UNIVERSIDAD POLITÉCNICA DE MADRID
Escuela Técnica Superior de Ingenieros de Telecomunicación

Deep Generative Models for Survival Analysis and Synthetic Data Generation in Healthcare

D O C T O R A L T H E S I S

Submitted for the degree of Doctor by:

Patricia Alonso de Apellániz
Máster Universitario en Ingeniería de Telecomunicación

Madrid, 2025


UNIVERSIDAD POLITÉCNICA DE MADRID
Escuela Técnica Superior de Ingenieros de Telecomunicación

Doctoral Degree in Communication Technologies and Systems

Deep Generative Models for Survival Analysis and Synthetic Data Generation in Healthcare

D O C T O R A L T H E S I S

Submitted for the degree of Doctor by:

Patricia Alonso de Apellániz
Máster Universitario en Ingeniería de Telecomunicación

Under the supervision of:
Dr. Juan Parras Moral

Madrid, 2025

Title: Deep Generative Models for Survival Analysis and Synthetic Data Generation in Healthcare

Author: Patricia Alonso de Apellániz

Doctoral Programme: Communication Technologies and Systems

Thesis Supervision:

Dr. Juan Parras Moral, Professor, Universidad Politécnica de Madrid (Supervisor)

External Reviewers:

Thesis Defense Committee:

Thesis Defense Date:

This thesis has been partially supported by the European Union's Horizon 2020 Research and Innovation Program under grant agreements No. 101017549 (GenoMed4All project) and No. 101095530 (SYNTHEMA project), the UPM Research Programme 'Programa Propio de I+D+i 2023', and also by the UPM for a teaching position (Assistant Professor).

A Nacho, por tener más ganas que yo de que esto terminara.
Y, por supuesto, a mí, porque es mi tesis.

Agradecimientos

Llegar hasta aquí ha sido como volver a aprender a caminar. Todo empezó en esa etapa en la que uno va a gatas (aunque, sinceramente, no lo tengo del todo claro, porque obviamente no me acuerdo de cómo aprendí a caminar). Como entonces —imagino— este proceso ha estado lleno de caídas, resbalones, narices contra el suelo y esa sensación de que la cabeza pesa demasiado y te hace caer hacia adelante. Pero, si he conseguido caminar otra vez, no ha sido solo mérito mío. Ha habido muchas manos que me ayudaron a dar los primeros pasos, muchas voces que me animaron a no rendirme cuando tropezaba y, por supuesto, un buen puñado de empujones estratégicos (algunos más sutiles que otros).

Para empezar, gracias a mi mejor amiga, Lau, por meterme en este lío. Sí, fuiste tú quien me convenció de que embarcarse en un doctorado era una gran idea. Pero también fuiste tú quien estuvo ahí siempre, guiándome con tu experiencia y tus consejos. Gracias por demostrarme que, al final, se sobrevive y que, además, uno puede sentirse orgulloso del camino recorrido. Esta tesis no sería lo que es sin ti, y yo tampoco.

A Juan, mi tutor, no podría haber pedido un mejor guía. Siempre dispuesto, pendiente y, sobre todo, relajado, justo el equilibrio que necesitaba para avanzar sin agobios. Tu manera de acompañar sin presionar ha hecho que este proceso sea llevadero e incluso gratificante. Y a Santi, gracias por ser un ejemplo de comprensión y por crear un entorno donde aprender, explorar y entender siempre son prioridades. Tu actitud sincera, tu coherencia y la forma en que valoras nuestras ideas me han hecho sentir escuchada y apoyada en cada paso.

A mis niños y niñas del GAPS, el grupo que empezó siendo solo dos y ahora es un puzle lleno de piezas únicas y perfectamente encajadas. En especial, gracias a Almo y Mats, que me habéis aguantado desde el principio, escuchando mis quejas (muchas y nada suaves). Vuestra paciencia y compañía han sido esenciales. A mi familia y amigos, gracias por estar ahí, a pesar de no tener ni la más remota idea de qué he estado haciendo exactamente. Gracias por preguntar *"¿y cuánto te falta?"* cuando apenas llevaba un par de meses, y por esas caras cada vez que intentaba explicar algo. Sé que siempre habéis estado orgullosos de mí.

A Nacho, gracias por estos cuatro años en los que te has doctorado en la gestión del doctorado de tu pareja. Ha sido un camino difícil, lleno de desafíos y muchas horas de paciencia (sobre todo por tu parte). Y, como siempre, lo has superado logrando un sobresaliente *cum laude*. Gracias por estar a mi lado en cada reto que emprendo, por soportar mis quejas y mis nervios. Pero, por fin, este capítulo se cierra. Ya puedes respirar tranquilo... ¿o quizá no?

Para ir acabando, podría escribir páginas agradeciéndome a mi misma por haber llegado hasta aquí, a pesar de las inseguridades que he arrastrado. Durante mucho tiempo pensé que no había empezado este camino en el mejor momento de mi vida, pero, a día de 15 de enero de 2025 (mientras escribo esto), estoy segura de que fue una de las mejores decisiones que tomé. Así que gracias a la Pati de hace cuatro años por atreverse, y a la Pati de estos cuatro años por no rendirse nunca. ¡Menos mal!

Y, por último y por encima de todo, gracias al tiempo. Aunque parecía detenerse, siguió avanzando. **Y aquí estamos: todo llega y esto también llegó a su fin.**

Abstract

Healthcare systems worldwide face persistent inequities, with disparities in access, representation, and quality disproportionately affecting marginalized populations. Addressing these challenges requires innovative solutions to overcome data scarcity, enhance collaboration, and improve predictive modeling in medical research. This doctoral thesis advances generative AI methodologies, focusing on tabular data—an essential yet underexplored type of healthcare information. Tabular data encompass patient demographics, clinical histories, and treatment outcomes, making them crucial for equitable healthcare delivery. The research leverages Variational Autoencoders (VAEs) as a foundational framework due to their ability to model complex, high-dimensional relationships and handle missing information. This thesis contributes across three interconnected domains: Survival Analysis (SA), Synthetic Data Generation (SDG), and Federated Learning (FL), demonstrating how these approaches collectively address key gaps in healthcare research.

In SA, VAE-based models such as SAVAE and CR-SAVAE address traditional limitations, including proportional hazard assumptions and censored data. These models improve time-to-event predictions and incorporate competing risks, enabling more precise analyses of patient outcomes and enhancing personalized care. In SDG, this thesis integrates VAEs with Bayesian Gaussian Mixtures, transfer learning, and meta-learning to generate high-quality synthetic tabular data. These methods tackle challenges such as mixed data types, small sample sizes, and class imbalances. Validation frameworks combining statistical and task-specific metrics ensure the reliability of synthetic data, empowering resource-limited institutions to contribute to medical research while preserving privacy. In FL, the Federated Synthetic Data Sharing (FedSDS) framework enables privacy-preserving collaboration across decentralized institutions. By generating synthetic data locally with VAE-based models, FedSDS mitigates data heterogeneity and imbalances, ensuring robust model training in IID and non-IID settings. This approach bridges the gap between data-rich and data-scarce institutions while safeguarding patient confidentiality. The contributions across SA, SDG, and FL are deeply interconnected, forming a cohesive framework to tackle systemic challenges in healthcare. By integrating these methodologies, the thesis demonstrates improved predictive accuracy, scalability, and equity in AI-driven healthcare applications. The research outcomes highlight the potential of generative AI to drive equity and innovation in medical research and practice.

Looking ahead, this thesis outlines key directions for future work, including integrating frailty models into SA to capture unobserved patient heterogeneity, extending methodologies to multi-modal datasets like imaging and genomics, and enhancing privacy in SDG through differential privacy or homomorphic encryption. It also highlights the importance of adaptive FL strategies and public repositories for high-quality synthetic datasets to drive equitable healthcare solutions globally.

This thesis lays a robust foundation for leveraging generative AI to reduce healthcare inequities by addressing key challenges in data scarcity, heterogeneity, and collaboration. Its contributions pave the way for meaningful applications, fostering inclusive, scalable, and globally accessible healthcare systems.

Resumen

La atención sanitaria enfrenta desafíos globales, especialmente en contextos con recursos limitados, donde las herramientas médicas y tecnológicas no siempre cubren las necesidades. Estas dificultades afectan de manera desproporcionada a poblaciones vulnerables, con datos que reflejan sesgos o carecen de representación adecuada. Superar estas barreras requiere soluciones innovadoras que aborden la escasez, heterogeneidad y necesidad de colaboración entre instituciones. Esta tesis desarrolla metodologías avanzadas de Inteligencia Artificial (AI) generativa, enfocándose en datos tabulares, esenciales en salud por su información sobre demografía, historiales médicos y tratamientos. Se emplean Autoencoders Variacionales (VAEs) por su capacidad para modelar relaciones complejas en datos de alta dimensionalidad y manejar información faltante. La tesis aporta avances en Análisis de Supervivencia (SA), Generación de Datos Sintéticos (SDG) y Aprendizaje Federado (FL), demostrando cómo estas metodologías abordan desafíos clave en la investigación en salud.

En SA, modelos basados en VAE como SAVAE y CR-SAVAE superan limitaciones tradicionales, mejorando la predicción del tiempo hasta el evento e incorporando riesgos en competencia para análisis más precisos y atención personalizada. En SDG, esta tesis combina VAEs con Mezclas Gaussianas Bayesianas, aprendizaje por transferencia y *meta-learning* para generar datos sintéticos de alta calidad, abordando la heterogeneidad de datos, el tamaño reducido de muestras y el desequilibrio de clases. Marcos de validación que integran métricas estadísticas y específicas de la tarea garantizan la fiabilidad de los datos sintéticos, permitiendo que instituciones con recursos limitados contribuyan a la investigación sin comprometer la privacidad. En FL, *Federated Synthetic Data Sharing* (FedSDS) facilita la colaboración descentralizada preservando la privacidad. Al generar datos sintéticos localmente con modelos VAE, FedSDS mitiga la heterogeneidad y los desequilibrios en los datos, garantizando un entrenamiento robusto en entornos IID y no-IID. Esta estrategia reduce la brecha entre instituciones con diferentes niveles de acceso a datos, promoviendo una colaboración equitativa sin comprometer la confidencialidad de los pacientes. Las contribuciones en SA, SDG y FL están interconectadas, formando un marco integral para abordar desafíos en salud. Al integrar estas metodologías, se mejora la precisión predictiva, la escalabilidad y la equidad en aplicaciones de AI para la atención médica, demostrando el potencial transformador de la AI generativa en la innovación y equidad en salud.

Esta tesis identifica varias líneas futuras de investigación, como la integración de modelos de fragilidad en SA para capturar heterogeneidad no observada y la extensión de las metodologías a datos multimodales, como imágenes médicas. También plantea el avance en garantías formales de privacidad en SDG mediante privacidad diferencial o cifrado homomórfico. Además, destaca la importancia de estrategias adaptativas en FL y la creación de repositorios públicos de datos sintéticos de alta calidad, impulsando soluciones sanitarias más equitativas a nivel global.

Al abordar la escasez de datos, la heterogeneidad y la necesidad de colaboración, esta tesis sienta las bases para aplicar la AI generativa en la reducción de desigualdades en salud, abriendo nuevas posibilidades para desarrollar aplicaciones transformadoras y fomentando una atención sanitaria más inclusiva, escalable y accesible.

Table of Contents

List of Figures

List of Tables

Abbreviations and acronyms

ABN Acyclic Bayesian Network

ADASYN Adaptive Synthetic Sampling

AE Autoencoder

AFT Accelerated Failure Time

AI Artificial Intelligence

BGM Bayesian Gaussian Mixture

BMI Body Mass Index

BN Bayesian Network

BRFSS Behavioral Risk Factor Surveillance System

BS Brier Score

CDF Cumulative Distribution Function

C-index Concordance Index

CI Confidence Intervals

CIF Cumulative Incidence Function

CNN Convolutional Neural Network

CoxPH Cox Proportional Hazards

CR Competing Risks

CR-SAVAE Competing Risks - Survival Analysis Variational Autoencoder

CRESA Competing Risks and Recurrent-Event Survival Analysis

CT Computed Tomography

CTGAN Conditional Tabular Generative Adversarial Network

$\boldsymbol{D}_{\mathbb{JS}}$ Jensen-Shannon Divergence

$\boldsymbol{D}_{\mathbb{KL}}$ Kullback-Leibler Divergence

DGM Deep Generative Model

DL Deep Learning

DNN Deep Neural Network

DP Differential Privacy

DRS Domain Randomized Search

DS Directional Symmetry

DT Decision Tree

EBMT European Society for Blood and Marrow Transplantation

EHR Electronic Health Record

ELBO Evidence Lower Bound

EM Expectation-Maximization

EMR Electronic Medical Record

EN Elastic Net

EU European Union

EUSIPCO European Signal Processing Conference

FDA Food and Drug Administration

FG Fine-Gray

FedAvg Federated Averaging

FedMA Federated Matched Averaging

FedNova Federated Normalized Averaging

FedSAVAE Federated Survival Analysis Variational Autoencoder

FedSDS Federated Synthetic Data Sharing

FedSGD Federated Stochastic Gradient Descent

FedVAE Federated Variational Autoencoder

FL Federated Learning

FLChain Free Light Chain

FTL Federated Transfer Learning

FWER Family-Wise Error Rate

GAN Generative Adversarial Network

GBSG German Breast Cancer Study Group

GBST Gradient-Boosted Survival Trees

GDPR General Data Protection Regulation

GM Generative Model

GMM Gaussian Mixture Model

GPU Graphics Processing Unit

HE Homomorphic Encryption

HFL Horizontal Federated Learning

HIPAA Health Insurance Portability and Accountability Act

IBS Integrated Brier Score

ICST International Conference on Software Testing, Verification and Validation

IID Independent and Identically Distributed

IoT Internet of Things

IPCW Inverse Probability of Censoring Weighting

JCR Journal Citation Reports

JSD Jensen-Shannon Distance

KDE Kernel Density Estimation

KM Kaplan-Meier

KNN K-Nearest Neighbors

KS Kolmogorov-Smirnov

LLM Large Language Model

LT Life-Table

MAE Mean Absolute Error

MAML Model-Agnostic Meta-Learning

MC Monte Carlo

MCMC Markov Chain Monte Carlo

MEDLARS Medical Literature Analysis and Retrieval System

Metabric Molecular Taxonomy of Breast Cancer International Consortium

MGUS2 Monoclonal Gammopathy of Unknown Significance

ML Machine Learning

MLE Maximum Likelihood Estimation

MLP Multilayer Perceptron

MMD Maximum Mean Discrepancy

MPC Multi-Party Computation

MRR Mean Reciprocal Rank

MRI Magnetic Resonance Imaging

NA Nelson-Aalen

NIH National Institutes of Health

NIMHD National Institute on Minority Health and Health Disparities

NLP Natural Language Processing

NMDS Non-Metric Multi-Dimensional Scaling

NN Neural Networks

nODE Neural Ordinary Differential Equation

NWTco National Wilm's Tumor Study

PAHO Pan American Health Organization

PATE Private Aggregation of Teacher Ensembles

PBC Primary Biliary Cholangitis

PCA Principal Component Analysis

pdf Probability Density Function

PEIRS Pathology Expert Interpretative Reporting System

PET Positron Emission Tomography

PII Personally Identifiable Information

PMF Probability Mass Function

PPC Pairwise Pearson Correlation

RBF Radial Basis Function

RDT Randomized Decision Tree

ReLU Rectified Linear Uni

RF Random Forest

RKHS Reproducing Kernel Hilbert Spaces

RNN Recurrent Neural Networks

RR Reciprocal Rank

RSF Random Survival Forest

SA Survival Analysis

SaMD Software as a Medical Device

SAVAE Survival Analysis Variational Autoencoder

SDC Statistical Disclosure Control

SDG Synthetic Data Generation

SGD Stochastic Gradient Descent

SMOTE Synthetic Minority Oversampling Technique

SSMTL Semi-Supervised Multi-Task Learning

STD Sexually Transmitted Diseases

STDG Synthetic Tabular Data Generation

STS Short Time-Series Distance

SUMEX-AIM Stanford University Medical Experimental Computer for AI in Medicine

Support Study to Understand Prognoses, Outcomes, and Risks of Treatment

SVM Support Vector Machine

SVRc Support Vector Regression for Censored Data

TVAE Tabular Variational Autoencoder

TRTS Train on Real, Test on Synthetic

TSTR Train on Synthetic, Test on Real

TVAE Tabular Variational Autoencoder

VAE Variational Autoencoder

VAE-BGM Variational Autoencoder - Bayesian Gaussian Mixture

VFL Vertical Federated Learning

WGAN Wasserstein Generative Adversarial Network

WHAS Worcester Heart Attack Study

WHO World Health Organization

XAI Explainable Artificial Intelligence

Chapter 1

Introduction

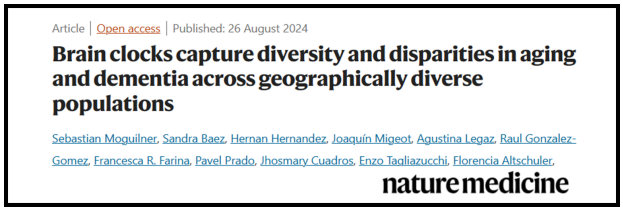
"Of all the forms of inequality, injustice in health is the most shocking and the most inhuman because it often results in physical death." —Rev. Dr. Martin Luther King Jr. (1966)

These words, spoken by Rev. Dr. Martin Luther King Jr. at the second convention of the Medical Committee for Human Rights, were a rallying cry against racial discrimination in healthcare. There are different versions of this documented quote. Some referred specifically to healthcare, while others excluded the 'physical death' part. There is no formal record of exactly what he said, only what was captured in news stories. Regardless of the specific words King used, which we may never know for sure, his message remains a powerful condemnation of systemic injustice that denied African Americans access to quality healthcare and cost countless lives.

In 2024, health inequities remain a pressing issue. While the explicit racial segregation of King's time has largely diminished (yet not faded), his message remains profoundly relevant, albeit in a different form. The inequities of the modern era are not defined by explicit denial of care but by invisible barriers: disparities in access to information and resources, representation in healthcare research, and quality of care provided. The headlines presented in Figure 1.1 underscore the diversity and urgency of these disparities. From gender-based inequities to geographic and economic disparities in cancer care, ageing, and dementia studies, these examples demonstrate the far-reaching impact of systemic health inequities. Similarly, Figure 1.2 shows the unmet healthcare needs in the European Union (EU) countries in 2023, revealing significant disparities for certain population groups despite universal coverage of basic services. Variations in service range and cost-sharing create affordability challenges, with unmet needs over three times higher among the lowest income groups than the highest. If this occurs in the EU, the situation will likely be far worse in developing countries. These disparities, drawn from the 'Health at a Glance: Europe 2024' report [1] by the European Commission, and the recent articles—published less than six months before the writing of this text—underscore the global impact of systemic health inequities. They are a stark reminder of the gaps that persist despite technological and scientific progress.

(a) World Health Organization (WHO)

(b) World Economic Forum



(c) Nature Medicine

(d) Cancer Journal

(e) Health

(f) Pan American Health Organization (PAHO)

Figure 1.1: Recent (2024) headlines highlighting global health inequities. These include disparities in cancer care, gender-based inequalities, geographic gaps in dementia research, and the disproportionate burden of early death among marginalized populations, emphasizing the need for equitable healthcare policies and systemic solutions.

Efforts to address these inequities have gained increasing attention on global platforms. The World Health Organization (WHO), a leading advocate for health equity, emphasizes the importance of harnessing data to advance health equity. Their upcoming book, 'Health Inequality Monitoring: Harnessing Data to Advance Health Equity,' highlights how robust data collection and analysis are indispensable for identifying and addressing health disparities[1]. Similarly, the National Institute on Minority Health and Health Disparities (NIMHD) underscores the critical role of data in tackling health inequities. As its director, Dr. Eliseo Pérez-Stable, recently stated, *"Ending health disparities starts with good data,"* stressing that **data are not just a resource but a tool for empowerment and change**[2]. This report elaborates on the profound connection between data and equitable healthcare. A pivotal statement, 'Communities deserve connections,' captures the essence of the issue: healthcare inequities cannot be addressed in isolation. Communities need access to accurate, timely, and representative data to understand their challenges and advocate for solutions. This notion resonates deeply with the philosophy behind Artificial Intelligence (AI)-driven healthcare innovation. By enabling communities to connect with their data, **AI has the potential to empower individuals and regions to overcome barriers** that have persisted for decades.

[1]Source: WHO Newsroom: Health Inequality Monitoring book (Accessed on December 5th, 2024)
[2]Source: UC Merced Newsroom: Health Disparities by NIMHD director (Accessed on December 5th, 2024)

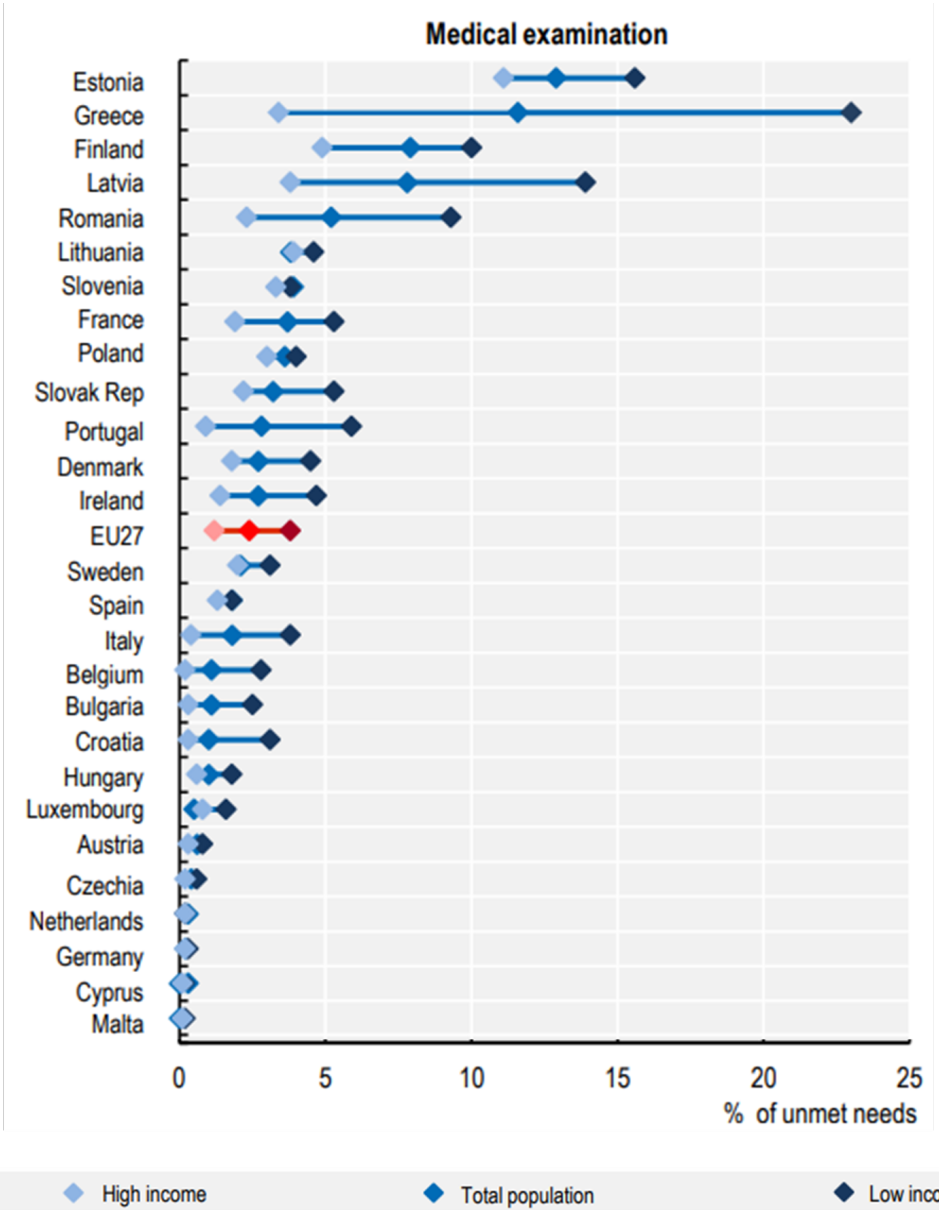


Figure 1.2: Percentage of unmet healthcare needs for medical examinations in EU countries in 2023 by income groups [1]. The chart shows significant disparities, with low-income groups consistently reporting higher unmet needs than high-income groups.

AI offers unprecedented opportunities to tackle the challenges of healthcare inequities. Central to mitigating these disparities is the role of data, as underscored by the Director of NIMHD, who aptly remarked *"Ending health disparities starts with good data."* However, the challenge extends beyond simply collecting data; it involves transforming this data into actionable insights. In this context, AI has become a critical tool, particularly generative AI models. Generative AI addresses one of the most enduring challenges in healthcare: the scarcity and incompleteness of representative and high-quality datasets.

