such as classification or segmentation. The output should conform also a structured evaluation report linkable from the HDH catalog. Data quality metrics should be also registered as part of the metadata to qualify data for public release or licensing.

5.3 DaaS and MaaS Toolbox Integration

The MaaS Toolbox complements the DaaS Toolbox by serving as the centralized catalogue and interaction layer for AI/ML models developed within the PHASE IV project. Both toolboxes are built on a shared, customized fork of the open-source Landscape2 tool, allowing for a harmonized user experience and technical interoperability.

While the DaaS Toolbox focuses on tools and workflows related to data preparation, generation, and quality assurance, the MaaS Toolbox organizes and presents trained machine learning models through a card-based interface. They will be accompanied by metadata, documentation, and secure access links and stored via framework such as OBiBa. Together, these two toolboxes form the foundation of an end-to-end AI pipeline interface, where data services and models are semantically and operationally linked. This guarantees a unified and easily supported common framework which combines work from WP3, WP4 and WP5 in an efficient, user friendly, and secure environment. For more information on MaaS please refer to Deliverable 4.6.

5.3.1 Shared Metadata Structures and Card Framework

Both toolboxes utilize a common metadata schema adapted to their specific domain (data services vs. models), allowing consistent rendering and filtering of content across the Landscape platform. Cards in the DaaS Toolbox may include references to models (hosted in MaaS) that use the corresponding data service or transformation pipeline, while MaaS cards may list upstream tools from DaaS that are required for proper data pre-processing or harmonization. For example, a MaaS card presenting a prostate cancer risk prediction model could reference specific DaaS tools used to harmonize EHR data or synthesize new anonymized lab results.

5.3.2 Cross-Linking Between Toolboxes

Model entries in the MaaS Toolbox can include dynamic links to related tools in the DaaS Toolbox, such as required pre-processing pipelines, input data harmonization services, or synthetic data generators for privacy-preserving model testing. Likewise, DaaS cards can indicate downstream models that rely on their outputs, giving users a full lineage from raw data to deployed model. This bidirectional linking not only improves the usefulness of the tool by also supports traceability, reproducibility, and transparency, especially important in clinical AI settings.

5.3.3 Co-Development and Extensibility

Because both toolboxes share a common software base, new features such as advanced filtering and tagging, dependency mapping, or model retraining triggers based on data updates, it can be developed once and deployed across both interfaces. This ensures low overhead for maintenance and facilitates cohesive evolution of the ecosystem.

5.3.4 Use Case–Driven Organization and Semantic Alignment

Both DaaS and MaaS toolboxes are organized around the PHASE IV Use Cases (UCs), providing users with a vertically integrated view of the pipeline, including data services supporting UC-specific preparation steps as well as models trained and validated for UC-specific predictions or classifications.

This structure facilitates domain-specific navigation, ensuring that healthcare professionals or developers working on a particular UC (e.g., nodule growth prediction to support diagnosis in lung cancer) can quickly find both the data tools and AI models relevant to their problem.

5.4 Interfacing via Health Data Hub (HDH)

Both DaaS and MaaS toolboxes will be progressively integrated with the HDH infrastructure. The MaaS Toolbox periodically imports model metadata and usage documentation from the HDH model catalogue API to ensure freshness and alignment with central repositories. The DaaS Toolbox, in turn, will transition from GitLab-based static metadata to a dynamic integration with the HDH catalogue API for service discovery. Through this shared backend integration, both toolboxes can remain synchronized with the HDH governance,

leveraged by its authentication and authorization services, data access and usage policy management, or audit trails for sensitive health data usage. This will enable secure and compliant model deployment workflows that use inputs from verified data preparation services.

5.4.1 Model Execution and Access Interfaces

Each MaaS card includes a secure HTTPS link to the model's execution endpoint or API, which may reside on the HDH infrastructure or another federated node. This access layer may support real-time inference, batch processing, interactive demo environments (e.g., notebooks). These endpoints are expected to operate on data that has passed through DaaS workflows, such as harmonized or anonymized patient records. MaaS model endpoints may even call DaaS services programmatically as pre-processing steps before applying the model—further tightening integration.

5.4.2 Metadata Flow and Alignment with HDH Catalog

The DaaS Toolbox should be connected with the HDH ecosystem through an Ingestion Connector, which queries or pulls metadata from DaaS tools (harmonization, anonymization, synthesis, etc.) and pushes metadata into the Catalog Service of the HDH, e.g. via the OpenMetadata non-public interface. This ensures that DaaS tools are indexed, discoverable, and queryable through the central HDH backend catalogue (like all other AI and data assets).

Metadata can be stored in a dedicated DaaS-specific warehouse, maintaining descriptive metadata for service discovery, some technical specifications, and linked to specific UCs and potentially other dependencies such as models or pipelines. A similar structure for MaaS modules and dataset ingestion interfacing with their own metadata layers will enable a harmonized ecosystem.

Instead of just cataloguing static tools, dynamic invocation of DaaS services from MaaS pipelines or HDH workflows can be enabled (e.g. calling an anonymization service before model execution). In this interaction, each execution should be logged and billed using smart contract. Furthermore, DaaS and MaaS should also be connected to traceable chains, e.g. provided by frameworks such as OBiBa, in a reliable environment showing which DaaS services were used to prepare the data that a MaaS model was trained or validated on.

5.4.3 Deployment and Access Interface via the Portal and Marketplace

The DaaS Toolbox requires an externally accessible HDH Portal UI, leveraging the same user entry point as the MaaS and Dataset ingestions modules. Once DaaS services are catalogued they are ready to be accessible in an offering catalogue generated with the Landscape tool. Also they can be exposed to authorized users (e.g. data scientists, clinicians) and include linked technical documentation and service-level details. From here, services can be accessed directly via secure endpoints, deployed on demand, and monetized or regulated via smart contracts.

6. Conclusion

In this initial version of the D3.3 deliverable, we present and benchmark the principle de-identification and synthetic data generation approaches. We also outline the advantages and limitations of state-of-the-art technologies, such as differential privacy. Furthermore, we emphasize the importance of establishing a comprehensive framework to evaluate various aspects of synthetic data, including privacy, fidelity, and utility. In addition, we discuss how de-identification, synthetic data generation, and evaluation technologies developed within WP3 will be integrated with other work packages.

Future versions of this deliverable will explore the applicability of these methods to the use cases in WP6 (UC1, UC2, UC3) and assess their potential for handling heterogeneous data in de-centralized way. The upcoming versions will also include a list of services as a separate Annex.