Generative AI is not only solving the problem of data scarcity but also fostering inclusivity in healthcare innovation. By democratizing access to high-quality synthetic datasets, these technologies empower institutions in resource-constrained settings to participate in cutting-edge research and development. For instance, Synthetic Data Generation (SDG) enables hospitals and research centers with limited patient records to contribute to training AI models for disease prediction, treatment optimization, and diagnostic accuracy. This is especially relevant in rare disease research, where small patient populations often exclude these conditions from mainstream medical advances. Similarly, generative AI offers a way to simulate and analyze data without compromising confidentiality in regions with strict privacy regulations, ensuring that innovation is accessible even in the most restrictive environments.

Beyond SDG, generative AI extends its huge potential to numerous other applications in healthcare, as summarized in Table 1.1.

These examples illustrate that generative AI is not merely a tool to address data scarcity but a revolutionary approach to solving diverse healthcare challenges. It democratizes access to advanced AI-driven tools, empowering institutions in resource-constrained environments to participate in cutting-edge medical research and innovation.

In the modern world, where information is power, inequitable access to data has created a new frontier of healthcare injustice. Machine Learning (ML) and AI hold extraordinary potential to revolutionize healthcare, offering tools for operational efficiency, disease prediction and diagnosis, and personalized patient care. Yet these advances often remain the privilege of developed regions with access to vast, high-quality datasets. In contrast, resource-limited settings—and even rare diseases within privileged regions—suffer from data scarcity, effectively excluding them from the benefits of these innovations. This inequity perpetuates a vicious cycle: without data, there can be no progress, and without progress, disparities in health outcomes persist. This reality drives the vision of this doctoral thesis. **Inspired by the need to bridge healthcare disparities, this research explores how advanced AI techniques—specifically generative AI—can reduce inequalities in access to medical knowledge and care.** By leveraging the power of Generative Models (GMs), this thesis aims to contribute to a future where healthcare is equitable, inclusive, and globally accessible. It is a step toward transforming healthcare for the privileged few, communities, and regions worldwide, ensuring that no one is left behind in pursuing health equity.

1.1 Research Objectives

Building on the vision outlined, the overarching goal of this thesis is to design and develop innovative, generative AI-driven methodologies to address the systemic inequities in healthcare. To do so, this research addresses three main objectives, including (1) **addressing a key medical task of significant importance**–Survival Analysis (SA)–to improve medical prognosis, decision-making, and, therefore, personalized patient care; (2) **overcoming the issue of data scarcity**, which is pervasive in healthcare due to strict privacy regulations, the rarity of certain diseases, and the incomplete and heterogeneous nature of medical databases;

[3]Source: Buoy Health Virtual Assistant (Accessed December 6th, 2024)

Application	Description	Example
Modeling Complex Distributions	Leveraging learned underlying data distribution for tasks such as probabilistic prediction, like estimating the likelihood of future events and anomaly detection, identifying patterns instances that deviate from the learned distribution.	Predicting the risk of disease relapse; estimating the time until a critical event (SA); diagnosing atypical conditions.
Missing Data Imputation	Plausibly filling incomplete datasets while maintaining statistical integrity.	Imputing missing variables in biostatistics or EHRs.
Image Enhancement	Improving medical image quality through denoising, resolution enhancement, and reconstruction. Also, leveraging AI to analyze medical images and flag abnormalities. medical images and flag abnormalities.	Denoising noisy MRI or CT scans; generating high-resolution images from degraded inputs; identifying tumors, fractures, or neurological conditions. Google DeepMind demonstrated performance in breast cancer detection exceeding human radiologists [2].
Pathology Analysis	Enhancing efficiency in analyzing pathology slides, often with accuracy comparable to experienced pathologists, and minimizing false negatives.	Computer-aided detection systems highlight areas of concern, assisting clinicians in diagnosing malignancies without replacing their expertise.
Sequential Modeling	Analyzing temporal data to make predictions or simulate patient trajectories.	Forecasting disease progression and modeling clinical pathways.
Optimizing Tasks	Exploring solution spaces for personalized treatments, drug discovery, and trial optimization.	Designing optimal treatment strategies; predicting molecular properties for drug discovery. The AI-driven DSP-1181 drug candidate developed by Exscientia reduced development time to less than 12 months, entering Phase I clinical trials in 2020 [3].
Virtual Assisstants and Telemedicine	Generating conversational responses and synthesizing clinical insights to support patient engagement and care delivery remotely.	Virtual assistants like Buoy Health[3] generating personalized recommendations for care; enabling AI-driven triage during teleconsultations.
Personalized Medicine	Customizing treatments based on genetic, environmental, and lifestyle factors to improve efficacy and minimize side effects.	Targeted cancer therapies tailored to tumor genetic profiles; identifying biomarkers for optimized drug responses in oncology, cardiology, and rare genetic disorders.
Generating Explanatory Content	Producing detailed, human-readable explanations or reports from technical data.	Automated medical reporting; generating conversational responses for virtual assistants.

Table 1.1: Examples of applications of generative AI in healthcare. The state of the art shows diverse applications of generative AI in healthcare, ranging from modeling data distributions and improving medical imaging to personalized treatment optimization and generating explanatory content for patients and clinicians.

and (3) **fostering collaborative efforts between institutions** to reduce inequalities across regions and organizations, breaking down barriers to equitable healthcare research.

A distinctive feature of this thesis is its emphasis on **tabular data**, a dataset foundational to healthcare but often overlooked in generative AI research. Unlike the extensive body of research on generative AI models for image-based datasets such as Computed Tomography (CT), Magnetic Resonance Imaging (MRI), or Positron Emission Tomography (PET) scans, tabular data remains unexplored [4]–[6]. However, it plays a crucial role in healthcare, encompassing patient demographics, clinical observations, laboratory results, and treatment outcomes, forming the backbone of many Electronic Health Records(EHRs) and clinical databases. These datasets are inherently complex, mixing numerical, categorical, and boolean variables with diverse distributions. They often include challenges such as missing values, outliers, and inconsistencies. Unlike data types such as images or text, tabular data lacks clear spatial or sequential patterns, presenting weak, non-linear, or conditional correlations that are harder to model. Nevertheless, ML techniques designed for tabular data demonstrate remarkable versatility. These methods can be applied to various data types, as input data—regardless of its original format—can be mapped into a lower-dimensional manifold abstraction of the input data and subsequently reconstructed back into its original input space. This adaptability underscores the potential of tabular data-focused models to address broader data representation and modeling challenges. Despite the inherent complexities of tabular data, its ubiquitous presence and real-world relevance make it highly valuable for advancing AI applications in healthcare [7]–[9].

Among the various generative AI models available in the state of the art, this research centers on Variational Autoencoders (VAEs) [10] as the primary model of choice due to their ability to learn meaningful latent representations that can be effectively employed in all three thematic areas explained below. The latent space in VAEs enables flexibility, interpretability, and scalability, making them particularly suited for addressing the challenges in healthcare outlined in this research. Specifically, VAEs allow for efficient handling of incomplete and heterogeneous data, robust generation of synthetic samples, and seamless integration in decentralized collaborative frameworks. These benefits position VAEs as an ideal choice for the objectives of this thesis, which aim to address key healthcare inequities by advancing predictive modeling, mitigating data scarcity, and fostering institutional collaboration.

1.1.1 Objectives Across Thematic Areas

Survival Analysis: Addressing a Foundational Medical Task

SA is crucial in medical prognosis, offering insights into time-to-event outcomes such as disease relapse, patient survival probabilities, and treatment efficacy. These insights directly inform treatment planning and follow-up strategies, making SA essential in improving patient outcomes and enabling personalized patient care tailored to the unique needs of individuals. However, traditional SA methods face significant limitations, including the reliance on large datasets, the assumption of proportional hazards, and the difficulty in capturing complex and non-linear associations within survival data. These challenges are particularly pronounced in rare diseases or underrepresented populations, where data is scarce.

This thesis employs VAEs to overcome these challenges, leveraging their latent representation framework to model the complexities of survival data and predict time-to-event outcomes. This work contributes to SA in several key ways:

1. Leveraging VAEs to handle censored data and complex relationships, enabling the development of flexible and accurate predictive models that generalize well across diverse populations.
2. Removing the proportional hazards assumption to offer more adaptable models that better fit a wide range of survival scenarios.
3. Integrating Competing Risks (CR) analysis to model multiple potential failure events, providing a nuanced understanding of risks associated with distinct outcomes, such as disease relapse or treatment complications.
4. Ensuring equity and scalability by designing models robust to data scarcity and accessible to underrepresented groups, broadening the applicability of SA tools.

By combining these contributions, this thesis significantly advances the precision, scalability, and inclusivity of survival modeling. These developments enhance the theoretical foundations and practical applications of SA, particularly in complex and diverse medical contexts, aiming to support personalized patient care and improve individual health outcomes.

Synthetic Data Generation: Mitigating Data Scarcity

Data scarcity poses a significant challenge in healthcare, driven by privacy concerns, fragmented datasets, and limited data availability for rare conditions. SDG is proposed as a solution to this problem, offering the ability to create synthetic datasets replicating the statistical and clinical properties of real-world data. This thesis focuses on Synthetic Tabular Data Generation (STDG), addressing the unique challenges and opportunities of generating high-quality synthetic tabular datasets. From this point forward, SDG will be used interchangeably with STDG, meaning that unless otherwise specified, any reference to SDG refers explicitly to the generation of synthetic tabular data.

VAEs for SDG are selected because they can generate high-quality synthetic samples from learned latent representations, ensuring that the generated data maintain clinical relevance and statistical integrity. By emphasizing STDG, this work aims to fill a critical gap in generative AI research, which has historically concentrated more on image and text than tabular data.

The objectives of STDG are:

1. To design robust frameworks for generating diverse and representative synthetic datasets, enabling researchers and institutions to overcome data limitations.
2. To propose validation methodologies for synthetic data, addressing a critical gap in the field where no universal standard exists. This thesis aims to establish efficient and reliable validation criteria to ensure that synthetic datasets are suitable for healthcare research and development.

3. To enhance the usability of synthetic data for AI model training, particularly in scenarios involving incomplete, scarce, sparse, or heterogeneous real-world datasets.

By addressing data scarcity through STDG and introducing validation standards, this work ensures that synthetic data can be a reliable resource for advancing medical AI, empowering resource-limited institutions to contribute to cutting-edge research.

Federated Learning: Enabling Collaborative Research

Collaboration across healthcare institutions is often limited by fragmented data systems, disparities in technical capabilities, and the challenges of working with heterogeneous datasets. Federated Learning (FL) offers a decentralized approach to training AI models, allowing institutions to collaborate without sharing sensitive patient data.

VAEs play a crucial role in this framework by providing structured latent representations that enable efficient handling of heterogeneous data during model training.

The FL-related objectives of this thesis include:

1. Designing a novel FL framework incorporating VAE-generated synthetic data to address heterogeneity and imbalance challenges.
2. Empowering institutions with limited resources to participate equitably in collaborative research, ensuring that advanced healthcare AI models are not confined to well-funded regions or organizations.
3. Demonstrating the feasibility and effectiveness of this framework in real-world healthcare scenarios, providing a scalable solution for global collaborative efforts.

This research aims to reduce barriers to collaboration and foster a more equitable distribution of healthcare AI advancements by integrating VAEs in FL.

1.1.2 Integrating Objectives Across Thematic Areas

The unifying aspect of this thesis is the synergistic integration of SA, STDG, and FL, with VAEs serving as the connecting thread. The latent representations generated by VAEs enable cohesion across these thematic areas:

- In SA, the latent space provides a compact and interpretable representation of survival data, supporting accurate time-to-event predictions.
- In SDG, these latent representations facilitate the generation of synthetic datasets that are both clinically relevant and statistically robust, ensuring usability across diverse healthcare applications.
- In FL, the latent space allows for efficient and decentralized training of AI models, overcoming challenges posed by heterogeneous and non-Independent Identically Distributed (non-IID) data.

The thesis ensures a coherent framework for addressing healthcare inequities through AI-driven solutions by aligning these objectives. The integration of VAEs across these areas highlights

the versatility and effectiveness of their latent representation capabilities in tackling complex healthcare challenges.

1.2 Thesis Overview

The thesis comprises seven chapters, each contributing to a cohesive narrative of motivation, research development, and evaluation. The chapters are thematically structured to facilitate readability, allowing each chapter to be understood independently while maintaining logical interconnections. This thematic organization aligns with the objectives outlined in Section 1.1.

The current chapter introduces the thesis, presenting the context of the research context, motivation, and objectives. It highlights the challenges in addressing healthcare inequities and emphasizes the evolving potential of AI, particularly generative AI, as a tool for achieving equitable healthcare. In addition, it provides an overview of the thesis structure.

The subsequent chapters are organized as follows:

Chapter 2 provides a comprehensive review of prior research and foundational concepts related to the thesis. It is divided into four sections. The first section examines the current applications, limitations, and ethical considerations of AI in healthcare, offering a contextual overview. The remaining sections delve into the primary thematic areas of this research: (1) SA, introducing key concepts and methodologies, focusing on gaps addressed by GMs; (2) SDG, exploring state-of-the-art data generation methods, particularly in medical contexts; and (3) FL, examining advances and challenges in privacy-preserving AI for collaborative healthcare research.

Chapter 3 details the development and evaluation of the proposed generative AI models for SA. It includes descriptions of the methodologies, experimental setup, and results, focusing on how these models improve predictive accuracy and address gaps identified in state-of-the-art approaches.

Chapter 4 focuses on the design, implementation, and validation of a novel GM for SDG, specifically on STDG. The proposed model is rigorously compared against existing state-of-the-art techniques, highlighting its strengths and addressing current gaps in the field. Additionally, this chapter introduces a standardized framework for evaluating the quality of synthetic tabular data, filling a critical gap in the literature where no universal standard currently exists. Finally, the chapter proposes a novel methodology for STDG under data scarcity, offering a practical solution for scenarios where data are limited or of low quality, thereby ensuring broader applicability in healthcare and beyond.

Chapter 5 explores the development of FL strategies for training AI models on decentralized, privacy-sensitive datasets. It includes case studies demonstrating the practical applications of FL in healthcare, particularly in the domains of SDG and SA. The chapter highlights how FL enables collaboration between institutions, particularly benefiting those with limited resources.

Chapter 6 contextualizes the contributions of the thesis within the broader field of AI in healthcare. It discusses the implications of the results, identifies limitations, and proposes

avenues for future research. It also suggests potential applications and improvements to the methods developed in the thesis.

Chapter 7 summarizes the contributions and outcomes of the thesis, emphasizing its impact on healthcare equity and innovation. It reiterates the importance of generative AI in addressing healthcare disparities and presents a vision for a more inclusive and equitable healthcare system.

Figure 1.3 illustrates the structure and dependencies between the chapters.

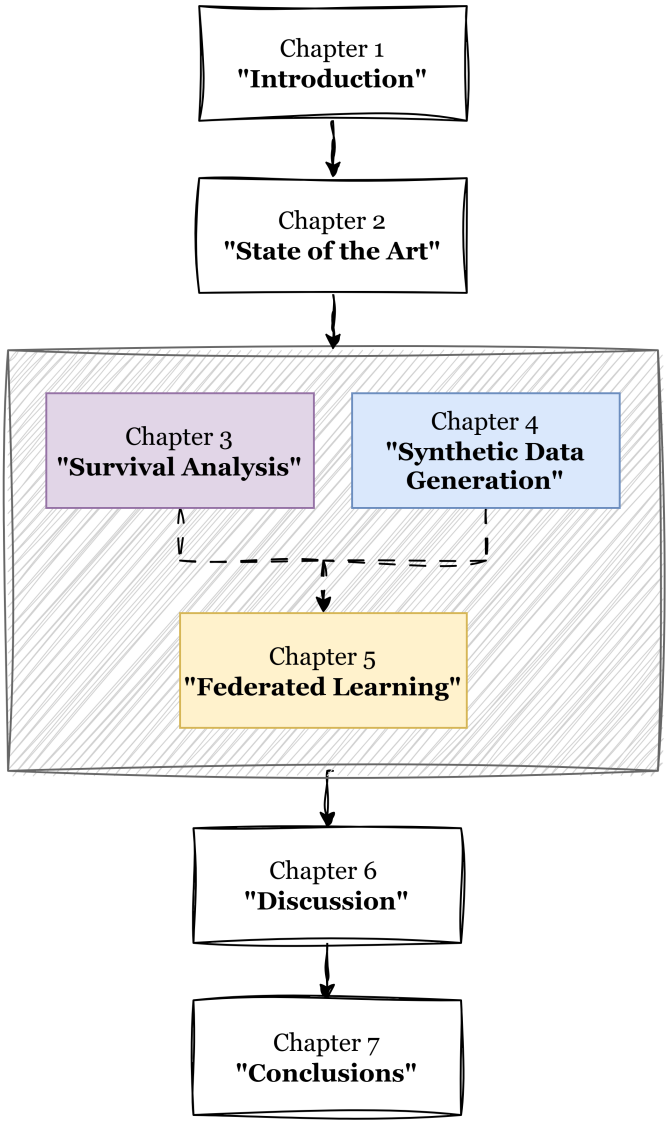


Figure 1.3: Block diagram depicting the structure and dependencies between chapters in the thesis. Solid arrows indicate chapters that are prerequisites for subsequent sections, while dotted arrows indicate recommended chapters that are not directly required. White blocks represent the introduction, related works, and conclusions, while colored blocks highlight the original contributions of the thesis.

1.3 Research Contributions

This section outlines the research contributions resulting from the studies conducted during the development of this thesis. These contributions are primarily reflected in the following publications, accompanied by journal metrics for those indexed in the Journal Citation Reports (JCR):

Journal Publications:

- **A. Apellániz P.,** Jiménez A., Arroyo Galende B., Parras J., and Zazo S., *'Synthetic Tabular Data Validation: A Divergence-Based Approach,'* in IEEE Access, vol. 12, pp. 103895-103907, 2024, doi: 10.1109/ACCESS.2024.3434582.
  Journal metrics (2023): IF 25.669, Rank Q2 (65.3 in Computer Science, Information Systems; 65.5 in Engineering, Electrical & Electronic; 60.9 in Telecommunications).
- **A. Apellániz P.,** Parras J., and Zazo S., *'Leveraging the variational Bayes autoencoder for survival analysis,'* in Scientific Reports, vol. 14, 24567, 2024, doi: 10.1038/s41598-024-76047-z.
  Journal metrics (2023): IF 3.8, Rank Q1 (81.7 in Multidisciplinary Sciences).
- **A. Apellániz P.,** Parras J., and Zazo S., *'Improving Synthetic Data Generation through Federated Learning in Scarce and Heterogeneous Data Scenarios,'* in Big Data and Cognitive Computing, vol. 9(2), 18, 2025, doi: 10.3390/bdcc9020018.
  Journal metrics (2023): IF 3.7, Rank Q2 (66.8 in Computer Science, Artificial Intelligence; 70.2 in Computer Science, Information Systems; 83.0 Computer Science, Theory & Methods).

Peer-Reviewed International Conference Publications:

- **A. Apellániz P.,** Parras J., and Zazo S., *'CR-SAVAE: A Parametric Method for Survival Analysis with Competing Risks,'* in the 32nd European Signal Processing Conference (EUSIPCO), Lyon, France, 2024, pp. 1526-1530, doi: 10.23919/EUSIPCO63174.2024.10715431.
- **A. Apellániz P.,** Parras J., and Zazo S., *'An Improved Tabular Data Generator with VAE-GMM Integration,'* in the 32nd European Signal Processing Conference (EUSIPCO), Lyon, France, 2024, pp. 1886-1890, doi: 10.23919/EUSIPCO63174.2024.10715230.

Submitted Journal Publications (Under Review):

- **A. Apellániz P.,** Jiménez A., Arroyo Galende B., Parras J., and Zazo S., *'Artificial Inductive Bias for Synthetic Tabular Data Generation in Data-Scarce Scenarios,'* under review in Pattern Recognition journal. Preprint available: 2407.03080.
  Journal metrics (2023): IF 7.5, Rank Q1 (88.1 in Computer Science, Artificial Intelligence, 93.1 in Engineering, Electrical & Electronic).

- **A. Apellániz P.,** Arroyo Galende B., Jiménez A., Parras J., and Zazo S., *'Advancing Cancer Research with Synthetic Data Generation in Low-Data Scenarios,'* under review in the IEEE Journal of Biomedical and Health Informatics.
  Journal metrics (2023): IF 6.7, Rank Q1 (89.8 in Computer Science, Information Systems, 87.9 in Computer Science, Interdisciplinary Applications, 93.2 in Mathematical & Computational Biology, 94.3 in Medical Informatics).
- **A. Apellániz P.,** Parras J., and Zazo S., *'Enhancing Survival Analysis Through Federated Learning in Non-IID and Scarce Data Scenarios,'* under review in Computers in Biology and Medicine.
  Journal metrics (2023): IF 7.0, Rank Q1 (94.0 in Biology, 89.1 in Computer Science, Interdisciplinary Applications, 87.4 in Engineering, Biomedical, 97.7 in Mathematical & Computational Biology).

Each of these contributions is discussed in detail in the corresponding chapters of the thesis. Figure 1.4 illustrates the alignment of each publication with the thematic areas and intersections of the research focus.

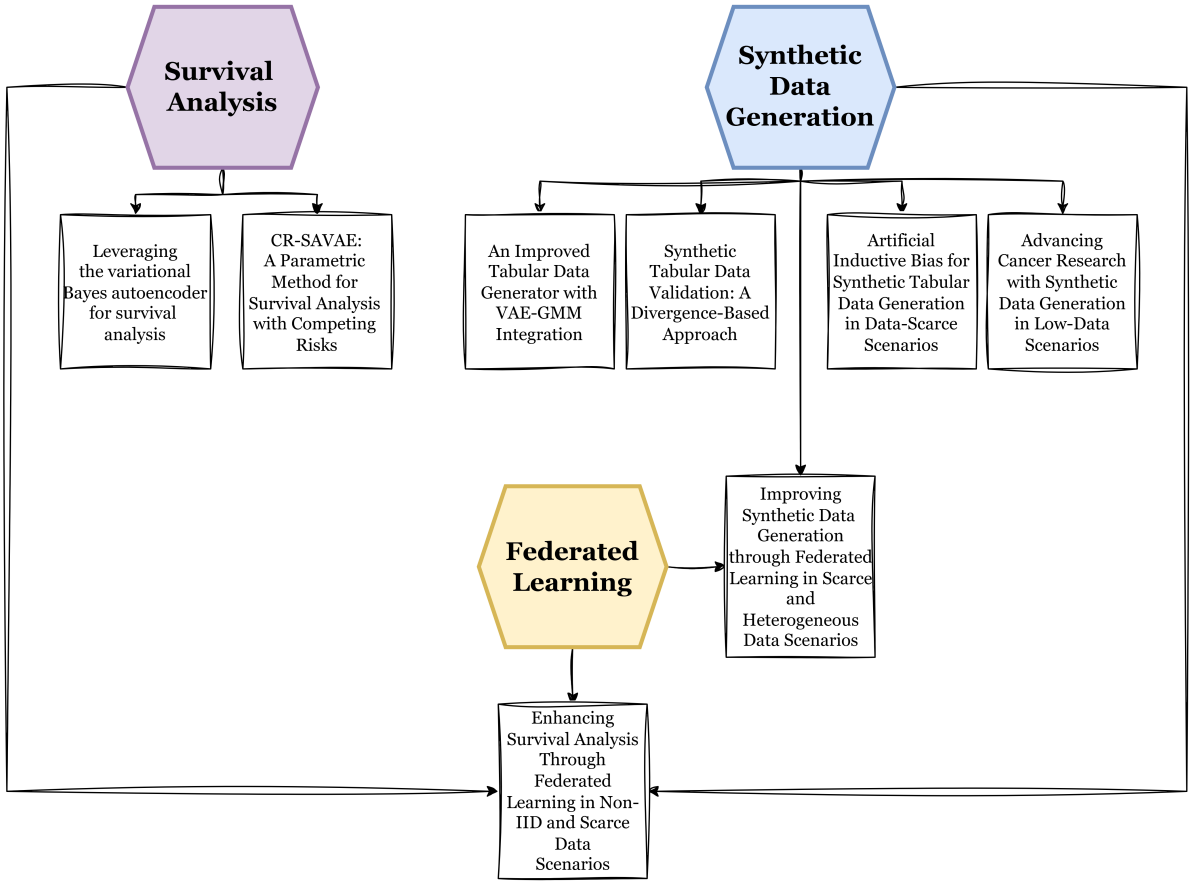


Figure 1.4: Mapping of publications to the thesis thematic areas.

Several additional publications, which indirectly stem from the work in this thesis, are also relevant to the thematic areas under research:

Journal Publications:

- D'Amico, S., Dall'Olio, L., Rollo, C., **Alonso, P.**, Prada-Luengo, I., Dall'Olio, D., et al., *'MOSAIC: An Artificial Intelligence–Based Framework for Multimodal Analysis, Classification, and Personalized Prognostic Assessment in Rare Cancers,'* in JCO Clinical Cancer Informatics, 8, e2400008, 2024, doi: 10.1200/CCI.24.0000
  Journal metrics (2023): IF 3.3, Rank Q2 (63.5 in Oncology).
- Lahoz Navarro, M., Jehle, J.S., **A. Apellániz, P.,** Parras, J., Zazo, S., Gerdts, M. *'Deep Learning as a New Framework for Passive Vehicle Safety Design Using Finite Elements Models Data,'* in Applied Sciences, vol. 14, 9296, 2024, doi: 10.3390/app14209296.
  Journal metrics (2023): IF 2.5, Rank Q2 (50.7 in Chemistry, Multidisciplinary, 75.7 in Engineering, Multidisciplinary, 41,4 in Materials Science, Multidisciplinary, and 51.7 in Physics, Applied).

Peer-Reviewed International Conference Publications:

- D'Amico, S., Dall'Olio, L., Rollo, C., **Alonso, P.**, Prada-Luengo, I., Dall'Olio, D., et al., *'Multi-modal analysis and federated learning approach for classification and personalized prognostic assessment in myeloid neoplasms,'* in Blood, 140 (Supplement 1), pp. 9828-9830, 2022, doi: 10.1182/blood-2022-166802.
  Journal metrics (2023): IF 21.1, Rank Q1 (98.5 in Hematology).
- Collado Gimbert A., Reidel S., **A. Apellániz P.**, Álvarez F., Arroyo Galende B., Beneitez D., et al., *'Data Driven Research through the European RADeep Registry and the Use of Artificial Intelligence Towards Personalized Medicine in Sickle Cell Disease,'* Poster Abstract published in Blood Journal, 144 (Supplement 1), p. 1138, 2024, doi: 10.1182/blood-2024-203331.
  Journal metrics (2023): IF 21.1, Rank Q1 (98.5 in Hematology).
- Asti G., D'Amico S., Carota L., Piscia D., Casadei F., Saha Cyrille Merleau N., **Alonso De Apellaniz P.**, et al., *'An Artificial Intelligence-Based Federated Learning Platform to Boost Precision Medicine in Rare Hematological Diseases: An Initiative By GenoMed4all and Synthema Consortia,'* Poster Abstract published in Blood Journal, 144 (Supplement 1), p. 4989, 2024, doi: 10.1182/blood-2024-205541.
  Journal metrics (2023): IF 21.1, Rank Q1 (98.5 in Hematology).
- Casadei, F., Carota L., Asti G., D'Amico S., Piscia D., Zazo S., **A. Apellániz P.**, Parras J., Sala C., Rollo C., S. C. Merleau N., Fariselli P., Della Porta M., Sanavia T., A. Garcia F., Castellani G., and Giampieri E., *'Survival Model Optimization via Federated Learning: A Study Combining Simulations and Experiments,'* in 2024 IEEE International Conference on Big Data (BigData), pp. 7658-7667, 2024, doi: 10.1109/BigData62323.2024.10825368.

Submitted Journal Publications (Under Review):

- Arroyo Galende B., **A. Apellániz, P.,** Parras, J., Zazo, S., Uribe S., *'Membership Inference Attacks, Differential Privacy, and their relation to Generative Models,'* 2024, under review in the IEEE Journal of the Computer Society.
  Journal metrics (2023): IF 5.7, Rank Q1 (89.0 in Computer Science, Hardware & Architecture; 86.6 in Computer Science, Information Systems; 83.8 Computer Science, Interdisciplinary Applications; 89.9 in Computer Science, Theory & Methods; 85.7 in Engineering, Electrical & Electronic).

Submitted Peer-Reviewed International Conference Publications (Under Review):

- Carota L., Casadeir F., Asti G., Piscia D., Biondi R., D'Amico S., Zazo S., **A. Apellániz, P.,** et al., *'Federated Learning to predict Silent Cerebral Infarction with Sickle Cell Disease for Small Datasets: Simulations and Experiments,'* 2024, under review in International Conference on Software Testing, Verification and Validation (ICST) 2025.

These contributions further demonstrate the broader impact of the methodologies and concepts developed during this research.

In addition to the publications derived directly and indirectly from this thesis, the research has also contributed academically through active participation in two research projects. These projects have provided valuable opportunities to collaborate with leading experts, address interdisciplinary challenges, and apply the methodologies developed in this thesis to real-world healthcare scenarios. The involvement in these initiatives has further enriched the academic and practical impact of the research, fostering innovation and alignment with European Union priorities in advancing healthcare equity and technological development. These projects are:

- **GenoMed4All.** This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 101017549.
- **SYNTHEMA.** This project is an initiative funded by the European Union's Horizon Europe Research and Innovation program under grant agreement N° 101095530.

Chapter 2

State of the Art

This chapter provides a comprehensive overview of the foundational areas underpinning the research conducted in this thesis, focusing on their historical evolution, benefits, and challenges. The objective is to establish the necessary context for understanding the methodologies and contributions presented in subsequent chapters.

Section 2.1 begins with a concise review of the role of AI in healthcare, tracing its historical progression from rule-based systems to modern ML and GMs. This section highlights the potential of AI in improving diagnostic accuracy, treatment planning, and healthcare efficiency while addressing ethical and technical challenges such as data privacy, algorithmic bias, and regulatory constraints.

The focus then shifts to SA in Section 2.2, a critical tool in medical prognosis that models time-to-event outcomes. Traditional approaches are discussed alongside recent advancements in ML-based methods. This section explores the strengths and limitations of these techniques, particularly in capturing complex, non-linear relationships and addressing CR.

The next section (Section 2.3) explores SDG, an area that has gained traction as a solution to data scarcity in healthcare. The discussion centers on techniques for generating synthetic datasets, particularly for tabular data, emphasizing their potential to balance statistical preservation with data utility. Challenges related to validation, similarity, and the generalizability of synthetic data are also analyzed.

Finally, FL is examined in Section 2.4 as a decentralized framework for training AI models. Its application in healthcare is explored, focusing on enabling collaborative research across institutions while safeguarding sensitive patient data. Technical challenges are outlined, such as data heterogeneity, communication costs, and security concerns.

This chapter aims to provide a comprehensive and systematic understanding of these thematic areas, forming the basis for the innovative methodologies and applications developed in the thesis.

2.1 Artificial Intelligence in Healthcare

AI is revolutionizing the healthcare industry, potentially transforming how clinicians diagnose, treat, and prevent diseases. By harnessing the power of ML, Computer Vision, and other advanced technologies, AI allows healthcare providers to make more informed decisions, improve patient outcomes through personalized treatment plans, and streamline bureaucratic processes.

2.1.1 Historical Evolution

Technological advancements, the increasing demand for data-driven solutions in clinical practice, and the evolving understanding of medical data have shaped the evolution of AI in healthcare. From the early development of rule-based systems to the current era of Deep Learning (DL) and integrated AI ecosystems, the trajectory of AI in healthcare reflects a continuous effort to improve diagnostic accuracy, treatment efficacy, and operational efficiency. This section provides a chronological overview of key milestones and challenges that have defined the adoption and impact of AI in the healthcare domain.

Figure 2.1 illustrates examples of key milestones in the evolution of AI in healthcare. These milestones, spanning foundational definitions, technological breakthroughs, and regulatory advancements, will be explored more deeply in the following subsections.

Foundations: Rule-Based Systems and Expert Systems (1950s-1970s)

The integration of AI into medicine is a relatively recent phenomenon, yet its roots trace back to the mid-20th century. While the term 'artificial intelligence' was first coined in the 1950s, practical applications in healthcare did not emerge until the early 1970s. During this formative period, AI research focused primarily on symbolic reasoning and logic-based systems, laying the groundwork for future innovations.

One of the earliest and most notable AI medical applications was MYCIN [13], developed at Stanford University during the early 1970s. MYCIN was an expert system designed to assist physicians in diagnosing bacterial infections and recommending appropriate antibiotic treatments. Using a knowledge base of rules and an inference engine, MYCIN could evaluate clinical input and provide treatment recommendations. Despite being a landmark achievement, MYCIN and other rule-based systems of the era faced significant limitations, including rigidity, inability to adapt to new data, and reliance on manually encoded knowledge.

In parallel with developments in expert systems, AI research in the 1960s witnessed advances in robotics and Natural Language Processing (NLP), which indirectly influenced its applications in medicine. Shakey the robot [14], developed at Stanford Research Institute in 1966, demonstrated the ability to process instructions and perform tasks autonomously, representing a major milestone in AI. Similarly, Eliza [15], an NLP program introduced in 1964 by Joseph Weizenbaum, showcased the potential for AI to interact with humans using pattern-matching and substitution methodologies. While not directly related to medicine, these innovations highlighted the potential for AI to simulate human reasoning and interaction, capabilities that would later be adapted for medical applications.

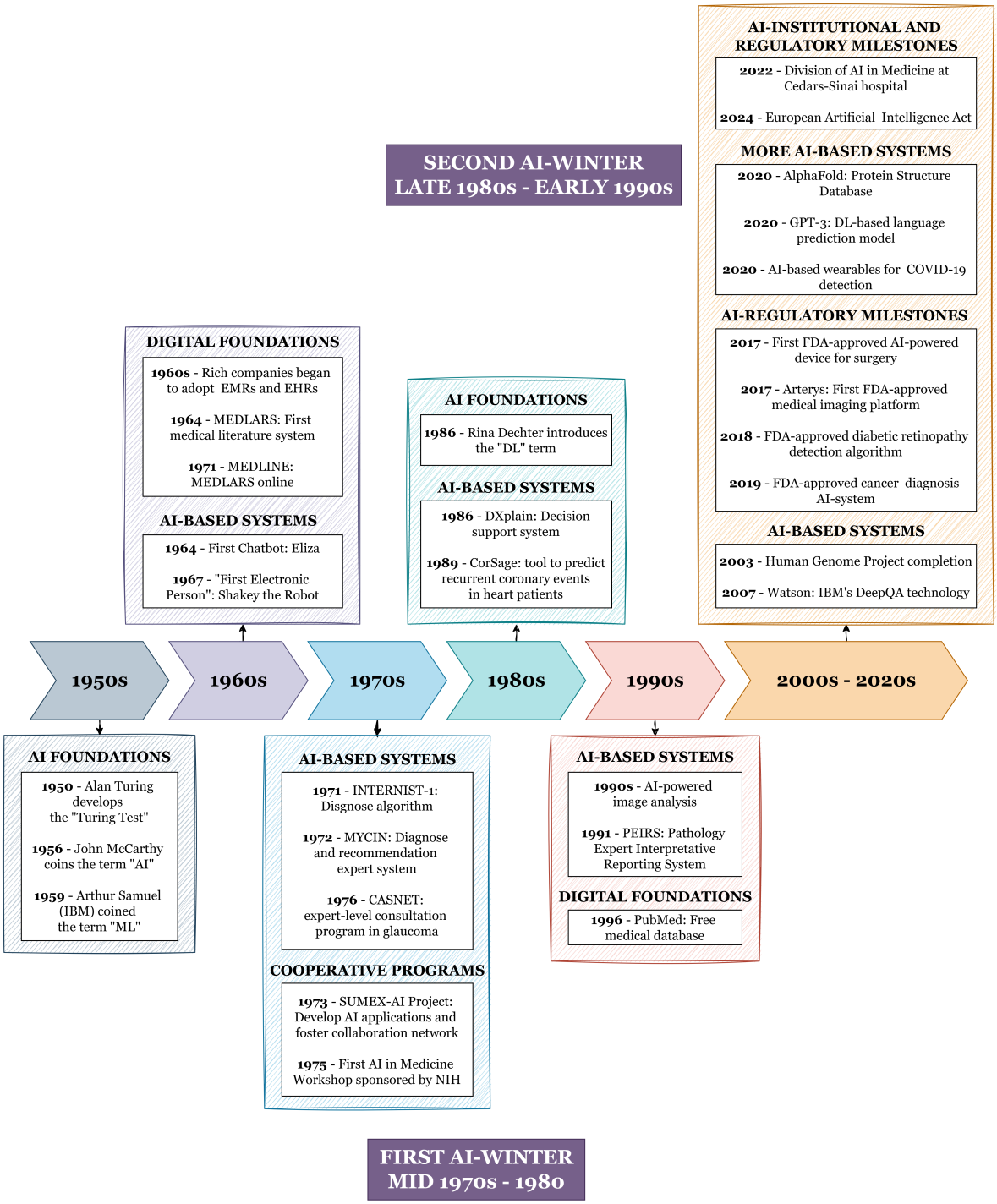


Figure 2.1: Timeline of key milestones in the evolution of AI in healthcare. It traces major developments from foundational concepts like the Turing Test [11] and early AI definitions [12] to applications in the 2000s and beyond. It also highlights AI Winters, periods of stagnation due to technical and funding challenges, and the subsequent resurgence driven by breakthroughs in regulation, institutional support, and innovation.

A critical advancement during this period was the digitization of medical information. The development of the Medical Literature Analysis and Retrieval System (MEDLARS) [16] and the subsequent web-based search engine PubMed[1] by the National Library of Medicine in the 1960s marked a turning point in biomedical research. These systems enabled efficient medical literature retrieval and provided the foundation for integrating digital resources into medical practice. In addition, early clinical informatics databases and Electronic Medical Records (EMRs) began to emerge, establishing the infrastructure for future AI-driven innovations in healthcare. EMRs provide a systematized, digital collection of patient health information, including medical histories, diagnoses, treatment plans, and clinician notes. By consolidating substantial amounts of healthcare data accumulated throughout the lifetime of a patient, EMRs offer new opportunities to improve the understanding of medical conditions, diagnoses, and treatments. While transitioning from traditional, paper-based systems to digitalized EMRs presented challenges, the long-term potential of such systems to enhance patient outcomes and streamline care pathways became evident. These digital foundations played an instrumental role in enabling the later growth of AI in medicine.

The 1970s saw interdisciplinary collaboration and resource sharing that advanced AI in medicine. The Stanford University Medical Experimental Computer for Artificial Intelligence in Medicine (SUMEX-AIM) [17] facilitated the development and distribution of AI systems, leading to innovations like CASNET [18], causal-associational network for managing glaucoma. CASNET demonstrated the feasibility of applying AI to provide disease-specific advice to physicians. During this time, the INTERNIST-1 system [19] was also developed, representing one of the first artificial medical consultants. Using search algorithms, INTERNIST-1 could generate differential diagnoses based on patient symptoms, paving the way for AI-assisted diagnostic tools. These early systems underscored the potential of AI to assist clinicians in decision-making, though widespread adoption remained constrained by technical limitations and skepticism within the medical community.

Trasition to Data-Driven Methods: Machine Learning and Decision Support (1980s-1990s)

The 1980s marked a shift in AI from rule-based systems to ML, which leveraged data-driven models capable of identifying patterns without explicit programming. Unlike earlier systems, ML algorithms adapted new information, offering greater flexibility and scalability. This transition was supported by the rapid accumulation of healthcare data, including medical images, genomic information, and EHRs, as well as advancements in computational capabilities.

One significant development of this era was the DXplain [20] system, introduced in 1986 by the University of Massachusetts. Building on earlier systems like INTERNIST-1, DXplain enabled clinicians to input patient symptoms and receive potential diagnoses from an expanded database of over 500 diseases at its inception. DXplain also served as an early electronic medical textbook, providing clinicians with detailed descriptions of diseases and references. This dual functionality highlighted the capacity of AI not only for decision support but also for augmenting clinical education.

[1]Source: PubMed (Accessed on December 6th, 2024)

Other innovations during this period included CorSage [21], a clinical tool introduced by Cedars-Sinai cardiologists in 1989, which combined AI and statistical techniques to predict recurrent coronary events in heart patients. Similarly, the Pathology Expert Interpretative Reporting System (PEIRS) [22] (1991) achieved nearly 95% diagnostic accuracy, demonstrating the growing precision of AI in specialized medical domains.

The 1990s also saw advancements in AI-powered image analysis, particularly in radiology and pathology. ML algorithms were applied to improve the accuracy and efficiency of interpreting medical images, a precursor to the computer-aided diagnostic tools widely used today.

Despite these achievements, AI in medicine faced significant setbacks during the 'AI winters' of the late 1970s and 1980s. These periods of reduced funding and interest stemmed from challenges such as unmet expectations, the high costs of developing and maintaining expert systems, and limited computational resources. However, interdisciplinary collaboration among pioneers persisted, sustaining progress through initiatives like the AI in Medicine workshop from the National Institutes of Health (NIH) at Rutgers University in 1975 [23]. These efforts fostered a resilient research community that laid the groundwork for later breakthroughs.

The Modern Era: Deep Learning and Advanced AI Techniques (2000s-Present)

By the early 2000s, AI in healthcare entered a groundbreaking phase, driven by advancements in data generation, computational technologies, and algorithmic innovation. Building on the foundational work of earlier decades, this era witnessed the transition of AI from theoretical research to practical, real-world applications in clinical workflows. From diagnostic tools to drug discovery, AI has become a central force reshaping healthcare.

→ *Key Drivers of Innovation*

Three main factors catalyzed the modern era of AI in healthcare. First, the widespread adoption of EHRs facilitated the systematic capture of patient data, encompassing clinical histories, imaging studies, and genomic information. This wealth of structured and unstructured data provided an unprecedented resource for training AI models. Second, computational advancements, including developing Graphic Processing Units (GPUs) and distributed computing platforms, allowed researchers to train complex DL algorithms on large datasets. Finally, breakthroughs in DL architectures, such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), enabled AI to tackle high-dimensional and sequential data accurately and efficiently, overcoming challenges like overfitting and poor generalization.

→ *Transformative Applications*

One of the most impactful applications of AI in the modern era has been in medical imaging. DL techniques, particularly CNNs, have revolutionized radiology and pathology by enabling healthcare professionals to detect subtle abnormalities in medical images precisely. For instance, in 2017, researchers at Stanford University developed a CNN capable of classifying skin lesions with diagnostic accuracy comparable to that of experienced dermatologists [24]. Similarly, Arterys[2], a medical imaging platform approved by the Food and Drug Administration (FDA), significantly reduced the time required to analyze cardiac MRI scans, exemplifying

[2]Source: Arterys Medical Imaging Platform (Accessed on December 6th, 2024)

the role of AI in accelerating clinical workflows.

NLP has also emerged as a critical domain. Tools like IBM Watson [25] have demonstrated the ability to extract meaningful insights from unstructured clinical text, enabling predictive analytics, automated documentation, and translational research. The DeepQA technology in Watson, introduced in 2007, highlighted the potential of NLP in identifying RNA-binding proteins associated with diseases like amyotrophic lateral sclerosis, showcasing the capacity of AI to bridge basic and clinical science.

Drug discovery represents another evolved area. By leveraging large molecular datasets and predictive modeling, AI platforms have expedited the identification of novel compounds and streamlined clinical trial designs. A milestone in this field was achieved in 2020 when Google DeepMind predicted protein structures from amino acid sequences with AlphaFold, addressing a decades-long challenge in biology and advancing targeted therapies [26].

AI-powered predictive analytics has enhanced clinical decision-making, particularly in identifying at-risk patients. Models capable of predicting sepsis onset [27], readmission risks [28], and cardiac arrest likelihood [29] have been integrated into care workflows, improving outcomes through timely interventions. In 2022, establishing the Division of Artificial Intelligence in Medicine at Cedars-Sinai hospital[3] underscored the importance of these tools for population-level analytics.

→ Seminal Milestones and Regulatory Advancements

The modern era has been marked by several milestones that signify the growing acceptance of AI in healthcare. In 2003, the completion of the Human Genome Project[4] provided critical data linking genetic variations to disease, catalyzing advancements in precision medicine. In 2017, the FDA approved the first AI-powered device for operating-room use[5], in 2018 a DL algorithm for detecting diabetic retinopathy with exceptional accuracy[6] and in 2019 an AI system for cancer diagnosis[7]. The last update on the trajectory of regulatory approvals in 2024 includes the authorization of 950 AI-powered devices[8].

→ Ethical Challenges and Future Directions

Despite its successes, the integration of AI into healthcare has raised ethical concerns that demand ongoing attention. Ensuring patient privacy, addressing biases in AI algorithms, and achieving transparency in decision-making processes are essential to maintaining trust in these technologies. Regulatory frameworks must evolve to balance innovation with safety, fostering the responsible use of AI in clinical practice.

The modern era of AI in medicine demonstrates the potential of the field to transform healthcare, offering new solutions to longstanding challenges. As technological advancements accelerate, the possibilities for AI to further enhance diagnostics, treatment planning, and

[3]Source: AI in Mecidine Division at Hospital in Los Angeles (Accessed on December 6th, 2024)
[4]Source: Completion of the Human Genome Project (Accessed on December 6th, 2024)
[5]Source: FDA approves First AI-powered Device for Surgery (Accessed on December 6th, 2024)
[6]Source: FDA approves DL for Detecting Diabetic Retinopathy (Accessed on December 6th, 2024)
[7]Source: Paige AI for Cancer Diagnosis (Accessed on December 6th, 2024)
[8]Source: FDA AI-Enabled Medical Devices (Accessed on December 6th, 2024)

operational efficiency remain vast and evolving.

2.1.2 Challenges in Implementing

Integrating AI into healthcare offers huge potential. However, its implementation is fraught with challenges arising from the complexity of medical data, stringent regulatory landscapes, interpretability concerns, and ethical dilemmas. Understanding these challenges is critical to unlocking the full potential of AI while ensuring its safe and equitable deployment in clinical practice.

Table 2.1 summarizes all challenges and solutions across the following subsections.

Challenge	Description	Possible Solution
Data Privacy and Security	1. Vulnerabilities to breaches and unauthorized access. 2. Ambiguity in applying existing regulations.	1. Implement advanced encryption, robust access controls, and real-time monitoring. 2. Develop AI-specific governance frameworks that balance privacy and innovation.
Insufficient and Biased Data	1. Limited, fragmented, or non-representative datasets. 2. Historical biases embedded in data.	1. Use fairness-aware algorithms; regularly audit model performance. 2. Foster collaboration between institutions; use SDG techniques to augment data.
Interpretability and Trustworthiness	Lack of transparency in 'black-box' models.	Introduce XAI frameworks to enhance trust and understanding.
Regulatory and Legal Complexities	Lack of AI-specific regulatory standards.	Harmonize global standards for validation and approval frameworks.
Ethical and Bias Concerns	Discriminatory outcomes from biased data.	Ensure inclusive datasets and fairness corrections.
Resistance to Adoption	Clinician skepticism about AI reliability.	Educate professionals on AI benefits and limitations; engage patients with transparent communication.
Financial Constraints	High initial costs of AI implementation.	Establish public-private partnerships and scalable AI solutions.

Table 2.1: Summary of challenges and proposed solutions for implementing AI in healthcare, categorized by key thematic areas.

Data Privacy and Security

Healthcare generates vast amounts of sensitive data, including personal health information, genomic sequences, and medical imaging. The reliance of AI on this data amplifies concerns about privacy and security. AI systems need data sharing across stakeholders, increasing vulnerabilities to breaches and unauthorized access. Cyberattacks targeting healthcare systems

can compromise patient confidentiality and safety, leading to identity theft, financial fraud, and disruptions in care delivery.

Regulatory frameworks such as the Health Insurance Portability and Accountability Act (HIPAA) in the United States of America and the General Data Protection Regulation (GDPR) in the European Union provide important guidelines for data protection. However, these regulations are not always specific to AI technologies, creating ambiguity in their application. Additionally, the competitive nature of healthcare markets and the adoption of disparate EHR systems exacerbate the challenge of secure and seamless data sharing.

To address these issues, healthcare organizations must invest in advanced encryption techniques, real-time monitoring, and robust access controls. Furthermore, policymakers must establish AI-specific data governance frameworks that balance privacy, innovation, and clinical utility.

Insufficient and Biased Data

The effectiveness of AI in healthcare depends on the availability of high-quality, diverse, and representative data. Many healthcare datasets are limited in scale, fragmented, or lack demographic diversity. This scarcity of data hinders the ability of AI to generalize across different populations, leading to models that perform well in controlled environments but fail in real-world scenarios.

Bias in data poses another significant challenge. Historical inequities embedded in healthcare systems often persist in training datasets, leading AI systems to perpetuate or exacerbate these disparities. For example, diagnostic models trained on predominantly light-skinned populations have demonstrated lower accuracy when applied to darker-skinned patients [30], [31]. Addressing these biases requires using inclusive datasets, fairness-aware algorithms, and regular audits to monitor model performance across different demographic groups.

Collaboration between institutions to responsibly share data, combined with SDG techniques, offers a potential solution to data scarcity. Additionally, fostering transparency in how datasets are curated and algorithms are trained is essential to ensure equitable outcomes.

Interpretability and Trustworthiness

One of the most significant barriers to adopting AI in healthcare is the lack of interpretability of complex models, particularly those based on DL. These models often function as 'black boxes,' providing outputs without explaining how decisions are made. In clinical settings, decisions usually involve life-or-death scenarios; this opacity undermines trust among healthcare professionals and patients.

Clinicians must be able to understand and validate AI-driven recommendations to integrate these tools into their workflows effectively. Explainable AI (XAI) frameworks are, therefore, crucial. They provide transparent reasoning for predictions and enable healthcare providers to trace the logic behind AI decisions. Developing AI systems that complement human expertise, rather than replacing it, ensures that clinicians retain accountability while benefiting from data-driven insights.

Regulatory and Legal Complexities

Healthcare is a highly regulated industry, and introducing AI technologies adds further complexity to an already stringent regulatory environment. Existing frameworks, such as HIPAA and GDPR, often fail to address the unique challenges posed by AI, such as algorithmic transparency and continuous learning systems. The lack of clear standards for AI-specific validation and approval delays the deployment of innovative solutions.

Globally, regulatory agencies are beginning to adapt. Initiatives such as the FDA framework for Software as a Medical Device (SaMD) and the guidelines from the European Medicines Agency for AI-based tools represent important steps. However, these frameworks are still evolving and require harmonization to support the international deployment of AI technologies. Developers must navigate diverse regulatory landscapes, ensuring compliance while maintaining the flexibility needed to innovate.

Ethical and Bias Concerns

The deployment of AI in healthcare raises profound ethical questions, particularly regarding fairness, autonomy, and accountability. AI systems trained on biased data can produce discriminatory outcomes, disproportionately impacting marginalized populations. For example, resource allocation algorithms can prioritize wealthier patients, reinforcing systemic inequities. Ethical concerns also extend to the use of AI in decision-making processes, such as diagnosis and treatment planning, where errors or biased recommendations can have severe consequences.

Ensuring the ethical deployment of AI requires transparency in how algorithms are developed, validated, and applied. Regular audits, fairness correction techniques, and stakeholder engagement are critical. Moreover, healthcare professionals must be equipped to understand the limitations of AI systems and maintain ultimate responsibility for patient care decisions.

Resistance to Adoption

Resistance from healthcare professionals and patients is a significant barrier to the widespread adoption of AI. Clinicians may be skeptical of the reliability of AI, concerned about disruptions to established workflows, or apprehensive about potential impacts on job security. Patients, meanwhile, may lack trust in AI-driven care, particularly in high-stakes scenarios.

Overcoming this resistance requires comprehensive education and training programs for healthcare professionals, emphasizing the benefits of AI and addressing concerns about its limitations. Engaging patients through transparent communication about the role of AI in their care can also build trust. Demonstrating measurable improvements in care quality through AI adoption is essential to foster stakeholder acceptance.

Financial Constraints

Integrating AI into healthcare systems is resource-intensive, requiring significant investments in infrastructure, training, and ongoing maintenance. High initial costs can discourage adoption, particularly for smaller institutions or those operating in resource-limited settings. In addition, data collection, cleaning, and storage expenses add to the financial burden.

Public-private partnerships, government subsidies, and scalable AI solutions can help address these financial challenges. Long-term cost savings from improved efficiency and better patient outcomes can further justify investments in AI technologies.

Conclusion

Implementing AI in healthcare is a complex endeavor that requires addressing significant technical, regulatory, ethical, and financial challenges. The healthcare industry can unlock the revolutionary potential of AI by prioritizing data security, fostering transparency, addressing biases, and building stakeholder trust. Collaboration between policymakers, healthcare providers, and technology developers is essential to navigate these challenges, paving the way for a more efficient, equitable, and patient-centered future.

2.1.3 Trends and Emerging Directions

As AI continues to evolve, novel paradigms and technologies are reshaping its role in healthcare. These emerging trends—from advanced GMs to decentralized data-sharing frameworks and wearable device integration—demonstrate the potential of AI to revolutionize diagnostics, personalized care, and health monitoring. This section explores key developments signaling the future trajectory of AI in healthcare.

Generative AI, particularly models like Generative Adversarial Networks (GANs) [32], Large Language Models (LLMs) [33], and VAEs are revolutionizing healthcare by synthesizing new data, creating realistic medical images and facilitating drug discovery. These models address data scarcity and variability challenges by generating synthetic yet realistic datasets for training and validating AI systems, all while maintaining patient privacy. For example, GMs are used to create synthetic medical images, such as abnormal brain MRIs, that aid in training diagnostic algorithms where real-world data are limited. These models accelerate innovation in drug discovery by designing novel molecular compounds and predicting their efficacy. Recent breakthroughs include generative approaches that create antibodies, paving the way for personalized therapies and new treatment modalities. Generative AI also extends its impact to clinical decision-making support. Studies evaluating tools like GPT-3.5 [34] in complex scenarios, such as therapy recommendations for brain gliomas [35], suggest that while these systems cannot replace medical experts, they offer valuable supplementary insights. As GMs advance, they are expected to enhance diagnostic accuracy further, streamline treatment planning, and improve patient and caregiver experiences. Figure 2.2 demonstrates an example from [36], showcasing the potential of generative AI in healthcare, particularly in medical imaging. The figure illustrates how state-of-the-art GAN models generate synthetic medical images from real datasets, highlighting the applicability of AI in overcoming data scarcity and advancing diagnostic research.

Integrating decentralized paradigms like FL into healthcare represents a paradigm shift, addressing critical challenges in data privacy and accessibility. Unlike traditional centralized models that require data pooling, FL enables AI systems to be trained collaboratively across decentralized datasets, ensuring that sensitive patient information remains secure and compliant with regulations like the GDPR. This approach is particularly impactful in

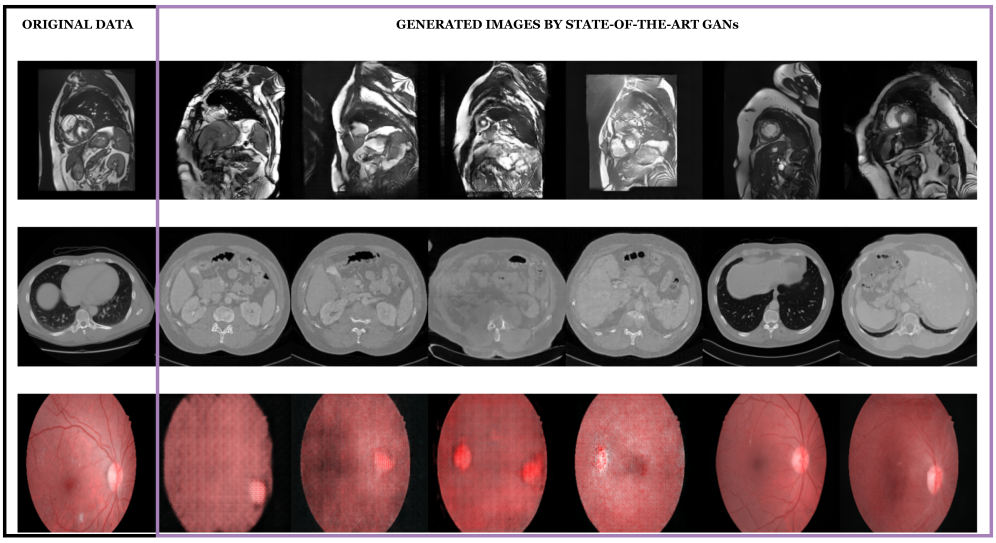


Figure 2.2: Examples of generative AI outcomes. This figure, taken from [36], illustrates the potential of generative AI in healthcare by creating synthetic medical images using state-of-the-art GAN-based models. The first column displays real medical images (MRIs, CT scans, and retinal scans), while the subsequent columns show synthetic images generated by these models.

multicenter collaborations, such as pooling data from diverse hospitals to develop more robust diagnostic tools or predictive models for rare diseases. FL promotes inclusivity by leveraging diverse datasets, mitigating biases, and improving model generalizability. In oncology, for example, it has been employed to predict treatment outcomes using data from geographically dispersed clinics [37]. As decentralized frameworks mature, they promise to drive large-scale innovation while preserving patient confidentiality.

Wearable devices and AI are transforming real-time health monitoring and management. These technologies collect continuous streams of physiological data, such as heart rate, oxygen saturation, and blood glucose levels, which AI algorithms analyze to detect anomalies and predict potential health issues. AI-powered wearables have demonstrated remarkable capabilities in managing chronic conditions and providing early warnings for acute health events. For instance, smartwatches equipped with AI can identify atrial fibrillation episodes and alert users to seek timely medical intervention [38]. During the COVID-19 pandemic, wearable technologies proved invaluable in remote rehabilitation, monitoring patient recovery, and supporting telemedicine initiatives [39]. Integrating wearable technologies with advanced AI algorithms is expected to expand personalized medicine. By analyzing data trends, these systems can provide tailored health recommendations, enabling users to proactively manage their health and reducing the need for frequent hospital visits. An example of recent advancements in AI-enabled wearable devices is shown in Figure 2.3[9]. This device is designed for pregnancy monitoring and tracking maternal and fetal health metrics like heart rate and uterine activity. Such technologies exemplify how integrating wearable devices with advanced AI algorithms is expected to expand personalized medicine. By analyzing data trends, these

[9] Image Source. Device Source (Accessed on December 6th, 2024)

systems can provide tailored health recommendations, enabling users to proactively manage their health and reducing the need for frequent hospital visits.

Figure 2.3: Example of AI-powered wearable technology. Modern AI-enabled wearable device designed for pregnancy monitoring. These advanced devices continuously track maternal and fetal health metrics, such as heart rate and uterine activity, providing real-time insights to enhance prenatal care and early detection of potential complications.

The future of AI in healthcare lies in developing multimodal AI systems capable of integrating diverse data types—from medical images and EHRs to genomic sequences and wearable device outputs. These systems aim to comprehensively understand the health of a patient, enabling clinicians to make more informed decisions. Recent innovations in multimodal models already demonstrate the ability to combine structured and unstructured data for tasks like predicting disease progression and recommending treatment plans. By synthesizing information from multiple sources, these systems improve diagnostic precision, optimize resource allocation, and personalize care delivery.

As AI technologies advance, ensuring their ethical and transparent deployment remains paramount. The development of XAI systems is a key focus, enabling clinicians to interpret and trust AI-driven recommendations. Ethical considerations also include addressing biases in AI algorithms, ensuring equitable healthcare outcomes, and fostering patient trust. Regulatory frameworks are adapting to this evolving landscape. Organizations such as the WHO emphasize the importance of AI systems, prioritizing safety, effectiveness, and equitable access. Regulatory innovations, including risk-based approaches and dynamic oversight mechanisms, facilitate deploying wearable AI solutions that uphold public safety while driving innovation.

The convergence of GMs, FL, wearable technologies, and ethical frameworks signals a synergistic future for AI in healthcare. These trends are not isolated; they complement each other, creating a robust ecosystem for innovation. GMs benefit from decentralized data enabled by FL; wearable technologies enhance real-time data collection, and ethical frameworks ensure these advances are deployed responsibly. Together, these trends promise to redefine healthcare by enabling earlier diagnoses, personalized treatments, and efficient health monitoring.

2.2 State of the Art of Survival Analysis in Healthcare

SA represents a key tool in the intersection of AI and healthcare, offering a statistical framework specifically designed to analyze and model time-to-event data. These 'events' could range from death and disease recurrence to recovery or other healthcare milestones. With the advent of EHRs, wearable technologies, and patient monitoring systems, the availability of rich, complex datasets has increased, enabling the application of SA to address pressing medical challenges. SA facilitates predictions of critical events such as disease progression and treatment outcomes but also aids in bridging statistical methodologies with real-world healthcare decision-making.

SA first appeared in the 17$^{\text{th}}$ century with the pioneering tables by John Graunt about mortality and life expectancy [40], which established the foundation for actuarial science and demographic studies. Initially focused on analyzing death rates and life spans, SA evolved over centuries to encompass various applications within biomedical research and fields like sociology, criminology, marketing, and institutional research. By the mid-20$^{\text{th}}$ century, advancements in statistical methodologies and computational capabilities allowed SA to expand its scope significantly. In healthcare, it became instrumental in evaluating the impact of interventions, understanding disease progression, and informing patient care strategies. Its importance has only grown with the integration of modern computational techniques, including AI, which enables the analysis of high-dimensional, heterogeneous, and often incomplete datasets.

The applications of SA in healthcare are extensive and diverse, playing a key role in prognosis, risk assessment, treatment evaluation, and operational planning. Prognostic modeling is one of its most prominent uses, where SA estimates survival probabilities and predicts the likelihood of events such as disease recurrence or death. For instance, survival curves, a core output of SA, enable oncologists to visualize and assess expected outcomes for cancer patients undergoing specific treatments. Similarly, SA supports risk stratification by identifying patients at varying levels of risk for adverse events. This stratification informs personalized care, such as guiding preventive strategies for individuals at high risk of cardiovascular disease.

In clinical research, SA is essential to evaluate treatments by comparing their effectiveness in controlled trials. For example, the Kaplan-Meier (KM) [41] survival curve and log-rank test are widely employed to determine differences in time-to-event outcomes between treatment groups. Furthermore, SA addresses complex scenarios involving CR, where multiple potential events (e.g., death from different causes) may occur. By accounting for interdependencies between these risks, SA provides accurate cause-specific risk assessments, which are critical for designing effective interventions and treatment plans.

Beyond individual patient care, SA informs predictions about the progression of chronic diseases. For instance, it models the time to significant milestones, such as kidney failure in chronic kidney disease or the onset of symptoms associated with Alzheimer's. This predictive capability is invaluable for long-term patient management. Additionally, SA contributes to healthcare operations by optimizing hospital resource allocation. SA supports efficient bed utilization and staffing strategies by modeling patient length-of-stay or predicting discharge times.

SA remains a cornerstone of modern healthcare analytics, empowering clinicians and researchers to extract meaningful insights from complex temporal data. Its integration with AI promises to expand its applicability further, making it a critical component of personalized medicine and healthcare optimization.

2.2.1 Fundamentals

The foundations of SA lie in its mathematical constructs, which enable robust modeling and interpretation of time-to-event data. These constructs include the survival function, hazard function, cumulative hazard, and considerations for censored data.

In a conventional time-to-event or SA setup, a dataset D consists of N observations. Each of these observations is described by triplets $D = (x_i, t_i, d_i)_{i=1}^{N}$, where $x_i = (x_i^1, ..., x_i^c, ..., x_i^{Cov})$ is an Cov-dimensional vector where $c = 1, 2, ..., Cov$ indexes the covariates, t_i is the time-to-event, and $d_i \in \{0, 1\}$ is the censoring indicator. When $d_i = 1$, the event of interest is observed at time t_i; conversely, $d_i = 0$ indicates that the event is censored and has not occurred within the observed time.

The primary objective of SA is to model the probability of event occurrence over time. Central to this are three mathematical constructs:

1. **Survival Function $S(t|x)$:** The survival function represents the probability that the event occurs after a given time t, conditional on the covariates x, and is defined as:

$$S(t|x) = P(T > t|x) = 1 - F(t|x), \tag{2.1}$$

   where $F(t|x)$ is the Cumulative Distribution Function (CDF) of the event times and T is the random variable representing time-to-event (we use capital T for the random variable and t for its realizations). The survival function monotonically decreases with t. This function provides a holistic view of survival probabilities over time, which is critical for prognosis and treatment planning.

2. **Hazard Function $h(t|x)$:** The hazard function quantifies the instantaneous rate of event occurrence at a specific time t, given survival up to that time:

$$h(t|x) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t, x)}{\Delta t}. \tag{2.2}$$

   The hazard function highlights periods of increased risk and is instrumental in identifying high-risk patient cohorts. Like $S(t|x)$, $h(t|x)$ is a non-negative function. While all the survival functions $S(t|x)$ decrease over time, the hazard function $h(t|x)$ can have a variety of shapes. *For ease of notation, we will drop the dependence of the survival and hazard functions with the covariates.*

   Using the relations between the survival and probability density functions, $h(t)$ can also be expressed as:

$$h(t) = \lim_{\Delta t \to 0} \frac{F(t + \Delta t) - F(t)}{\Delta t \cdot S(t)} = \frac{p(t)}{S(t)}, \tag{2.3}$$

where $p(t) = \frac{dF(t)}{dt}$ is the time Probability Density Function (pdf). The time pdf provides the density of the event times and encapsulates the interplay between the instantaneous risk and the probability of surviving up to that time as $p(t) = h(t)S(t)$.

3. **Cumulative Hazard Function $H(t)$:** $H(t)$ aggregates the hazard over time, providing a cumulative measure of risk:
$$H(t) = \int_0^t h(u)du. \tag{2.4}$$
The integral captures the total 'amount of risk' accumulated from the start of the observation to time t. While less commonly used directly in analysis, $H(t)$ establishes an essential link between the hazard and survival functions.

4. **Relations between $S(t)$, $H(t)$ and $h(t)$:** The cumulative hazard function $H(t)$ and the survival function $S(t)$ are connected as follows:
$$S(t) = e^{-H(t)}. \tag{2.5}$$
This relation arises from the definition of the hazard function and the survival function:
$$h(t) = -\frac{d}{dt} \ln S(t). \tag{2.6}$$
Integrating both sides from 0 to t gives:
$$\ln S(t) = -\int_0^t h(u)du = -H(t). \tag{2.7}$$
Exponentiating both sides yields:
$$S(t) = e^{-H(t)}. \tag{2.8}$$
This exponential relation illustrates how the cumulative hazard encapsulates the survival probability in a multiplicative decay framework.

Figure 2.4 visually illustrates the relationships between these functions, providing a conceptual understanding of their interplay.

Among these functions, $S(t)$ and $h(t)$ are the most widely used in practice due to their interpretability and direct applicability to clinical and operational decision-making. The cumulative hazard function $H(t)$, while mathematically significant, is less commonly used directly, as it primarily serves to bridge $h(t)$ and $S(t)$. The probability density function $p(t)$ is often a byproduct of these analyses, used primarily in model fitting and estimation.

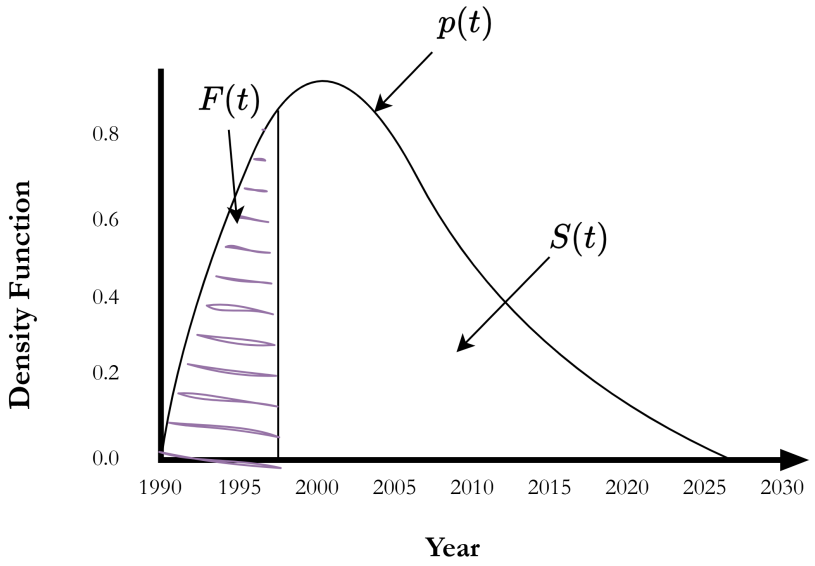


Figure 2.4: Relationships between key SA functions. Interplay between the survival function $S(t)$, the CDF $F(t)$, and the probability density function $p(t)$ in the context of time-to-event modeling.

A distinct feature of SA is its ability to handle censored data where the event of interest is not observed during the study period [42]. Censoring is ubiquitous in real-world healthcare scenarios, such as clinical trials, longitudinal studies, or chronic disease management, and arises due to various reasons [43]:

- **Right Censoring:** Occurs when the event time of a subject is unknown because they are lost to follow-up, withdraw from the study, or the study ends before the event occurs. For example, a patient surviving beyond the study period has a right-censored observation. This type of censoring is the most common form [44]. Mathematically, if a sample i is right-censored at time t_i, the true event time T_i is unknown but satisfies $T_i > t_i$.
- **Left Censoring:** Occurs when the event of interest has already occurred before the subject enters the study. For instance, patients diagnosed with a disease before enrollment may have a left-censored time-to-diagnosis. Mathematically, if a sample i is left-censored at time t_i, the true event time T_i is unknown but satisfies $T_i < t_i$.
- **Interval Censoring:** In this case, the exact event time is unknown but lies within a known interval. This is common in periodic follow-ups or scheduled clinical visits. Mathematically, for interval censoring, the true event time T_i satisfies $t_{i1} < T_i \geq t_{i2}$, where t_{i1} and t_{i2} are the bounds of the observed interval.

Figure 2.5 illustrates the different types of censoring commonly encountered in SA, highlighting the defined scenarios.

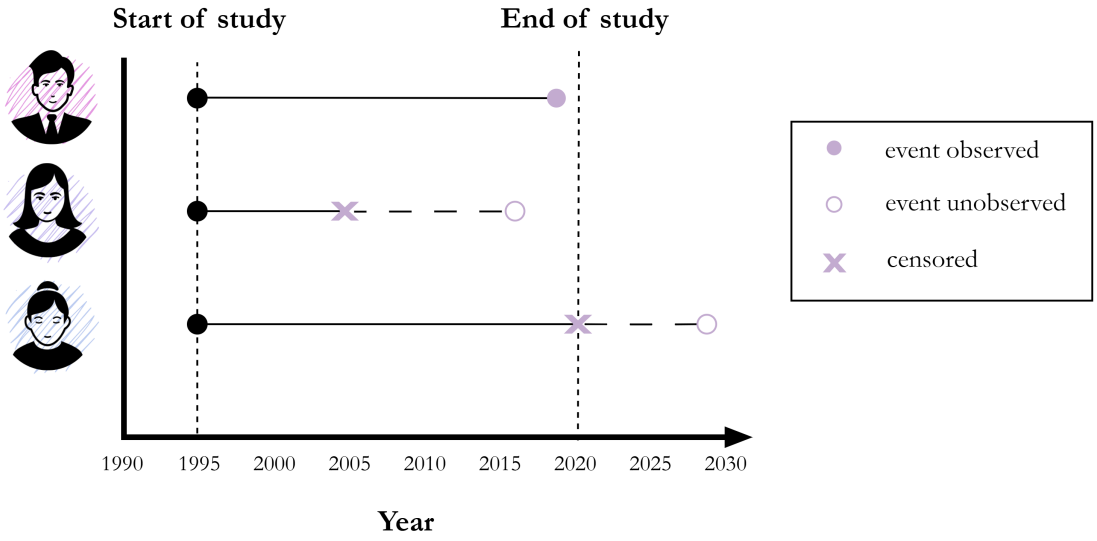


Figure 2.5: Timeline of a study with three patients showing SA outcomes. Solid circles indicate observed events, open circles indicate unobserved events, and 'X' marks denote censored data points. The timeline spans from the start to the end of the study.

2.2.2 Classical Approaches

SA has a rich history rooted in classical statistical methodologies that remain foundational for time-to-event data analysis. Among these, the KM estimator and Cox proportional hazards model (CoxPH) [45] provide robust tools for modeling and interpreting survival data. Despite the emergence of more complex ML and DL techniques, these classical approaches retain relevance due to their interpretability, simplicity, and effectiveness in many applications.

While both classical statistical methods and ML-based approaches share the target of predicting survival times and estimating survival probabilities, their emphases differ. Classical methods focus on characterizing the distributions of event times and the statistical properties of parameter estimation, typically visualized through survival curves. Conversely, ML approaches prioritize predictive accuracy, often leveraging the power of traditional SA frameworks in conjunction with advanced algorithms to improve the prediction of event occurrences at specific time points. This distinction underscores the complementary nature of traditional and modern methods in advancing the field of SA.

Classical statistical methods in SA can be classified into three categories based on the underlying assumptions and how model parameters are used: non-parametric, semi-parametric, and parametric.

Non-Parametric Methods

Non-parametric methods are particularly valuable in SA when no assumed underlying distribution for event times exists or when the proportional hazards assumption does not hold. These methods allow flexibility and adaptability in modeling survival data without imposing

restrictive assumptions.

Among these methods, the KM estimator remains the most widely used non-parametric approach to estimate the survival function $S(t)$, which represents the probability that an event of interest occurs after time t as defined previously. Introduced in 1958, the KM estimator [41] is particularly useful for datasets with censored observations, as it accounts for incomplete information about time-to-event data.

Let $T_1, T_2, T_3, \ldots, T_K$ denote the distinct ordered event times for the population of N observations. At a specific event time T_j, there are $\delta_j \geq 1$ observed events, and $risk_j$ individuals considered 'at risk', where $risk_j$ accounts for those who have not experienced the event or been censored before T_j. The conditional survival probability beyond T_j is given by:

$$p(T_j) = \frac{risk_j - \delta_j}{risk_j}. \tag{2.9}$$

The KM estimator of $S(t)$ is then defined as the product of these conditional probabilities:

$$\hat{S}(t) = \prod_{T_j \leq t} \left(1 - \frac{\delta_j}{risk_j}\right), \tag{2.10}$$

where the product accumulates overall event times T_j up to t.

The KM curve provides a stepwise estimate of the survival probability over time, with each step corresponding to an event. Its key advantages include:

1. **Non-parametric Nature:** The KM estimator does not require assumptions about the underlying distribution of survival times, making it highly flexible.
2. **Handling of Censoring:** The estimator accommodates right-censored data, ensuring accurate survival probability estimates even when some individuals are lost to follow-up.

However, this method presents two main limitations: (1) it cannot directly incorporate covariates to adjust for individual differences in risk factors, and (2) it assumes that censored individuals have the same survival probability as those who remain uncensored, which may not always hold.

The Nelson-Aalen (NA) estimator [46], [47], another non-parametric method, provides a cumulative hazard function estimate based on modern counting process techniques. It is particularly useful for analyzing the hazard of events over time. The NA estimator is defined as:

$$\hat{H}(t) = \sum_{T_j \leq t} \frac{\delta_j}{risk_j}, \tag{2.11}$$

where δ_j is the number of events that occurred at time T_j, and $risk_j$ is again the number of individuals at risk (not censored) before T_j. While less commonly used for direct survival probability estimation than the KM estimator, the NA method provides valuable insights into the event intensity over time. It serves as a basis for more advanced SA techniques.

The Life-Table (LT) method [48] is a variant of the KM estimator designed for interval-grouped survival data. Instead of estimating survival probabilities continuously, the LT

method aggregates data into time intervals, estimating survival probabilities for each interval based on grouped events and risk counts. This approach is particularly advantageous for large populations or datasets with interval-censored data. However, the interval-based nature of LT reduces precision compared to the continuous-time KM method.

These non-parametric methods form the foundation of SA, offering robust tools to estimate survival and hazard functions without relying on stringent distributional assumptions. Their adaptability makes them indispensable when flexibility and minimal assumptions are paramount.

Semi-Parametric Models: Cox Proportional Hazards Model

CoxPH [45], developed in 1972, is one of the most influential tools in SA. It provides a robust framework for evaluating the effects of multiple covariates on survival outcomes. As a hybrid between parametric and non-parametric approaches, CoxPH offers the flexibility of not requiring a predefined distribution for survival times while enabling precise estimation of covariate effects through the proportional hazards assumption. This balance makes CoxPH particularly valuable in clinical investigations where multiple interacting factors influence survival.

Unlike simpler methods like the KM estimator, which considers a single covariate at a time, the CoxPH model accommodates multivariate data, allowing researchers to analyze the effect of several covariates simultaneously. For instance, factors like age may also play a significant role when comparing survival outcomes for patients with different genotypes. CoxPH adjusts for these additional variables, providing a comprehensive estimate of the effect size of each covariate while controlling for confounding factors.

The CoxPH model specifies the hazard function $h(t)$, representing the instantaneous rate of event occurrence at time t, as:

$$h(t) = h_0(t) \exp(\beta^T x), \tag{2.12}$$

where $h_0(t)$ is called baseline hazard and represents the hazard value if all covariates are equal to zero (the quantity $\exp(0)$ equals 1), and β is the vector of regression coefficient quantifying the effect of covariates x on the hazard. The proportional hazards assumption ensures that the hazard ratio HR between any two individuals remains constant over time, independent of the baseline hazard. For two individuals with covariate vectors x_1 and x_2, the hazard ratio is given by:

$$HR = \frac{h(t|x_1)}{h(t|x_2)} = \exp(\beta^T(x_1 - x_2)). \tag{2.13}$$

This formulation emphasizes the relative risk of event occurrence between individuals while abstracting the absolute hazard rate.

The CoxPH model employs a semi-parametric approach, leaving $h_0(t)$ unspecified, which makes it more flexible than fully parametric models. The regression coefficients β are estimated using the partial likelihood method, which depends only on the observed event times and avoids

explicitly modeling the baseline hazard. For each covariate, the hazard ratio $HR = \exp(\beta)$ provides a direct interpretation:

- A hazard ratio of 1 ($\beta = 0$) indicates no effect of the covariate on the hazard.
- A hazard ratio greater than 1 ($\beta > 0$) signifies an increased hazard (i.e., a higher likelihood of the event and shorter survival time).
- A hazard ratio less than 1 ($\beta < 0$) denotes a decreased hazard (i.e., a lower likelihood of the event and longer survival time).

The survival function $S(t)$, which represents the probability of survival beyond time t, can be derived from the hazard function as:

$$S(t) = S_0(t)^{\exp(\beta^T x)}, \tag{2.14}$$

where $S_0(t) = \exp(-H_0(t))$ is the baseline survival function, and $H_0(t)$ is the cumulative baseline hazard function. The estimator introduced by Breslow [49] is commonly used to approximate $H_0(t)$, given by:

$$\hat{H}_0(t) = \sum_{t_i \leq t} \frac{1}{\sum_{j \in Risk_i} \exp(\beta^T x_j)}, \tag{2.15}$$

where $Risk_i$ denotes the set of individuals at risk at time t_i.

Extensions of the CoxPH model

Given the limitations of the basic CoxPH model, such as its reliance on the proportional hazards assumption, various extensions and adaptations have been developed:

- **Regularized Cox models:** These include Lasso-Cox [50], Ridge-Cox [51], and Elastic Net (EN) [52] Cox models, which incorporate penalty functions to handle high-dimensional data and select significant features while mitigating overfitting. For example, Lasso applies an l_1-norm penalty for sparse feature selection, while Ridge uses an l_2-norm for handling correlated features.
- **CoxBoost [53]:** This method applies gradient boosting to Cox regression, enabling the inclusion of mandatory covariates while improving prediction performance in high-dimensional settings.
- **Time-Dependent Cox Models [54]:** These models handle covariates that change over time, such as blood pressure or temperature, by modifying the hazard function to incorporate time-varying effects.

The CoxPH model is extensively used in healthcare to evaluate survival outcomes in diseases such as cancer, cardiovascular conditions, and chronic illnesses. It is particularly effective in assessing the impact of demographic, clinical, and genetic factors. However, the model has limitations, including its reliance on the proportional hazards assumption, which may not hold in all datasets. This has motivated the development of robust variants and extensions to ensure its continued applicability in modern SA.

Parametric Models

Parametric survival models provide a robust alternative to non-parametric (e.g., KM) and semi-parametric (e.g., CoxPH) methods by assuming that survival times, or their logarithmic transformations, follow a specific theoretical distribution, such as exponential, Weibull, log-normal, or log-logistic [43]. These models are particularly valuable when the assumed distribution aligns well with the underlying data, offering a simple and efficient framework for predicting survival probabilities and time-to-event outcomes.

Parametric models assume that the survival time T, or its logarithm $\ln(T)$, adheres to a known probability distribution. For instance:

- The Exponential distribution assumes a constant hazard rate over time, characterized by a single parameter λ. Its survival function is given by:

$$S(t) = \exp(-\lambda t), \tag{2.16}$$

  where $\lambda > 0$ is the rate parameter. This simplicity makes it ideal for scenarios where the event risk is time-invariant.

- The Weibull distribution, one of the most versatile parametric models, introduces a shape parameter $\alpha > 0$ to allow for varying hazard rates:

$$S(t) = \exp\Big(-\Big(\frac{t}{\lambda}\Big)^{\alpha}\Big), \tag{2.17}$$

  where $\lambda > 0$ is the scale parameter and $\alpha > 0$ determines the shape of the hazard function. If $\alpha = 1$, the Weibull model reduces to the exponential model. When $\alpha > 1$, the hazard rate increases over time, while $\alpha < 1$ implies a decreasing hazard rate.

- The Log-normal and Log-logistic distributions model survival times using transformations to log space, allowing for non-monotonic hazard functions. These distributions are particularly useful when hazard rates initially increase and decrease over time, capturing more complex survival patterns.

Parametric models utilize the Maximum Likelihood Estimation (MLE) technique to estimate model parameters. Suppose the dataset consists of N with v censored and $N - v$ uncensored observations. Let $\beta = (\beta_1, ..., \beta_P)^T$ denote the parameter vector. For a given instance i, if the event occurs at time T_i, the likelihood is driven by the probability density function $f(T_i, \beta)$. Conversely, if the instance is censored at time V_i, the likelihood incorporates the survival function $S(V_i, \beta)$, reflecting the probability that the event has not yet occurred.

The combined likelihood function for all instances is given by:

$$L(\beta) = \prod_{d_i=1} f(T_i, \beta) \prod_{d_i=0} S(V_i, \beta). \tag{2.18}$$

The MLE framework optimizes this likelihood to estimate the parameters β, enabling predictions for survival probabilities and hazard rates.

A special class of parametric models is the Accelerated Failure Time (AFT) model [54], which assumes a linear relationship between the logarithm of survival time and covariates:

$$\ln(T) = X\beta + \sigma\epsilon, \tag{2.19}$$

where $\sigma > 0$ is an unknown scale parameter, and ϵ is an error term following a specified distribution. AFT models are particularly appealing as they provide direct estimates of time acceleration or deceleration due to covariates, allowing for intuitive interpretations of how covariates influence the time to the event of interest.

Parametric survival models are widely used in healthcare, engineering, and reliability analysis due to their efficiency and ability to directly estimate survival probabilities, hazard rates, and median survival times. For example, the Weibull model is often applied in cancer prognosis studies, while the log-normal model analyzes treatment durations for chronic diseases.

However, the primary limitation of parametric models is their reliance on the assumed distribution. The resulting estimates can be biased and misleading if the chosen distribution does not align well with the underlying survival times. Non-parametric or semi-parametric methods are more appropriate for datasets with no theoretical distribution.

Despite these limitations, parametric models remain essential in SA. They offer a powerful and interpretable framework for understanding time-to-event data. Their mathematical rigor and computational efficiency combination ensures their continued relevance across diverse domains.

2.2.3 Machine Learning Approaches

ML has significantly advanced SA by addressing the limitations of traditional statistical models such as KM estimators and CoxPH models. While classical methods provide interpretable and foundational tools, they often struggle with high-dimensional, nonlinear, and multimodal data. ML methods, ranging from tree-based algorithms to DL architectures, enable the modeling of complex patterns in survival data and the integration of diverse data sources, such as genomic profiles, medical imaging, and EHRs.

Tree-Based Methods

Tree-based approaches are among the earliest ML methods adapted for SA. They extend decision tree algorithms to accommodate censored data, recursively partitioning the data based on covariates to model time-to-event outcomes.

- **Random Survival Forests (RSFs)** [55]: RSF improves upon survival trees by creating ensembles of survival trees using bootstrap aggregation (bagging) and random feature selection. This method enhances predictive accuracy and robustness while effectively handling complex interactions among covariates and right-censored data.
- **Gradient-Boosted Survival Trees (GBST)** [53]: GBST refines survival predictions by iteratively optimizing a survival-specific loss function, offering improved accuracy over single-tree models.

Bayesian Methods

Bayesian approaches provide a probabilistic framework for SA, integrating prior knowledge with observed data. Bayesian Networks (BNs) model dependencies among covariates, while

Naïve Bayes classifiers offer a simplified yet effective probabilistic approach to survival prediction [56]. These methods excel in scenarios with limited data or where uncertainty quantification is critical, such as clinical decision-making.

Support Vector Machines

Support Vector Machines (SVMs) have been adapted for SA by incorporating methods to handle censored data. For example, Support Vector Regression for Censored Data (SVRc) [57] modifies the traditional SVM loss function to account for censored observations, enabling robust predictions in high-dimensional spaces. However, these methods may face computational challenges with large datasets due to their quadratic time complexity.

Deep Learning

DL methods have transformed SA by introducing models that capture complex, nonlinear relationships in high-dimensional datasets. They provide increased flexibility and predictive accuracy compared to traditional approaches. The key DL contributions to SA include:

- **Cox-Based Extensions:** The integration of Neural Networks (NNs) with CoxPH models began with the Faraggi and Simon model [58], which introduced a feed-forward NN to replace the linear predictor of the CoxPH model, enabling the capture of nonlinear relationships between covariates and hazard risks. Building on this foundation:
  - **DeepSurv [59]**: Extends the Faraggi model by parameterizing the log-risk function with a deep feed-forward NN, enabling the modeling of nonlinear covariate relationships while maintaining the proportional hazards assumption. Optimized using stochastic gradient descent, DeepSurv is particularly effective in large datasets with complex, nonlinear interactions.
  - **Cox-Time [60]**: Introduces time-varying effects by allowing the hazard function to depend on both covariates and time. Relaxing the proportional hazards assumption offers greater flexibility but increases computational demands.
  - **Cox-nnet [61]** and **Cox-PASNet [62]**: These models adapt DeepSurv for high-dimensional data, such as genomics, using regularization techniques to prevent overfitting.
  - **Luck's Model [63]:** Proposes an extension of DeepSurv with additional regularization techniques to improve interpretability and robustness in high-dimensional datasets, particularly in genomics and imaging data.
- **Discrete-Time Models:**
  - **DeepHit [64]:** A groundbreaking model that directly estimates the survival distribution without relying on proportional hazards or specific parametric assumptions. Its architecture includes a shared sub-network for learning general representations and additional task-specific sub-networks for survival predictions. DeepHit is particularly suited for modeling nonlinear relationships between covariates and survival times.

  - **Dynamic-DeepHit [65]:** Builds on DeepHit by incorporating RNNs to handle longitudinal data with time-varying covariates. This model captures temporal dependencies and includes an auxiliary prediction loss to enhance its ability to predict future covariate values, making it ideal for time-series survival datasets.

- **Fully Parametric Models:** DL has also advanced fully parametric survival models by employing architectures such as RNNs to predict survival distributions. RNN-Surv [66] sequentially predicts survival distributions using RNNs, capturing dependencies across time. It is particularly effective in longitudinal datasets with time-varying covariates. Weibull Time-to-Event RNN [67], based on the Weibull distribution: this model offers flexibility in estimating survival times while preserving parametric assumptions about the hazard function.

ML, particularly DL, has revolutionized SA by enabling flexible, accurate predictions in high-dimensional and multimodal data settings. While traditional tree-based and Bayesian methods remain valuable for interpretability and computational simplicity, DL approaches like DeepSurv, DeepHit, and Dynamic-DeepHit provide unparalleled capacity to model nonlinear and time-varying relationships, making them indispensable tools in precision medicine and beyond.

2.2.4 Competing Risks

Building upon SA advancements, CR represents a natural extension of traditional single-risk models to address more complex and realistic scenarios. While state-of-the-art models like DeepSurv [59] and Cox-Time [60] have achieved remarkable success in estimating survival probabilities and hazard functions, these approaches are limited to single-risk frameworks. In contrast, CR scenarios involve multiple mutually exclusive events, where the occurrence of one event precludes the observation of others. Addressing CR scenarios requires specialized models that capture the interplay between risks and estimate the probability of each specific event. This challenge remains underexplored in the current landscape of SA research.

In traditional SA, the focus is on estimating the survival function $S(t)$, which represents the probability of an individual surviving beyond time t. The survival function operates under the assumption of a singular event of interest, with censoring accounting for the absence of an observed event or loss to follow-up. However, this framework becomes insufficient in CR settings, where individuals are exposed to competing events. For example, in a medical study, a patient may die due to cardiovascular disease or cancer, but the occurrence of one event precludes the observation of the other.

To accommodate these complexities, the data representation changes. Each instance is now characterized as (x_i, t_i, r_i), where r_i denotes the event type:

- $r_i = 0$: The instance is censored.
- $r_i = k$ ($k \in \{1, 2, ..., R\}$): Event k occurred, where R is the total number of possible events.

This expanded notation emphasizes the need to model the probability of each specific event

while accounting for the competing nature of other risks.

The Cumulative Incidence Function (CIF) replaces the survival function as the primary modeling target in CR scenarios. The CIF for event k is defined as:

$$CIF_k(t|x) = P(T \leq t, r = k|x), \tag{2.20}$$

where T is the time-to-event random variable, and $r = k$ specifies that event k occurred. Unlike the survival function, which aggregates all risks into a single metric, the CIF explicitly disentangles the probability of each specific event. Key properties of the CIF include:

1. The CIF for any event k does not reach 1 as $t \to \infty$ due to the presence of CR.
2. The sum of CIFs over all events gives the probability of experiencing any event by time t:
$$\sum_{k=1}^{R} CIF_k(t|x) = P(T \leq t|x), \tag{2.21}$$
which excludes censoring.

Classical Methods for CR: The Fine-Gray Model

The Fine-Gray (FG) model [68] is one of the most widely used classical approaches for CR. FG directly models the subdistribution hazard for each event type, allowing for estimating the CIF without explicitly modeling the cause-specific hazards. It achieves this by reformulating the hazard function to reflect the probability of an event occurring in the presence of CR. While effective, the FG model, like the Cox proportional hazards model, relies on strong assumptions about the underlying stochastic processes and the form of the hazard function. These assumptions, such as proportional hazards, often fail to align with the complexities of real-world data, motivating the need for more flexible, data-driven methods.

Deep Learning Approaches to CR

The emergence of DL in SA has led to significant advancements in addressing CR. While most DL-based models focus on single-risk scenarios, a few state-of-the-art methods explicitly incorporate CR into their frameworks. These include:

- **DeepHit** [64]: DeepHit is a foundational model in DL-based CR SA, directly estimating the CIF for each competing event. Already defined for single-event SA, DeepHit can also be extended to handle CR. Its architecture features a shared sub-network to learn general latent representations from covariates and cause-specific sub-networks that model CIFs for each competing event. DeepHit operates discretely and uses a composite loss function combining likelihood-based terms and a ranking loss to improve the C-index. While highly flexible, its discrete-time framework requires time discretization, which can lead to information loss. Moreover, DeepHit produces numerical CIF estimates without analytical expressions, limiting statistical computations and precision.
- **DeepComp** [69]: DeepComp extends the discrete-time approach by employing RNN-based cause-specific sub-networks. Unlike DeepHit, DeepComp outputs cause specific

discrete hazards for each time interval rather than CIFs. These hazards can be aggregated to estimate the cumulative incidence of each risk. RNNs enable DeepComp to capture temporal dependencies more effectively than feed-forward networks. However, the reliance on discrete hazards shares similar limitations with DeepHit regarding continuous-time modeling and precision.

- **CRESA** [70]: The CR and Recurrent-Event Survival Analysis (CRESA) model also employs RNN architectures and discrete-time modeling. Like DeepHit, it generates a final distribution over all CR using a loss function based on recurrent CIFs. Additionally, it incorporates a ranking component to improve performance on metrics like the C-index. The RNN-based design from CRESA is particularly advantageous for datasets with sequential or recurrent events, although its discrete-time framework imposes similar constraints to DeepHit and DeepComp.
- **SSMTL** [71]: Semi-Supervised Multi-Task Learning (SSMTL) transforms CR SA into a multi-task classification problem. Each time point (or competing event) is treated as a separate binary (or multi-class) classification task. SSMTL uses a custom composite loss function that includes a classification loss for uncensored data, a semi-supervised loss for censored data, and regularization losses. While innovative, SSMTL operates in discrete time, similar to DeepHit, and its reliance on binary classification tasks limits the granularity of its predictions.
- **DeepCompete** [72]: DeepCompete introduces a novel continuous-time approach for CR modeling. It employs neural ordinary differential equation (nODE) blocks within each cause-specific sub-networks cause-specific sub-network, which outputs cumulative hazard functions. Unlike discrete-time methods like DeepHit and CRESA, the continuous-time framework of DeepCompete allows for more precise modeling of survival times and CIFs. However, using nODE blocks increases the computational complexity of the model.

Despite these advancements, modeling CR remains challenging due to disentangling overlapping risks and effectively managing censoring. Current state-of-the-art methods often require large datasets and computational resources, which may limit their applicability in real-world scenarios. Future research should focus on developing scalable and interpretable models that integrate CR seamlessly into SA frameworks while addressing the limitations of discrete-time and proportional hazard assumptions.

CR highlights the complexities of real-world SA, necessitating specialized models that go beyond traditional single-risk frameworks. By leveraging the flexibility and capacity of DL, researchers are making significant strides in this domain, with methods like DeepHit paving the way for more comprehensive CR modeling. These innovations hold the potential to transform applications ranging from personalized medicine to risk management, where understanding the interplay of competing events is critical.

2.2.5 Validation Techniques

Model evaluation is a critical component of SA, as it provides insights into the performance and reliability of predictive models, especially in the presence of censored data and varying time horizons. Unlike traditional regression or classification tasks, SA involves the dual challenge of handling incomplete data due to censoring and accurately predicting time-to-event outcomes. Effective validation techniques assess model accuracy and ensure robustness across diverse datasets and scenarios. The evaluation process typically focuses on two key aspects: discrimination (the ability to rank individuals by risk) and calibration (the agreement between predicted probabilities and observed outcomes).

As outlined previously, each dataset is characterized by triplets $D = (x_i, t_i, d_i)_{i=1}^{N}$, where x_i represents the covariate vector, t_i denotes the time-to-event, and $d_i \in \{0, 1\}$ indicates the censoring status.

The Concordance Index (C-index) is one of the most widely used metrics to evaluate the discriminatory power of a survival model. It measures the rank correlation between predicted risks and observed event times, offering an intuitive understanding of whether higher predicted risks correspond to shorter observed survival times. The general formulation of the C-index is:

$$C_{index} = P\Big(\hat{F}(t|x_i) > \hat{F}(t|x_j)\Big| d_i = 1, t_i < t_j, t_i \leq t\Big), \tag{2.22}$$

where $\hat{F}(t|x_i)$ is the estimated CDF for the covariates x_i at time t. The C-index evaluates pairwise concordance among comparable samples, emphasizing the ability of the model to rank correctly.

The C-index is adapted in CR settings to evaluate the concordance of the predicted CIF for specific competing events. It considers the predicted CIF of the event of interest at a given time, compared against those of other individuals who have not yet experienced the event:

$$C_{index} \approx \frac{\sum_{i \neq j} A_{k,i,j} \cdot \mathbb{I}\Big(CIF_k(t|x_i) \leq CIF_k(t|x_j)\Big)}{\sum_{i \neq j} A_{k,i,j}}, \tag{2.23}$$

where $A_{k,i,j}$ is the indicator function for comparable pairs (i, j) for event k. This formulation evaluates the ability of the model to rank correctly for each specific competing event, considering the unique challenges introduced by multiple outcomes.

However, the original C-index [73] assumes time-invariant risk and cannot accommodate dynamic changes. To address this limitation, the time-dependent C-index [74] extends the metric by accounting for evolving risks.

The Brier Score (BS) evaluates the squared error between predicted survival probabilities and observed outcomes. It incorporates the Inverse Probability of Censoring Weighting (IPCW) [75] to adjust for censored observations, assigning higher weights to uncensored instances. The BS at a specific time t is given by:

$$BS(t) = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{(S(t|x_i))^2}{G(t_i)}\cdot\mathbb{I}(t_i < t, d_i = 1) + \frac{(1-S(t|x_i))^2}{G(t)}\cdot\mathbb{I}(t_i \geq t)\right], \tag{2.24}$$

where $G(t)$ is the survival function for censoring.

The BS measures the ability of the model to rank individuals accurately by risk (discrimination) and the alignment of predicted survival probabilities with observed event frequencies (calibration).

To provide a time-agnostic assessment, the Integrated Brier Score (IBS) calculates the average BS over a range of times:

$$IBS(t_{max}) = \frac{1}{t_{max}}\int_0^{t_{max}} BS(t)dt. \tag{2.25}$$

The IBS is particularly valuable for comparing models across varying time horizons and assessing their overall predictive performance.

In CR settings, the BS and IBS are calculated for the CIF of each competing event k, providing a measure of the discrepancy between predicted and observed CIFs over time:

$$BS_k(t) = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{(1-CIF_k(t|x_i))^2}{G(t_i)}\cdot\mathbb{I}(t_i < t, d_i = k) + \frac{(CIF_k(t|x_i))^2}{G(t)}\cdot\mathbb{I}(t_i \geq t)\right]. \tag{2.26}$$

The IBS for CR iBS_k is then calculated as:

$$iBS_k(t_{max}) = \frac{1}{t_{max}}\int_0^{t_{max}} BS_k(t)dt. \tag{2.27}$$

This metric evaluates the overall alignment between predicted and observed CIFs, accounting for the inherent complexity of multiple events.

Time-dependent metrics, like the time-dependent C-index and time-dependent BS, are essential for evaluating models that account for changes in risk over time. These metrics are especially relevant for modern survival models, such as RNNs and dynamic survival models, which explicitly model time-varying covariates. Time-dependent metrics ensure the evaluation reflects the dynamic nature of risks and outcomes in longitudinal datasets.

Censoring introduces unique challenges in SA, as it prevents the direct observation of event times for all individuals. This complicates the calculation of traditional metrics and necessitates specialized techniques like IPCW to handle censored observations. Key challenges include:

- **Biased Estimates:** Without proper adjustments, censored data can lead to biased model evaluations.
- **Comparable Pairs:** Metrics like the C-index rely on comparing pairs of individuals. Censored data reduce the number of comparable pairs, potentially lowering metric reliability.

- **Extrapolation:** For heavily censored datasets, models may overfit or extrapolate survival probabilities inaccurately, leading to unreliable calibration.

SA validation techniques ensure that models perform well in realistic scenarios characterized by censored data and dynamic risk profiles. Metrics like the time-dependent C-index and IBS provide comprehensive insights into the discrimination and calibration capabilities of a survival model. However, the presence of censoring necessitates thoughtful adaptations and robust statistical techniques to ensure unbiased and reliable evaluation. As SA models become more sophisticated, leveraging advanced metrics and addressing challenges with censored data will remain pivotal in advancing the field.

2.2.6 Challenges and Future Directions

Despite the significant progress in SA, particularly with the integration of ML and DL, several challenges and research gaps remain. These challenges, summarized in Figure 2.6, include handling high-dimensional data, modeling complex relationships, addressing censoring, and ensuring flexibility in training. Additionally, advancing methods for CR and incorporating parametric models into DL frameworks are critical areas for future research.

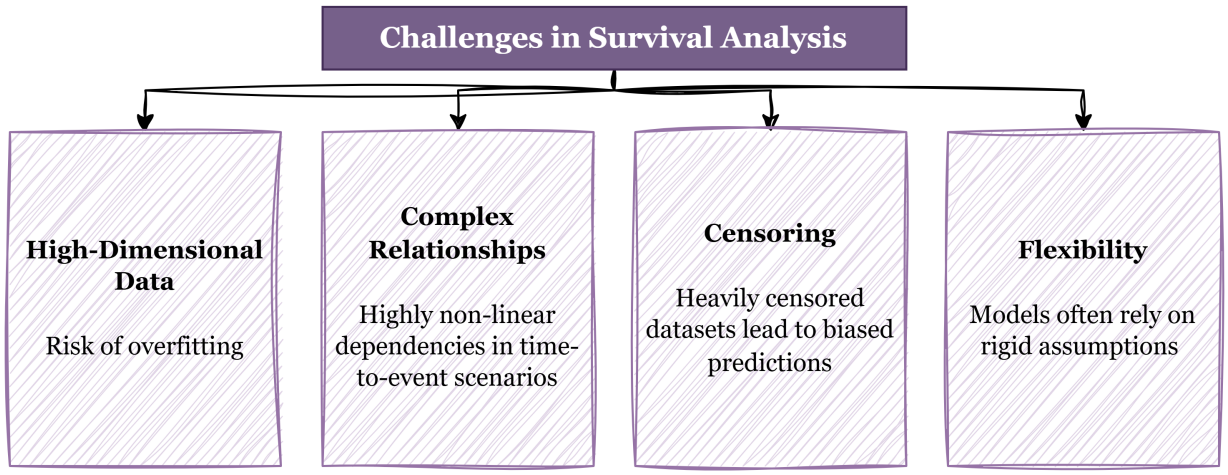


Figure 2.6: Challenges in Survival Analysis. A summary of key limitations, including handling high-dimensional data, modeling complex relationships, addressing censoring, and ensuring flexibility in training.

Current Limitations

1. **High-Dimensional Data:** Modern SA applications, such as those in genomics and medical imaging, often involve datasets with numerous covariates relative to the sample size. DL models, while powerful, can struggle with overfitting in such settings. This limitation necessitates robust regularization techniques and effective feature selection to ensure meaningful and reliable model performance.
2. **Complex Relationships Between Covariates:** DL models are adept at capturing intricate relationships within data; however, modeling highly non-linear and intricate

dependencies, especially in time-to-event scenarios, remains challenging. Ensuring these models adequately capture interactions without introducing biases or artifacts is difficult.

3. **Censoring:** Heavily censored datasets reduce the amount of observable information for model training, leading to biased predictions and reduced reliability of validation metrics. Current approaches, such as IPCW, offer partial solutions but are not universally robust.
4. **Flexibility:** Current approaches often require specific assumptions about time distributions or pre-defined functional forms, reducing their adaptability to diverse data settings. Flexible models that do not rely on rigid time assumptions are crucial for broad applicability across real-world scenarios.

Research Gaps

1. **Integration of CR:** Although CR modeling has seen notable advancements in DL-based SA, current methods often struggle to integrate multiple competing events into a unified framework seamlessly. Future research should focus on developing more effective architectures for CR modeling that balance accuracy, interpretability, and computational efficiency.
2. **Parametric Models with DL:** The integration of parametric survival models within DL architectures is underexplored. Parametric models offer interpretable survival estimates and flexible, functional forms, but combining these strengths with the representation learning capabilities of DL could unlock powerful hybrid approaches for SA. Future work should design methods that effectively merge these paradigms, ensuring flexibility and interpretability.

Addressing these challenges requires a multidisciplinary approach combining optimization innovations, neural architecture design, and statistical theory. Future efforts should prioritize methods that balance interpretability with adaptability, ensuring robustness in real-world, complex datasets. By addressing these limitations and research gaps, SA can become a reliable tool for critical healthcare domains.

2.2.7 State-of-the-Art Models Summary

The development of SA methods spans classical, ML, and DL paradigms, each offering unique advantages and limitations. Classical methods like the KM estimator and CoxPH provide interpretable and foundational tools for modeling time-to-event data. ML approaches like RSF and Bayesian models enhance predictive capabilities by capturing complex patterns in the data. DL methods, including DeepSurv and Dynamic-DeepHit, further expand the modeling capacity to handle nonlinear relationships, high-dimensional covariates, and CR.

Table 2.2 provides a comprehensive summary of key SA models discussed in Section 2.2. The table categorizes the models into classical, ML, and DL approaches while detailing their type (e.g., parametric, non-parametric, semi-parametric), support for CR, covariate handling, and distinguishing features. This overview highlights the progression from interpretable statistical models to advanced neural networks capable of addressing the complexities of real-world datasets.

Category	Model	Type	CR Support	Covariate Handling	Key Features
Classical	KM estimator [41]	Non-parametric	No	None	Stepwise survival estimation
	NA estimator [46], [47]	Non-parametric	No	None	Cumulative hazard estimation
	LT estimator [48]	Non-parametric	No	None	Interval-grouped survival estimation
	CoxPH [45]	Semi-parametric	No	Multivariate	Proportional hazards; interpretable hazard ratios
	Cox Regularized (Lasso [50], Ridge [51], EN [52])	Semi-parametric	No	High-dimensional	Handles correlated/large covariates with penalties
	CoxBoost [53]	Semi-parametric	No	Multivariate	Gradient boosting; feature selection
	TD-Cox [54]	Semi-parametric	No	Time-varying covariates	Models dynamic risk factors
	FG [68]	Semi-parametric	Yes	Multivariate	Subdistribution hazards; CR-specific
	Exponential [43]	Parametric	No	Multivariate	Constant hazard; simple and interpretable
	Weibull [43]	Parametric	No	Multivariate	Flexible hazard (increasing/ /decreasing) widely used
Machine Learning	RSF [55]	Non-parametric	No	Multivariate	Ensemble of trees; handles censoring
	GBST [53]	Non-parametric	No	Multivariate	Survival-specific loss function
	Bayesian Methods [56] (BN, Naïve Bayes)	Parametric	No	Multivariate	Probabilistic modeling; quantifies uncertainty
	SVRc [57]	Non-parametric	No	Multivariate	Adapts SVMs for censored data
Deep Learning	Faraggi and Simon's model [58]	Semi-parametric	No	Multivariate	First DL extension of CoxPH; captures nonlinear effects
	DeepSurv [59]	Semi-parametric	No	Multivariate	Nonlinear relationships; proportional hazards
	Luck's Model [63]	Semi-parametric	No	High-dimensional	Improved interpretability for genomics data
	Cox-Time [60]	Semi-parametric	No	Multivariate + Time	Time-varying hazards
	Cox-nnet [61] and Cox-PASNet [62]	Semi-parametric	No	High-dimensional	Optimized for high-dimensional genomic data
	DeepHit [64]	Non-parametric	Yes	Multivariate	Direct CIF estimation; discrete-time modeling
	Dynamic-DeepHit [65]	Non-parametric	Yes	Time-varying covariates	RNN-based temporal dependencies; CR modeling
	DeepComp [69]	Non-parametric	Yes	Multivariate	Cause-specific discrete hazards; RNN-based
	CRESA [70]	Non-parametric	Yes	Multivariate	Recurrent CIFs; RNN for sequential data
	SSMTL [71]	Non-parametric	Yes	Multivariate	Semi-supervised; multi-task classification
	DeepCompete [72]	Non-parametric	Yes	Multivariate	Continuous-time CR with nODE blocks

Table 2.2: Comparison of SA models described in Section 2.2. The table summarizes key models for SA, indicating their type, CR support, covariate handling, and distinguishing features.

2.3 State of the Art of Synthetic Data Generation in Healthcare

As discussed in earlier sections, the integration of AI in healthcare has shown remarkable potential in revolutionizing patient care, research, and operational efficiency. However, one of the key challenges hindering the full realization of these advancements is the limited availability of high-quality, large-scale datasets. This issue is especially critical in SA, where small sample sizes, high-dimensional covariates, and significant levels of censoring often characterize datasets. These limitations restrict developing, validating, and deploying robust ML and DL models.

The scarcity of survival data (and medical data in general) is compounded by stringent privacy regulations, such as GDPR and HIPAA, which impose restrictions on data sharing across institutions. These legal frameworks, designed to protect patient confidentiality, further limit access to the diverse and representative datasets needed to develop generalizable models. Moreover, the sensitive nature of medical data and the risk of re-identification create significant barriers to its widespread use and distribution.

SDG offers a promising solution to these challenges. By creating synthetic datasets that mimic the statistical and structural properties of real-world data, synthetic data can address key issues in SA and healthcare, including:

- **Augmenting existing datasets** to improve model training and validation.
- **Enabling data sharing** across institutions without violating patient privacy.
- **Simulating rare or hypothetical scenarios** to test model robustness and expand research possibilities.

This section delves into the state of the art of SDG, focusing on STDG and its application to healthcare. Key methodologies, such as VAEs and GANs, are explored, highlighting their ability to overcome the data scarcity challenges that constrain progress in this field. SDG is crucial in advancing AI-driven SA in healthcare by bridging the gap between data demand and real-world availability constraints.

2.3.1 Fundamentals

SDG has emerged as a critical tool in healthcare, addressing challenges inherent to using real-world data [5], [76]. In healthcare, patient data are among the most sensitive, with strict privacy regulations such as the GDPR and the HIPAA limiting their accessibility. SDG enables the creation of synthetic datasets that replicate the statistical properties, distributions, and structural relationships of real data without revealing sensitive information. This capability allows researchers and practitioners to leverage data-driven innovations while safeguarding privacy and maintaining regulation compliance.

The concept of SDG has its roots in Statistical Disclosure Control (SDC), which aims to protect the confidentiality of respondents in public-use datasets while preserving their inferential utility. This process balances the trade-off between preventing the disclosure of sensitive

information and maintaining the analytical value of the data. The idea of synthetic data as a tool for achieving disclosure control was first proposed in 1993 [77]. These early developments laid the foundation for modern SDG, highlighting its potential to enable valid statistical inferences while mitigating privacy risks. By generating synthetic datasets, researchers could maintain privacy and confidentiality without compromising the utility of the data—a principle that remains central to SDG today.

Several factors drive the need for synthetic data in healthcare. First, privacy concerns often prevent the sharing of real patient data, impeding research collaboration, algorithm development, and clinical innovations. Synthetic data can mitigate these concerns by removing Personally Identifiable Information (PII) while preserving the utility of the data for analysis. Second, data scarcity remains a significant issue, particularly in rare or neglected diseases, where insufficient data hampers the development of robust ML models. Synthetic data can augment real datasets, enabling better model training and improved generalization. Third, real-world datasets often exhibit biases or underrepresentation of specific demographic groups, leading to inequities in healthcare outcomes. Researchers can mitigate these biases and promote fairness in AI applications by generating balanced and representative synthetic datasets.

Healthcare involves a wide variety of data types, all of which can benefit from SDG. Structured tabular data, such as EHRs, contain patient demographics, diagnoses, and treatment information. Time-series data, often collected from wearable devices or biomedical sensors, provide longitudinal insights into patient health. Medical imaging data, such as CT scans and MRIs, are crucial for diagnostic and treatment planning. Text data, including clinical notes and unstructured EHR content, capture qualitative aspects of patient care. Each of these data types presents unique challenges and opportunities for SDG. For example, generating synthetic EHRs requires maintaining complex dependencies between variables, while generating synthetic medical images demands high-resolution similarity to reflect anatomical structures accurately.

The overall process of SDG involves four key stages, as depicted in Figure 2.7. These stages include (1) securely accessing and acquiring real-world data, (2) preparing and preprocessing the data to ensure it is suitable for modeling, (3) developing and training GMs to produce synthetic datasets, and (4) evaluating the generated data for resemblance, utility, and privacy. This systematic approach ensures that synthetic datasets are not only realistic but also preserve the privacy of the original data sources. This section will focus on steps (3) developing GMs and (4) evaluating synthetic data, as these are the core areas of innovation and methodology in synthetic data generation.

Synthetic data applications in healthcare are diverse and life-changing. Providing large, high-quality datasets that mirror real-world scenarios can support the development of AI models for diagnostics, drug discovery, and personalized treatment plans. Synthetic data also accelerates clinical trials by simulating patient populations, enabling faster treatment efficacy and safety evaluation. Furthermore, it democratizes access to data, empowering researchers across institutions and geographic locations to conduct experiments without the risks of sharing sensitive patient information. Advances in GMs, such as Stable Diffusion [78], DALL-E [79], and large language models like GPT [33] and ChatGPT [80], have revolutionized

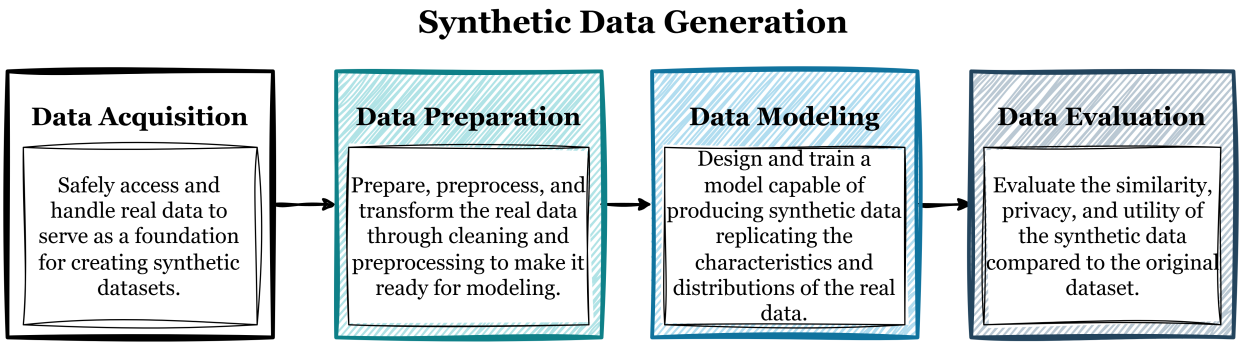


Figure 2.7: Synthetic data generation workflow. The key stages of the SDG process include data acquisition, preparation, modeling, and evaluation. Each step ensures that the resulting synthetic data maintains utility, privacy, and similarity to the original dataset.

data generation in images, text, and video. These advancements demonstrate the potential of SDG in creating realistic and high-quality synthetic datasets, which could accelerate medical research.

While synthetic data holds significant promise, maintaining a balance between realism and privacy is paramount. High-quality synthetic data must accurately replicate the statistical and structural properties of real data to ensure its utility in AI and ML applications. At the same time, it must guarantee privacy by preventing the re-identification of individuals. Striking this balance is particularly critical in healthcare, where the accuracy of AI models has direct implications for patient outcomes.

This thesis focuses specifically on STDG, as this data type is ubiquitous in healthcare and presents unique methodological challenges. Structured tabular data, such as EHRs, encompass rich patient-level information critical for predictive modeling, clinical decision support, and resource allocation. Tabular datasets often contain interconnected features, including explicit identifiers, quasi-identifiers, and sensitive attributes, necessitating robust SDG models capable of handling their intricate and longitudinal nature. While synthetic data can be generated for other types of healthcare data (e.g., imaging, time series, and text), providing a comprehensive review of all these domains would exceed the scope of this work. Therefore, our review exclusively centers on STDG in healthcare, laying the groundwork for a deeper exploration of methodologies, applications, and challenges in this domain.

2.3.2 Classical Statistical Approaches

Classical approaches to SDG represent some of the earliest methodologies developed to facilitate data sharing and analysis, particularly in sensitive domains such as healthcare. Rooted in statistical modeling and probabilistic techniques, these methods aim to replicate the distributions and relationships of real-world data while ensuring privacy protection. These foundational techniques have enabled researchers and practitioners to use artificial datasets for various applications, even though their inherent limitations have driven the development of more sophisticated approaches.

The origins of SDG are closely tied to anonymization techniques. These early approaches focused on protecting sensitive information by obfuscating or removing identifiable dataset features. Common strategies included adding noise, generalizing values, or suppressing data entirely. Methods such as k-anonymity [81] and l-diversity [82] were pivotal in establishing baseline privacy guarantees by ensuring that individual records could not be uniquely identified. However, while effectively reducing privacy risks, these techniques often suffered from significant information loss, limiting their utility in downstream applications like ML. Additionally, they struggled to capture complex dependencies in data, particularly in healthcare datasets where inter-variable relationships, such as those between patient demographics, diagnoses, and treatments, are critical. Furthermore, such techniques were prone to de-anonymization or re-identification attacks [83], [84], further emphasizing the need for advanced approaches.

Classical statistical models marked the transition from anonymization to SDG, focusing on recreating the statistical properties of real data through parametric and non-parametric methods. These models aim to estimate data distributions and generate synthetic samples that closely align with the characteristics of the original dataset. While effective for certain analytical and planning tasks, their limitations often render them unsuitable for advanced clinical applications [85]. Below, we explore several key approaches, emphasizing their applications, strengths, and limitations in healthcare contexts.

- **Mixture Models:** Mixture models decompose complex datasets into simpler components, enabling accurate representation of heterogeneous data. This makes them particularly well-suited for medical datasets, which often consist of distinct patient subgroups. For example, Gaussian Mixture Models (GMMs) have been applied to synthesize medical data, effectively capturing continuous features but initially struggling with discrete variables in [86]. A subsequent study [87] extended this approach to mixed data types, achieving more robust performance across a broader range of features. These models have successfully modeled patient populations but require careful tuning to maintain inter-feature relationships.
- **Copulas [88], [89]:** Copulas provide a mathematical framework to model multivariate dependencies by separating marginal distributions from their dependency structures. This technique has been widely applied to healthcare datasets to simulate linear and non-linear relationships. For example, [88] demonstrated the ability of copulas to capture complex correlations, while [89] employed t-Copulas [90] to generate synthetic datasets that preserved both univariate and multivariate dependencies. Despite their versatility, copulas face computational challenges and are often constrained by simplistic privacy methodologies.
- **Monte Carlo Simulation:** Monte Carlo methods, including Markov Chain Monte Carlo (MCMC), generate synthetic data by iterative sampling from estimated distributions. Gibbs sampling, a common MCMC approach, has been employed in [91], [92] to approximate the joint distribution of high-dimensional medical data. The resulting datasets strongly aligned with the original data in predictive modeling and correlational analyses. However, these methods are largely confined to discrete features and partially synthetic datasets, limiting their applicability to broader healthcare contexts.

- **BNs:** BNs excel in modeling probabilistic relationships among variables, making them a popular choice for healthcare applications. PrivBayes [93] represents a notable effort to synthesize differentially private data using BNs. This approach efficiently handles high-dimensional datasets by generating noisy marginal distributions and constructing synthetic samples. However, its reliance on low-dimensional approximations can result in inaccuracies. In another study, Kaur et al. [94] used an Acyclic Bayesian Network (ABN) to generate synthetic records, achieving high realism scores across uni- and multivariate metrics. Nonetheless, limitations in modeling temporal dependencies and continuous data were observed. These studies highlight the potential of BNs to capture rare patterns in healthcare data while addressing some of their computational and representational challenges.
- **Kernel Density Estimation (KDE):** KDE is a flexible, non-parametric approach to estimating the distribution of a population from limited samples. Platforms like MDClone[10] [95] have employed KDE to generate high-similarity synthetic health data, with applications ranging from sepsis prediction to geospatial analysis. Studies using the U.S. National COVID Cohort Collaborative dataset demonstrated strong alignment between KDE-generated synthetic data and real datasets in clinical and non-clinical tasks [96]. While KDE offers significant advantages for health data modeling, it can be computationally expensive for large-scale datasets and often lacks robust privacy mechanisms.

Two important oversampling techniques have also played a significant role in classical SDG. Synthetic Minority Oversampling Technique (SMOTE) [97], originally developed to address class imbalance in datasets, generates synthetic samples for underrepresented classes by interpolating between existing instances. While primarily designed for classification tasks, SMOTE has been adapted for healthcare applications such as disease prediction and risk stratification. However, its focus on minority classes limits its general utility for comprehensive dataset synthesis. Adaptive Synthetic Sampling (ADASYN) [98] builds on SMOTE by dynamically generating synthetic samples in regions where the minority class is underrepresented, focusing on harder-to-learn examples. This adaptive approach improves model performance on imbalanced datasets, making it particularly useful for healthcare scenarios where rare disease cases often lack sufficient representation.

Statistical methods have been applied widely in healthcare, addressing challenges such as privacy preservation, data augmentation, bias mitigation, and clinical simulation. By generating synthetic datasets that comply with regulations like GDPR and HIPAA, these methods enable data sharing and collaboration while protecting sensitive information. They have also been instrumental in expanding small datasets to improve predictive model performance, correcting demographic imbalances to enhance fairness in AI solutions, and supporting simulations for drug trials and treatment planning.

Despite their significant contributions, these methods face several challenges. Many statistical models rely on parametric assumptions, which may not align with the complexities of real-world healthcare data, limiting their ability to capture non-linear and heterogeneous

[10]Source: MDClone Platform for Synthetic Medical Data Generation (Accessed December 8th, 2024)

relationships. Scalability is another critical issue, as these methods often struggle with high-dimensional datasets common in healthcare, where variables can number in the hundreds or thousands. Privacy challenges persist, with early anonymization techniques and synthetic models sometimes failing to meet rigorous standards, particularly in datasets containing rare or unique cases. Moreover, synthetic data generated by classical methods may lack the diversity and similarity required for robust AI model training, particularly when dealing with datasets exhibiting intricate dependencies.

In conclusion, classical statistical methods have laid the groundwork for SDG in healthcare. Their generalist approach makes them broadly applicable across various data types, but they lack the specialization necessary for addressing the unique challenges of tabular data. While they remain valuable for privacy preservation and initial data augmentation, their limitations highlight the need for advanced methodologies tailored to structured data, which will be explored in subsequent sections.

2.3.3 Machine Learning Approaches

Significant advancements have been made in generating synthetic tabular data by applying DL methodologies, particularly GANs, VAEs, and, more recently, diffusion models. Each approach brings unique advantages and challenges to generating realistic and utility-preserving synthetic data. Below, we discuss these categories in detail, providing examples and insights into their development and applications.

Decision Trees

Decision Trees (DTs) are widely recognized for their simplicity, interpretability, and adaptability, making them a valuable tool in SDG. Their human-readable structure aligns with the increasing demand for XAI, particularly in sensitive domains like healthcare, where interpretable models can build trust and provide actionable insights. This interpretability enhances model transparency and facilitates a deeper understanding of data patterns, ultimately enabling iterative improvements in data synthesis methods.

Randomized Decision Trees (RDTs) have been employed in SDG to create synthetic datasets with a strong resemblance to real data. For instance, [99] used RDTs to generate synthetic data that performed well in classification and regression tasks. The intrinsic Differential Privacy (DP) mechanisms in RDTs ensure minimal risk of sensitive data leakage while preserving utility.

Sequential decision tree-based approaches enhance data realism by conditioning the generation of each feature on previously synthesized ones. [100] demonstrated that the order of feature generation significantly impacts the quality of synthetic data. Their method produced high realism scores and utility, validated through predictive tasks like hospital readmission and treatment arm predictions. This approach was later extended to generate synthetic COVID-19 datasets, confirming its utility as a privacy-preserving alternative for sensitive data [101].

Decision trees are central to non-parametric SDG models, which rely on empirical data distributions instead of predefined assumptions. Their flexibility makes them ideal for

handling heterogeneous data patterns, enabling SDG that mirrors real-world complexities. Recent studies, such as [102] and [103], have highlighted decision tree-based methods as scalable and robust solutions for diverse datasets.

Decision trees offer a robust, interpretable, and privacy-preserving approach to SDG. Their adaptability to complex data structures and ability to generate high-quality synthetic datasets position them as a vital tool in SDG, particularly in domains requiring utility and privacy, such as healthcare. Continued advancements in decision tree-based methods promise to enhance their applicability and effectiveness in addressing real-world challenges.

Generative Adversarial Networks

GANs [32], introduced in 2014, have revolutionized SDG by providing a framework for implicit modeling of complex multidimensional distributions. GANs consist of two NN: a generator, which creates synthetic data samples, and a discriminator, which evaluates whether a sample is real or synthetic. These networks engage in a zero-sum game, wherein the generator aims to fool the discriminator, and the discriminator strives to distinguish real samples from fake ones. Over successive iterations, this adversarial process enables the generator to learn the data distribution and produce realistic synthetic samples [104]. GANs have demonstrated outstanding capabilities across various domains, including image synthesis, medical data generation, and tabular data modeling. However, challenges such as mode collapse, non-convergence, and vanishing gradients remain critical areas of ongoing research [105].

The initial application of GANs for medical data synthesis was marked by the advent of MedGAN in 2017 [106]. MedGAN extended the vanilla GAN architecture by incorporating an Autoencoder (AE) to handle better discrete data—a pivotal feature in EHRs. Despite its groundbreaking nature, MedGAN suffered from several limitations, including its inability to support temporal data and mixed feature types. Subsequent studies, such as MedBGAN [107] and MedWGAN [108], sought to address these issues by leveraging boundary-seeking GANs and Wasserstein GANs (WGANs) with gradient penalties. These variants demonstrated improved performance across realism metrics but continued to rely on inherited privacy guarantees without explicit empirical validation.

Tabular datasets pose unique challenges for GANs due to their mixed data types, imbalanced class distributions, and complex inter-feature dependencies. CTGAN, introduced in 2019 [109], is one of the most significant advancements in tabular data synthesis. This model employs conditional generation to handle the multimodal and imbalanced nature of tabular datasets, effectively capturing both discrete and continuous distributions. Its successor, CTAB-GAN [110], models mixed data types and effectively generates imbalanced categorical variables and continuous variables with complex distributions.

Another notable model is TableGAN [111], which enhances vanilla GANs by introducing classification and information loss terms to improve semantic integrity and feature-wise similarity between real and synthetic data. However, while TableGAN demonstrated comparable utility for privacy-preserving data sharing, it lacks robust support for categorical feature imbalances. Researchers have also explored dependency-preserving models like HGAN [112],

which penalizes generators for producing synthetic samples that violate standard constraints (e.g., gender-specific disease codes), thereby achieving higher record-level similarity.

Privacy has become a central concern in SDG, particularly for sensitive domains such as healthcare. DP mechanisms have been increasingly integrated into GAN architectures to mitigate the risk of sensitive data leakage. For example, PATE-GAN [113] combines GANs with the Private Aggregation of Teacher Ensembles (PATE) framework [114], ensuring DP by introducing noise during the discriminator training phase. Similarly, DP-CTGAN [115] extends CTGAN with DP guarantees, showing superior utility for downstream ML tasks. Privacy-preserving models such as ADS-GAN [116] incorporate identity masking and additional loss constraints to minimize re-identification risks while maintaining the structural realism of synthetic data.

While many GAN-based models excel in generating static tabular data, synthesizing temporal and sequential data has proven more complex. CorGAN [117], for instance, employs CNNs to better capture temporal dependencies in longitudinal medical records. Similarly, SC-GAN [118] introduces dual-generator components to model patient state-medication interactions, enabling the generation of realistic time-series data. Despite these advancements, maintaining the temporal coherence of synthetic records remains an open research challenge.

Several innovations have been introduced to improve the stability and diversity of GAN training. For example, PacGAN [119] addresses mode collapse by utilizing pairs of samples in the discriminator, resulting in more diverse outputs. WGAN-GP [120] introduces gradient penalties to enhance convergence and prevent vanishing gradients. These innovations have significantly expanded the applicability of GANs to complex datasets, including high-dimensional medical and tabular data.

GANs have emerged as a powerful tool for SDG, offering flexibility and scalability across diverse data modalities. GAN-based models like MedGAN, CTGAN, and PATE-GAN have laid the foundation for privacy-preserving and utility-driven data synthesis in healthcare. However, challenges such as temporal dependency modeling, privacy-risk quantification, and categorical imbalance remain areas of active exploration. Future research should focus on hybrid architectures that integrate GANs with other GMs, such as VAEs, to address these limitations and further enhance the similarity, privacy, and utility of synthetic datasets.

Variational Autoencoders

VAEs are a powerful tool in SDG, leveraging a probabilistic framework to model data distributions. Unlike traditional deterministic AEs [121], VAEs learn a latent representation of data by imposing a probabilistic structure, enabling the generation of new, statistically similar data samples. Their adaptability to mixed data types and incomplete datasets makes them well-suited for structured tabular data. A mathematical definition and detailed explanation of the VAE model can be found in Appendix A.

Various adaptations of VAEs have been developed to address the unique challenges posed by tabular data, which often includes a combination of numerical, categorical, and missing data. For instance, TVAE [109] has been specifically designed for tabular data, adeptly handling mixed data types while capturing statistical relationships within the data to generate

synthetic datasets that align closely with real data distributions. Similarly, VAEM [122] focuses on heterogeneous and incomplete data, integrating solutions for missing values while maintaining high-quality generative performance across diverse feature types. Another model, HI-VAE [123], addresses inconsistencies introduced by missing data by balancing marginal distributions, thereby enhancing the realism of synthetic datasets.

Expanding on general-purpose applications, specialized VAEs have been designed to tackle specific challenges. RTVAE [124] enhances robustness by employing β-divergence, making it particularly effective for tabular data with mixed categorical and continuous features. Similarly, TabVAE [125] addresses the issue of over-pruning and posterior collapse, which often arise in high-dimensional latent spaces, by introducing a novel sampling technique for categorical variables. This refinement ensures that multi-class categorical data are modeled accurately, even in complex tabular datasets. On the other hand, OVAE [126] integrates differentiable oblivious decision trees with VAEs, introducing a strong inductive bias for tabular data. Designed to handle privacy concerns, OVAE generates high-similarity synthetic data suitable for sensitive domains such as healthcare and finance while maintaining data confidentiality.

VAEs are also well-suited for time-series applications, as demonstrated by the development of CR-VAE [127]. This causal recurrent VAE is designed to generate time-series data while learning Granger causality graphs. With a multi-head decoder, where each head corresponds to a dimension of the input data, CR-VAE has shown superior performance in medical applications such as electroencephalography and functional MRI analysis. Despite its robust performance, the reliance of the model on an isotropic Gaussian assumption for latent factors may limit its generative capabilities, suggesting potential areas for future improvement.

Innovative architectures have further expanded the utility of VAEs. TS-VAE [128] uses a two-stage training paradigm to refine latent space representations, ensuring better generative performance for continuous data. VampPrior VAE [129], with its mixture of Gaussians as a prior for latent variables, increases the flexibility and adaptability of the model, making it well-suited for intricate datasets such as those found in genomics and medical imaging.

VAEs have demonstrated exceptional adaptability in healthcare and finance, where structured data are sensitive and often include varying feature types and missing values. VAEs have been utilized in healthcare to generate synthetic EHRs that retain the statistical and structural properties of real patient data while safeguarding privacy. VAEs have been employed in the financial sector to generate synthetic datasets for privacy-preserving analytics and predictive modeling [130].

In conclusion, VAEs provide a robust and flexible approach to SDG. Their ability to handle mixed data types and incomplete datasets, coupled with advancements in their architecture, positions them as an essential tool for generating high-quality synthetic tabular data. Continued innovation in VAE frameworks, such as incorporating more sophisticated priors or multi-stage training processes, holds significant promise for further improving their utility in complex real-world applications.

Other Techniques and Innovations

SDG frameworks have advanced significantly, integrating diverse methodologies and models to address data scarcity, privacy, and the demand for high-similarity datasets in sensitive fields like healthcare. Among the latest innovations, diffusion models have emerged as a promising approach for generating synthetic data and addressing data imputation tasks, particularly for mixed-type tabular data.

Diffusion models [131] leverage a two-step process that begins with degrading a complex data distribution into a simple known distribution, such as Gaussian noise, and restoring the original structure through a reverse generative process. This approach has been adapted for tabular data in models such as the following. STaSy [132] utilizes a score-based generative framework, incorporating self-paced learning and fine-tuning strategies to enhance the diversity and quality of generated data by stabilizing the denoising training process. Similarly, CoDi [133] employs dual diffusion models to handle mixed data types effectively; one model processes continuous features, while the other manages categorical variables. This approach ensures that interactions between feature types are captured using a conditioning mechanism. TabSyn [134] combines VAEs and diffusion models by encoding mixed-type data into a continuous latent space and then applying diffusion in this latent space. By employing transformer architectures during the encoding process, TabSyn captures intricate feature relationships, which are leveraged during the generative phase. TabDDPM [135] adapts Gaussian diffusion for continuous features and multinomial diffusion for categorical ones, employing preprocessing steps such as scaling and one-hot encoding to harmonize data types. Its hybrid optimization objective combines mean squared error and KL divergence, making it effective for classification and regression tasks. Meanwhile, TabCSDI [136] extends diffusion technologies initially designed for time-series data to handle tabular datasets. Through feature tokenization and transformer-based learning, missing data are effectively imputed while maintaining the statistical properties of the dataset. Building on these principles, MTabGen [137] introduces a dynamic masking mechanism and an encoder-decoder transformer architecture. By masking random features during training and reconstructing them in the denoising phase, MTabGen excels in data imputation and conditioned synthetic data generation. This adaptability underscores its potential for tackling diverse tabular data challenges with a single unified model.

LLMs represent a promising advancement in STDG, leveraging their generative capabilities to produce high-quality synthetic datasets. These models reformat tabular data into natural language or simplified sequences, allowing the application of pre-trained architectures such as GPT-2 [138]. Approaches like GReaT [139] and TabuLa [140] demonstrate the effectiveness of LLMs in generating tabular data by sequentially predicting feature values as tokens. Despite their strength in capturing contextual and semantic relationships, LLMs struggle to fully exploit the relational structure inherent in tabular data. Furthermore, they pose privacy concerns, as they may inadvertently memorize and expose sensitive information from training data. Addressing these privacy risks through robust protection mechanisms is crucial to realizing the full potential of LLMs in STDG.

Beyond diffusion models and LLMs, other frameworks have also demonstrated significant contributions to SDG. SynSys [141], for example, employs hidden Markov models and regression techniques to generate synthetic data for healthcare, specifically in smart home environments.

This framework excels at modeling intricate sequences of human behavior, sensor events, and their corresponding timestamps. However, its scalability to more diverse real-world scenarios remains an area for further exploration. The TF-GAN[11] library, built on TensorFlow, provides a flexible and robust framework for developing advanced GMs using GANs. Its versatility makes it suitable for creating synthetic data across various domains. However, the steep learning curve and significant computational resources required for its implementation limit its accessibility to organizations with constrained resources or technical expertise. In contrast, Gretel Synthesis[12] offers a more accessible solution with a focus on user-friendly workflows and privacy preservation. This platform simplifies the process of synthetic data generation, making it particularly attractive for non-technical users. However, a trade-off exists between privacy and data similarity, as the generated datasets may lack the complexity required for nuanced applications. Another significant framework is DataSynthesizer [142], which employs differential privacy techniques to protect sensitive information during synthetic data generation. While its focus on privacy is commendable, the resulting datasets can sometimes lack the variability and richness necessary for downstream analysis, limiting its effectiveness in high-similarity applications.

These advancements illustrate the versatility and innovation of SDG frameworks. Whether through GAN or VAE-based approaches or cutting-edge diffusion models, these tools continue to evolve, addressing the nuanced demands of SDG while balancing privacy, scalability, and data similarity. Future research should aim to optimize these methods further to enhance their applicability across diverse domains, particularly in healthcare and other sensitive fields.

Machine Learning State-of-the-Art Models Summary

The models in Table 2.3 concisely summarize the latest innovation, diffusion models and the SDG approaches discussed in previous sections, categorized into Decision Trees, GANs, VAEs, and Diffusion Models. Each category addresses specific challenges in SDG, such as handling mixed data types, preserving privacy, and ensuring data similarity. Decision trees focus on interpretability, GANs excel in modeling complex distributions, and VAEs offer probabilistic frameworks for generating heterogeneous and time-series data. As the latest innovation, diffusion models deliver robust solutions for mixed data synthesis and imputation, exemplified by TabDDPM and MTabGen. These models collectively showcase the advancements in SDG methodologies, highlighting their potential for privacy-compliant and realistic data generation in healthcare and beyond.

[11]Source: TF-GAN Library for TensorFlow 2.0 (Accessed December 13th, 2024)
[12]Source: Gretel: Synthetic Data Generation (Accessed December 13th, 2024)

Category	Model	Key Features
Decision Trees	RDTs [99]	Privacy-preserving via DP mechanisms; high utility for classification and regression tasks.
	Sequential DTs [100]	Enhanced realism through sequential feature generation; validated on tasks like hospital readmission and COVID-19 dataset generation.
	Non-parametric SDG Trees [102], [103]	Empirical data distribution modeling; scalable and robust for diverse datasets.
GANs	MedGAN [106]	Handles discrete data in EHRs; limited support for mixed or temporal data types.
	CTGAN [109]	Conditional generation for mixed and imbalanced tabular data; captures discrete and continuous feature distributions effectively.
	PATE-GAN [113]	Combines GANs with PATE to ensure DP.
	CorGAN [117]	Focuses on temporal dependencies in longitudinal medical records using CNNs.
VAEs	TVAE [109]	Designed for tabular data; handles mixed data types while preserving statistical relationships.
	VAEM [122]	Focused on heterogeneous and incomplete datasets; high-quality generative performance across feature types.
	RTVAE [124]	Employs β-divergence for robustness; suitable for tabular data with mixed features.
	TabVAE [125]	Addresses over-pruning and posterior collapse in high-dimensional spaces; introduces novel sampling techniques for categorical variables.
	OVAE [126]	Integrates differentiable oblivious DTs for privacy-preserving SDG.
	CR-VAE [127]	Designed for time-series data; learns Granger [143] causality graphs; multi-head decoder for handling multivariate data.
Diffusion Models	STaSy [132]	Score-based generative approach with self-paced learning; stabilizes denoising training for high-quality outputs.
	CoDi [133]	Manages mixed data types using separate diffusion processes for continuous and categorical features.
	TabSyn [134]	Uses latent diffusion with transformer encoding for feature relationships; optimized for tabular data.
	TabDDPM [135]	Adapts Gaussian and multinomial diffusion for continuous and categorical data; hybrid loss optimization.
	TabCSDI [136]	Adapts time-series diffusion for tabular data; uses feature tokenization and transformer-based learning.
	MTabGen [137]	Dynamic masking with encoder-decoder transformer; excels in conditioned synthetic data generation and imputation.
LLMs	GReaT [139]	Reformats tabular data as text for LLM-based generation; sequential feature prediction for structured datasets.
	TabuLa [140]	Leverages pre-trained LLMs for tabular SDG; contextualized feature synthesis for improved realism.

Table 2.3: Summary of SDG state-of-the-art models. SDG models categorized by type and distinctive features highlighted.

2.3.4 Validation Techniques

Following the impressive advances in generative AI for data generation, a critical aspect emerges: ensuring the quality and effectiveness of synthetic data. Replicating the statistical properties of the source data, as well as accuracy, consistency, and domain-specific relevance, is essential. This challenge becomes significantly pronounced as GMs gain traction within engineering design communities. The need for robust and standardized validation metrics becomes paramount. The current literature highlights a significant gap in this area. Although there are standardized evaluation metrics for the generation of synthetic images [144], [145], and the generation of text [146], [147], measuring the quality of synthetic tabular data presents unique challenges. Unlike image data, qualitative evaluation through visual inspection is not feasible. Additionally, relying solely on expert insight can be highly inefficient. The existing landscape for tabular data validation often focuses primarily on the efficacy or utility in ML tasks [148], [149]. However, a consistent approach to assessing the similarity between synthetic and real tabular data remains sparse. Studies employ a diverse set of metrics, including pairwise correlation difference [150], support coverage [151], likelihood fitness [109], alongside various statistic values [152], [153]. Reviews such as [8], [76] attempt to categorize these techniques. Still, a genuinely standardized approach or a single metric that captures the full spectrum of statistical information from a distribution remains challenging.

Generally, synthetic tabular data validation techniques can be divided into three categories: resemblance, utility, and privacy. Resemblance focuses on how well the synthetic data mirrors the statistical characteristics of the original dataset, utility evaluates the applicability of the synthetic data to specific tasks, and privacy measures the risk of sensitive information leakage. This thesis concentrates on resemblance and utility validation, leaving privacy validation—due to its broad scope and complexity—for future exploration. As noted in the literature [76], the degree to which synthetic data achieves high resemblance and utility determines its capacity to substitute real data in practical applications.

Figure 2.8 provides a schematic overview of the various validation techniques, categorizing them into resemblance and utility metrics. This visual summary highlights the distinct methods used to assess synthetic data quality and their alignment with specific validation goals.

Resemblance Metrics

Resemblance validation ensures that synthetic data faithfully reproduces the statistical and structural properties of the original dataset. This step is fundamental to establishing the realism and practical utility of synthetic data. The resemblance is typically assessed at two levels: univariate resemblance, which focuses on individual feature distributions, and multivariate resemblance, which examines the relationships and dependencies between features. This distinction is critical because even when synthetic data demonstrates high univariate resemblance, it may fail to preserve the intricate inter-feature dependencies essential for many real-world applications, particularly in fields like healthcare, where correlations and covariances between variables are paramount.

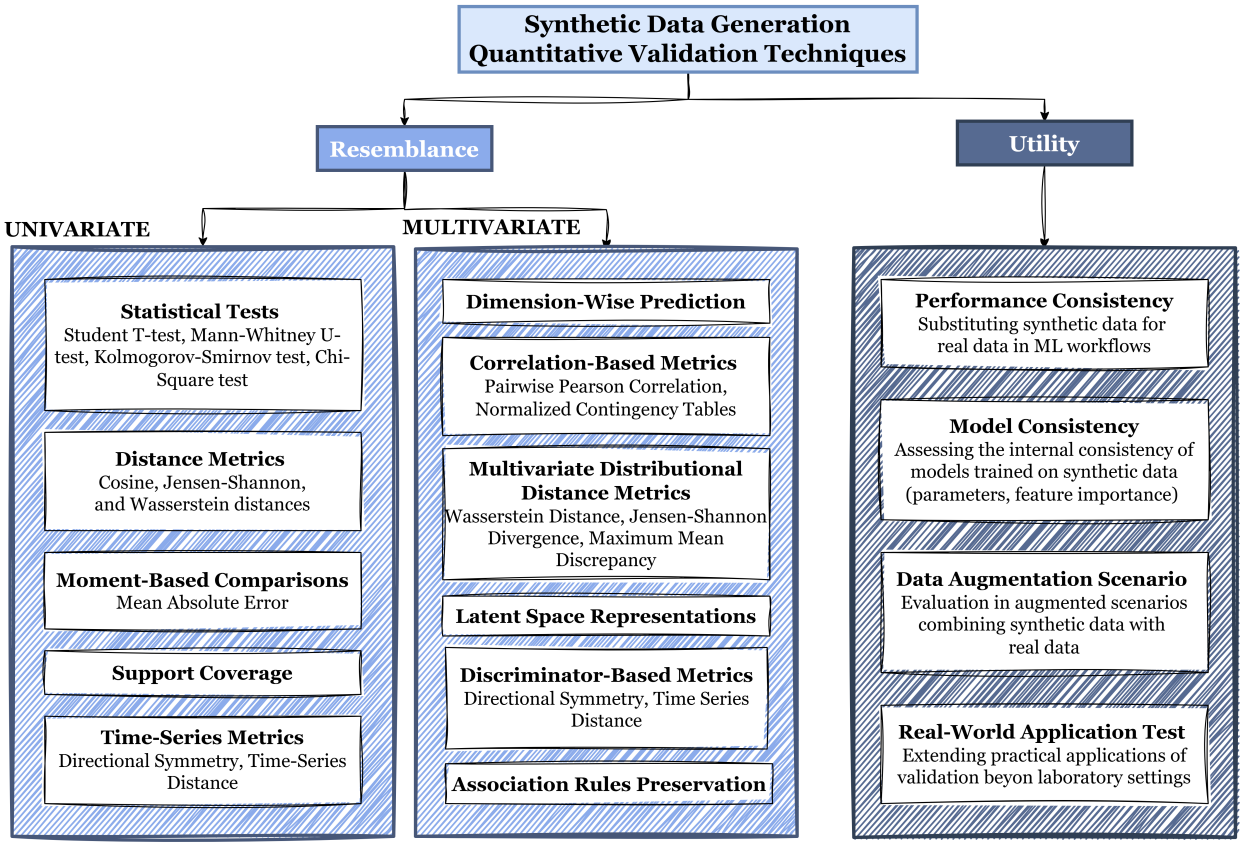


Figure 2.8: Overview of quantitative validation techniques for SDG, categorizing methods into resemblance and utility metrics. Resemblance metrics assess how closely synthetic data aligns with real data, focusing on univariate and multivariate evaluations. Utility metrics evaluate the practical effectiveness of the synthetic data in ML and real-world applications, emphasizing performance consistency, model alignment, augmentation scenarios, and broader applications.

→ ***Univariate Metrics:***

Univariate metrics evaluate the similarity between individual feature distributions in real and synthetic datasets. This initial validation step ensures that each feature in the synthetic dataset retains the statistical properties of its counterpart in the real dataset.

1. **Statistical Tests:** Statistical tests are widely employed to compare the properties of individual features in real and synthetic data. These tests determine whether the statistical characteristics of synthetic data align with those of real data. Commonly applied tests include:
   - **Student T-test:** Compares the means of real and synthetic attributes to evaluate statistical equivalence [116].
   - **Mann-Whitney U-test:** Assesses whether real and synthetic features originate from the same population distribution [95].
   - **Kolmogorov-Smirnov (KS) test:** Compares the cumulative distributions of

real and synthetic features to detect significant deviations [108], [154], [155].

For categorical features, the Chi-Square test $\mathcal{X}^2$ evaluates the independence of attributes between real and synthetic data [95], [156]. The null hypothesis assumes no statistical relationship between real and synthetic features. High p-values suggest that the synthetic data adequately preserves the properties of the real data.

2. **Distance Metrics:** Distance metrics quantify the divergence between real and synthetic features. Lower values indicate a closer resemblance [110], [116], [151], [157]. Being x_r the real data feature and x_g the synthetic data feature to be compared, common metrics include:

   - **Cosine Distance:** Measures angular similarity between feature vectors of real and synthetic data. Values near zero indicate high resemblance.

   $$CD(x_r, x_g) = 1 - \frac{\sum_{i=1}^{N} x_{r_i} \cdot x_{s_i}}{\sqrt{\sum_{i=1}^{N} x_{r_i}^2} \cdot \sqrt{\sum_{i=1}^{N} x_{s_i}^2}}. \tag{2.28}$$

   - **Jensen-Shannon Distance (JSD):** Assesses the similarity between probability distributions. Values below 0.1 indicate strong alignment. JSD is the square root of the Jensen-Shannon Divergence ($D_{\mathbb{JS}}$). This divergence measures the similarity between two probability distributions:

   $$D_{\mathbb{JS}}(p, q) = \sqrt{\frac{D_{\mathbb{KL}}(p||m) + D_{\mathbb{KL}}(q||m)}{2}}, \tag{2.29}$$

   where p is the probability distribution of the real data feature, q is the probability distribution of the synthetic data feature, m is the point-wise mean of p and q and $D_{\mathbb{KL}}$ is the Kullback-Leibler Divergence defined as:

   $$D_{\mathbb{KL}}(p||q) = \sum_{x \in X} p(x) \log \frac{p(x)}{q(x)}. \tag{2.30}$$

   Both $D_{\mathbb{KL}}$ and $D_{\mathbb{JS}}$ play a significant role in evaluating the resemblance of probability distributions and are of particular interest in this thesis. Their theoretical properties and practical applications will be explored in greater depth in later sections as they form a key component of the proposed methodology.

   - **Wasserstein Distance:** Reflects the minimum cost to transform the real data distribution into the synthetic distribution. Thresholds around 0.3 are typically used to ensure acceptable resemblance.

   $$WD(x_r, x_g) = \int_{-\infty}^{+\infty} |R_{CDF} - S_{CDF}|, \tag{2.31}$$

   where R_{CDF} and S_{CDF} are the CDFs of the real and synthetic data features, respectively.

3. **Moment-Based Comparisons [107], [116]:** These analyses compare statistical moments (e.g., mean, variance, skewness) of features in real and synthetic datasets.

Higher-order moments are examined for features with complex distributions. Metrics such as Mean Absolute Error (MAE) are used to quantify differences between the moments of real and synthetic features [149], [150].

4. **Support Coverage [151]:** This metric evaluates the overlap between real and synthetic data regarding categorical or discrete value ranges. A specific variant, rare feature occurrence metrics, ensures that rare categories in real data are adequately represented in synthetic data. This is particularly critical in datasets with imbalanced classes or sparse categories.

5. **Time-Series Metrics [158]:** For temporal data, specialized metrics such as Directional Symmetry (DS) and Short Time-Series Distance (STS) are used. DS examines whether the directional trends (e.g., upward or downward movements) in real and synthetic data are harmonized. STS measures the average distance between trends in real and synthetic data to assess temporal alignment.

While univariate metrics provide a foundational assessment of the statistical similarity of synthetic data, they cannot capture dependencies between features. For instance, high univariate resemblance may overlook critical co-occurrences or impossible feature combinations (e.g., demographic or clinical anomalies). Studies have shown that high univariate resemblance does not necessarily translate to sufficient multivariate similarity [151]. Consequently, multivariate metrics are indispensable for ensuring the structural realism and functional utility of synthetic datasets.

→ ***Multivariate Metrics:***
Multivariate metrics are essential for evaluating how well synthetic data captures the inter-feature relationships and dependencies inherent in real datasets. These metrics are critical for ensuring that synthetic data can support complex tasks where correlations, covariances, and higher-order interactions between variables play a significant role. The following approaches are commonly employed:

1. **Dimension-Wise Prediction [106]–[108], [151]:** This method iteratively selects each feature in the dataset as a target variable and predicts it using the remaining features as predictors. High prediction accuracy indicates that the synthetic data effectively preserves the inter-feature dependencies in the real dataset.

2. **Correlation-Based Metrics:** These metrics assess the preservation of pairwise relationships between features. For numerical attributes, Pairwise Pearson Correlation (PPC) matrices compare correlation coefficients between feature pairs in real and synthetic data [94], [159], [160]. For categorical variables, normalized contingency tables are used to analyze co-occurrence patterns [149]. Visual tools like heatmaps often complement these metrics, providing intuitive comparisons of the correlation structures between real and synthetic datasets.

3. **Multivariate distributional distance metrics:** Metrics such as the Wasserstein Distance and $D_{\mathbb{JS}}$ quantify the alignment between the overall multivariate distributions of real and synthetic data [116]. These measures evaluate the global structural similarity of the datasets, offering insights into their alignment at a distributional level.

Additionally, Maximum Mean Discrepancy (MMD) [161] provides a kernel-based approach to compare probability distributions without making parametric assumptions. Unlike divergence-based metrics like $D_{\mathbb{KL}}$ and $D_{\mathbb{JS}}$, MMD leverages reproducing kernel Hilbert spaces (RKHS) to assess the difference between feature distributions [162]. Given real and synthetic datasets, X_r and X_g, sampled from the respective distributions p and q, MMD is computed as:

$$MMD^2(p,q) = \mathbb{E}_p[\hat{k}(X_r, X_r)] - 2\mathbb{E}_{p,q}[\hat{k}(X_r, X_g)] + \mathbb{E}_{\|}[\hat{k}(X_g, X_g)], \tag{2.32}$$

where $\hat{k}(x,y)$ is a kernel function, commonly the Gaussian kernel. Lower MMD values indicate a higher resemblance between the real and synthetic distributions, making it a valuable metric for validating the fidelity of synthetic data.

4. **Latent space representations [112], [163]:** Dimensionality reduction techniques like Principal Component Analysis (PCA) and Non-Metric Multi-Dimensional Scaling (NMDS) project high-dimensional data into lower-dimensional latent spaces. Comparing these representations can reveal whether synthetic data captures the core correlational structures of the original dataset. Additional measures, such as log-cluster coherence [151], assess the similarity in clustering patterns between real and synthetic data within these latent spaces, further validating the preservation of structural relationships.

5. **Discriminator-based metrics [100], [101], [164]:** This approach combines real and synthetic datasets into a single labeled dataset, with labels indicating the origin of each record (e.g., 0 for real data, 1 for synthetic data). The combined dataset is split into training and testing subsets. ML classifiers such as Random Forest (RF), K-Nearest Neighbors (KNN), DT, SVMs, and Multilayer Perceptron (MLP) are trained to differentiate between real and synthetic records. Classifier performance close to random (e.g., accuracy near 50%) suggests that synthetic data are indistinguishable from real data, reflecting strong multivariate resemblance.

6. **Association rules preservation [108], [112]:** This metric evaluates whether the association rules observed in real data, such as feature co-occurrences or disease comorbidities, are maintained in synthetic data. For domain-specific tasks, domain rule preservation ensures that known feature associations or dependencies—critical for the domain—are accurately replicated in the synthetic dataset [112], [165].

By leveraging these multivariate metrics, researchers can comprehensively evaluate the ability of the synthetic data to replicate the intricate interdependencies and relationships in real-world datasets. These metrics provide a more robust validation framework than univariate metrics, emphasizing the structural realism and functional utility of synthetic data for downstream applications.

Utility Metrics

Utility validation focuses on assessing the practical applicability and effectiveness of synthetic data in real-world tasks, particularly in ML and statistical modeling. Unlike resemblance metrics, which evaluate how closely synthetic data replicates the structural and statistical properties of real data, utility metrics emphasize the functional outcomes achieved when

synthetic data are used. This approach underscores the importance of aligning synthetic data with specific application goals. Studies have shown that synthetic data often experience less utility loss in ML tasks and can perform well even if their resemblance to real data are moderate [166].

A primary utility validation method involves substituting synthetic data for real data in ML workflows and analyzing performance. This can be done through setups such as Train on Synthetic, Test on Real (TSTR), and Train on Real, Test on Synthetic (TRTS) [167]. High performance in these configurations—measured by accuracy, F1-score, precision, recall, and ROC-AUC—indicates that the synthetic data effectively captures the patterns and relationships present in real data. Such results demonstrate its viability for practical applications [100], [113], [168], [169].

Another critical aspect of utility validation is assessing the internal consistency of models trained on synthetic data. This includes comparing model parameters, feature importance rankings, or decision boundaries between models trained on synthetic versus real data [100], [170]. Such analyses are particularly valuable in contexts where interpretability is crucial or feature selection plays a significant role in model performance. For instance, response agreement [111], [171] evaluates whether models trained on synthetic and real data produce similar predictions for identical test cases, providing insights into the ability of the synthetic data to generalize and support real-world decision-making processes.

Synthetic data can also be evaluated in augmentation scenarios, combining it with real data to address issues such as data scarcity or imbalance. The effectiveness of synthetic data in improving model performance when used for augmentation validates its contribution to enhancing the training process. This is especially significant in domains like healthcare, where data availability may be restricted due to privacy concerns or resource constraints. Successful augmentation demonstrates the capacity of synthetic data to enhance the representativeness and robustness of real datasets [89], [172], [173].

Practical applications of utility validation extend beyond laboratory settings. Synthetic data has been employed in real-world tasks such as clinical decision support [174], replicating findings from previous studies, and even training models for educational purposes [175]. These use cases highlight the ability of synthetic data to generate actionable and meaningful outcomes, affirming its value as a tool for supporting research, training, and operational decision-making.

In summary, utility validation ensures that synthetic data are not merely a statistical mimic of real data but an effective, functional resource capable of supporting diverse applications. By emphasizing performance, consistency, and real-world usability, this validation approach establishes the credibility of synthetic data as both a substitute for and a complement to real-world datasets.

Qualitative Assesment Approach

While much focus has been placed on quantitative evaluations, a complementary qualitative assessment of synthetic data also exists. This form of evaluation involves human judgment to determine the plausibility and validity of synthetic records. It is widely recognized as an

essential step following quantitative analysis to detect potential discrepancies or anomalies within the data. The qualitative assessment process is generally manual and relies heavily on the expertise and experience of the evaluators. This type of evaluation can be divided into two main categories: (1) Expert Review and (2) Visual Inspection.

Expert review entails domain specialists, such as clinicians in the case of healthcare data, manually examining synthetic records. The goal is to assess the clinical plausibility of the data by verifying adherence to logical constraints, acceptable value ranges, feature correlations, and frequency distributions. Many studies underline the importance of this step as a final validation phase to ensure the reliability and diversity of synthetic data. However, while essential for assessing the plausibility of synthetic data, it presents several challenges. It is inherently time-consuming and prone to human error due to its manual nature. Moreover, as datasets grow in size and dimensionality, comprehensive evaluations become increasingly impractical, limiting scalability. The process is also highly subjective, with outcomes varying depending on the expertise of the evaluator, skill level, and commitment. This variability can lead to inconsistencies, particularly when different reviewers analyze the same dataset. These limitations highlight the need for more efficient and standardized approaches to qualitative assessment.

Visual inspection involves comparing real and synthetic data using data visualization techniques to evaluate their alignment and realism. Various visualization methods, such as histograms, correlation heatmaps, and PCA plots, are commonly employed in the literature. Practitioners can tailor visualization techniques depending on the data format and the requirements for assessing realism. Although visualizations are not explicitly recognized as a formal metric for evaluating synthetic data quality, they are invaluable tools for providing a high-level overview of data plausibility. Additionally, they can offer critical insights for selecting appropriate quantitative validation metrics for further analysis. Visual inspection serves as a quick and effective means to gauge the overall quality of synthetic data and often guides deeper, more detailed evaluations.

2.3.5 Challenges and Future Directions

Figure 2.9 provides an overview of the key challenges in STDG, categorizing them into four major areas: data complexity, data scarcity, validation, and privacy and ethical concerns. Each category encompasses specific issues that hinder the development and application of high-quality synthetic datasets. These challenges highlight the multifaceted nature of STDG and underscore the need for continued innovation and standardization in the field.

STDG is a rapidly evolving field driven by the increasing demand for high-quality, privacy-preserving datasets across healthcare, finance, and education domains. Despite notable advancements, several inherent challenges continue to hinder the development, validation, and application of synthetic tabular data. These challenges can be broadly categorized into data complexity, generation quality, validation methodologies, and ethical considerations.

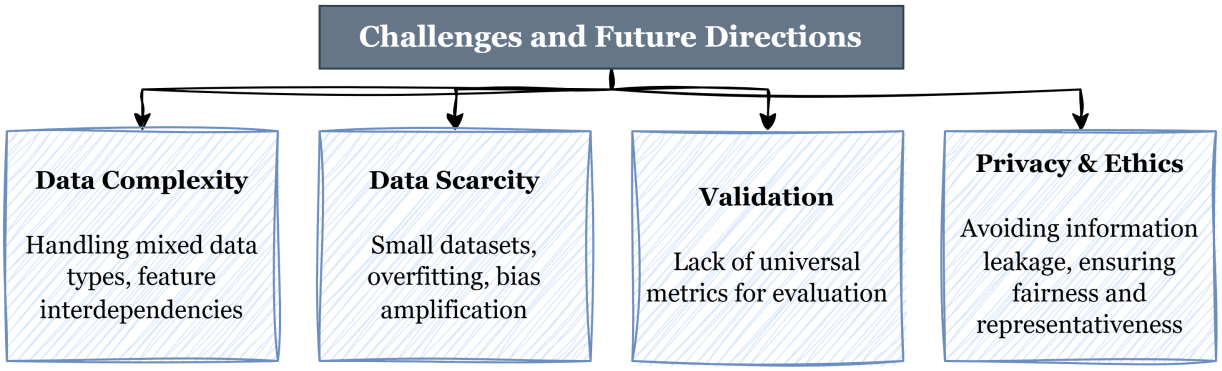


Figure 2.9: Key Challenges in STDG. The diagram summarizes the primary challenges in synthetic tabular data generation, including data complexity, data scarcity, validation, and privacy concerns, as discussed in this section.

Data Complexity and Diversity

Tabular data generation faces significant challenges due to its inherent complexity, diversity, and the need to preserve statistical and structural realism. Unlike other data modalities such as images or text, tabular datasets often combine numerical, categorical, ordinal, and temporal variables, creating a highly heterogeneous structure. This heterogeneity makes it challenging to model and synthesize data accurately reflecting the intricate interdependencies and correlations between features. For example, healthcare datasets often include patient demographics (categorical), clinical measurements (numerical), and treatment timelines (temporal), which together form complex, multi-type relationships that synthetic models must capture.

Another challenge lies in the sparsity and imbalance frequently present in tabular datasets. Rare events or minority classes, such as patients with specific genetic conditions, are often underrepresented, making it difficult for GMs to appropriately synthesize these cases without overfitting. For instance, a healthcare dataset tracking rare diseases may have a handful of positive cases among thousands of records. Failing to generate accurate synthetic representations of such rare occurrences can limit the usability of synthetic data for niche analyses or downstream applications.

Maintaining the statistical and structural realism of synthetic tabular data compounds these challenges. While models may excel at replicating univariate distributions—ensuring alignment for individual features—they often fail to preserve multivariate relationships, such as feature covariances or hierarchical dependencies. For example, while synthetic data might correctly replicate the age distribution of a patient cohort, it may fail to preserve critical correlations, such as the relationship between age, comorbidities, and treatment outcomes. This lack of multivariate similarity can significantly impact the usability of synthetic data for predictive modeling or inferential studies.

Furthermore, handling extreme or rare values adds another layer of complexity. These values are often vital for understanding outliers or edge cases but are challenging to synthesize

without risking overfitting. For instance, a financial dataset with extreme credit scores or healthcare data with anomalous clinical measurements can distort model training if not adequately represented. GMs must balance capturing these outliers without disproportionately weighting their influence.

Addressing these challenges requires GMs that can effectively learn from and replicate the intricate structures of tabular data while maintaining similarity to univariate and multivariate properties. Advanced techniques, such as incorporating hierarchical Bayesian methods or feature-specific embeddings, have shown promise in tackling these issues.

Handling Data Scarcity

One of the most significant challenges in STDG is the inherent scarcity of data in many real-world scenarios, particularly in specialized fields like healthcare, finance, and rare disease research. Unlike image and text datasets, which often comprise millions of samples for training GMs, tabular datasets are frequently limited in size due to privacy concerns, the cost of data collection, or the rarity of specific conditions.

For example, in healthcare applications, patient datasets for diseases such as multiple myeloma or rare genetic disorders often consist of only a few hundred or thousand records. This contrasts the datasets typically used to evaluate state-of-the-art GMs like MedGAN and CTGAN, which include 20,000 to 500,000 samples [106], [109]. Such large datasets are unrealistic for many medical contexts, where ethical considerations, institutional barriers, and patient privacy regulations severely limit the amount of available data.

The scarcity of data introduces several issues for STDG:

1. **Overfitting and Limited Generalization:** GMs trained on small datasets are prone to overfitting, capturing noise and outliers instead of learning generalizable patterns. This limits the utility of synthetic data for downstream tasks such as ML model training or predictive analytics.
2. **Bias Amplification:** Small datasets often do not represent the broader population, resulting in synthetic data exacerbating biases in the original data. For instance, underrepresented demographic groups in the real data may be further marginalized in the generated data, leading to ethical concerns and reduced model performance.
3. **Inadequate Diversity:** Data scarcity often leads to GMs producing synthetic datasets with limited variability. This can undermine their ability to represent the full range of scenarios required for robust applications, particularly in fields like healthcare, where patient heterogeneity is critical.
4. **High Dimensionality with Low Sample Size:** Tabular data often features high dimensionality, with hundreds of variables, while datasets may only contain a few hundred records. This imbalance between feature size and data volume complicates the training of GMs, as they struggle to learn meaningful relationships without sufficient data.

Overcoming data scarcity in STDG remains an active area of research. It requires designing

models and techniques that are robust to small datasets and capable of generating high-quality synthetic data that faithfully represent the complexities of the original data. Addressing these challenges is vital for the broader adoption of STDG in domains like healthcare, where high-quality synthetic data can drive innovation, support decision-making, and enhance research in resource-constrained settings.

Evaluation and Validation

Validation of synthetic tabular data are one of the most pressing challenges, as there is no universal metric to assess its quality comprehensively. Unlike image or text data, where qualitative evaluation and standardized benchmarks are available, tabular data requires rigorous quantitative and qualitative assessment to ensure utility and resemblance.

- **Quantitative Challenges:** Resemblance metrics assess statistical alignment but often fail to capture the practical utility of synthetic data in downstream tasks. Conversely, utility metrics, such as TSTR and TRTS, are task-specific and lack generalizability, limiting their applicability across diverse datasets and domains.
- **Qualitative Validation:** While expert evaluation and visual checks can identify discrepancies missed by quantitative metrics, these methods are subjective and impractical for large datasets, making scalable qualitative validation difficult.

Developing a standardized framework integrating resemblance, utility, and privacy metrics is essential to comprehensively evaluate synthetic data, ensuring its reliability and broader applicability across real-world scenarios.

Privacy and Ethical Concerns

SDG becomes particularly complex in domain-specific contexts, such as healthcare, where strict regulatory and contextual constraints govern datasets. Synthetic data in such settings must preserve clinically meaningful relationships, such as comorbidities or treatment responses, while avoiding the creation of unrealistic or invalid patient records., while avoiding creating Ethical considerations compound these challenges, as synthetic data must protect privacy and ensure fairness and representativeness.

While synthetic data aims to address privacy concerns by obfuscating original data, there remains a risk of information leakage. If GMs inadvertently memorize or replicate sensitive records from the original dataset, it undermines the very purpose of SDG. Furthermore, synthetic data must be carefully generated to avoid introducing or amplifying biases inherent in the original data. Failing to do so can lead to skewed analyses, reinforce systemic inequities, and compromise the utility and trustworthiness of downstream applications. Balancing these domain-specific and ethical challenges is critical to ensuring synthetic data are reliable, fair, and privacy-preserving tools for real-world use.

Lack of Standardized Benchmarks

The absence of standardized benchmarks and datasets for STDG makes comparing the performance of different GMs difficult. While some benchmarks exist in the image and NLP

domains, the tabular data space lacks equivalent resources. This limitation hampers the systematic evaluation of GMs and their applicability to real-world scenarios.

Addressing the Challenges

While these challenges present significant hurdles, ongoing research continues to develop innovative solutions. Techniques such as transfer learning, meta-learning, and domain-specific inductive biases have shown promise in addressing data scarcity and heterogeneity. Efforts to standardize validation metrics and establish benchmarks for synthetic tabular data quality will further advance the field. Additionally, integrating privacy-preserving techniques with GMs, such as DP, can enhance privacy guarantees while maintaining data utility. By addressing these challenges, synthetic tabular data can achieve its full potential as a powerful tool for data-driven innovation in various domains.

2.4 State of the Art of Federated Learning in Healthcare

FL has become a groundbreaking framework in ML, addressing essential issues concerning data availability, privacy, and diversity. As highlighted by Google researchers, *"The real-world performance of your ML model depends on the relevance of the data used to train it."*[13] This highlights the need for diverse, high-quality, and representative data to train effective models. However, in domains such as healthcare, the collection and centralization of such data face significant obstacles, including stringent privacy regulations, ethical considerations, and logistical constraints.

Traditional centralized ML approaches involve aggregating data from multiple sources into a single repository for training. While effective in environments where data centralization is feasible, this approach proves unsuitable in healthcare, where data are distributed across hospitals, clinics, and research institutions, each governed by strict privacy laws. Centralization amplifies the risks of data breaches and poses practical limitations in handling non-IID and unbalanced datasets that are common in medical contexts. These limitations necessitate innovative approaches that safeguard privacy while leveraging the diversity inherent in decentralized datasets.

FL addresses these challenges by enabling collaborative training of ML models without centralizing sensitive data [176]–[179]. Instead, local models are trained on distributed datasets, and only model parameters or updates are shared with a central server for aggregation. This distributed framework preserves privacy, ensures compliance with regulations such as HIPAA and GDPR, and allows models to benefit from the diversity of decentralized data. Moreover, the FL architecture integrates seamlessly with advancements in Edge AI, leveraging distributed computing resources for localized AI operations without violating data regulations [180], [181]. The convergence of Edge AI and FL has further enhanced the applicability of FL in healthcare. Using the computational capabilities of edge devices, Edge AI facilitates localized model training and inference, minimizing latency and compliance risks [180]. FL builds on this foundation by enabling collaboration across multiple edge devices or institutions, creating robust and generalizable models incorporating insights from diverse and distributed datasets.

FL is particularly impactful in healthcare, allowing institutions to collaborate while safeguarding patient confidentiality. It has been adopted in tasks ranging from disease prediction to personalized treatment recommendations, proving its efficacy in academic research and real-world scenarios. This decentralized approach also addresses the practical challenges posed by non-IID and unbalanced data distributions—common in healthcare—by enabling diverse entities to contribute to a robust, generalizable model without requiring data pooling.

To provide a comprehensive understanding of FL in healthcare, this section reviews its foundational concepts, techniques, and current state of adoption. The discussion includes defining characteristics, advantages, and challenges of FL, methodologies tailored to healthcare applications, and their role in addressing tasks such as predictive modeling, diagnostics, and personalized care. Furthermore, the section explores the tangible benefits of FL, such as enhanced privacy, improved collaboration, and equitable access to advanced AI technologies,

[13] Source: Google FL Blog (Accessed December 11th, 2024)

while also highlighting limitations and areas for future research in this rapidly evolving field.

2.4.1 Fundamentals

This paradigm was introduced in 2016 as a collaborative learning framework to train models on decentralized data sources without transferring raw data to a central repository. It was initially conceived to address the limitations of edge and mobile device applications, described as an *"unbalanced and non-IID (identically and independently distributed) way of partitioning data across a huge number of unreliable devices with limitations such as bandwidth"* [182]. While originally designed to enable collaborative model training for mobile and edge devices, FL has rapidly evolved, extending its applications to finance, telecommunications, and healthcare domains, where centralized data processing is either impractical or undesirable due to privacy, regulatory, or logistical challenges.

At its core, FL enables multiple entities, often referred to as clients, to collaboratively solve a machine-learning task under the coordination of a central server or service provider. [183] define FL as a *"machine learning setting where multiple entities collaborate in solving a machine learning task under the coordination of a central server or service provider."* Unlike traditional centralized learning paradigms, where raw data are aggregated in a single repository, FL operates on decentralized datasets. Each client retains its raw data locally and transmits only model updates, such as weights or gradients, to a central server for aggregation. This architecture inherently addresses critical data privacy, security, and regulatory compliance issues while reducing data centralization costs.

Mathematically, FL aims to optimize the following global objective function [184]:

$$\min_{\omega} \mathcal{L}(\omega) = \sum_{c=1}^{C} f_c \mathcal{L}_c(\omega), \tag{2.33}$$

where ω represents the parameters of the global model, C is the number of participating clients, $\mathcal{L}_c(\omega)$ is the local loss function for client c, and f_c denotes the relative importance or weight of the contribution of each client.

The global model $Model_{global}$ is iteratively refined by aggregating updates from local models $Model_c$ trained on client-specific datasets D_c. The aggregation process is denoted as:

$$Model_{global}^{(t+1)} = \text{AGG}(Model_1^{(t)}, Model_2^{(t)}, ..., Model_C^{(t)}), \tag{2.34}$$

where AGG$(\cdot)$ represents the aggregation function, such as Federated Averaging (FedAvg) [182].

Architecture of Federated Learning

The FL architecture is built around a decentralized framework that ensures data privacy and security while enabling collaborative model training. A typical FL system comprises two main components: the central server and the clients. Depending on the implementation, these elements can be configured in one of three primary scenarios:

- **Central Server:** This component, although not always necessary, plays a pivotal role in coordinating the entire FL process. It handles tasks such as model initialization, aggregation, and redistribution of the global model. The central server is critical in centralized FL setups, where it acts as the core entity facilitating communication with the clients. However, in decentralized or hierarchical setups, its role is either replaced by peer-to-peer communication or augmented by intermediate aggregators.
- **Clients:** Representing individual entities, such as devices, institutions, or organizations, the clients perform local model training using their private datasets. Depending on the scenario, clients may interact with the central server or directly with each other. In a centralized FL scenario, clients communicate exclusively with the central server, sending updates and receiving aggregated model parameters. In a decentralized FL setup, clients engage in peer-to-peer communication, collaboratively exchanging and aggregating updates without a central server. For hierarchical FL, clients communicate with intermediate aggregators that consolidate updates before interacting with the central server, enhancing scalability and reducing communication overhead.

The workflow of FL generally follows an iterative process comprising the following steps:

1. **Initialization:** The central server initializes a global model ω^0, which can either be randomly initialized or pre-trained. This model is broadcast to all participating clients.
2. **Local Training:** Each client trains the global model using its local dataset D_c. The updated local model parameters $\omega_c^{(t+1)}$ are computed as follows:

$$\omega_c^{(t+1)} = Train(\omega^t, D_c), \tag{2.35}$$

   where $Train$ represents the optimization algorithm (e.g., Stochastic Gradient Descent (SGD)) applied to minimize the local loss function $\mathcal{L}_c(\omega)$.
3. **Model Upload:** After local training, clients send their model updates (e.g., weights, gradients) to the central server. These updates are often encrypted or anonymized to ensure privacy and security.
4. **Aggregation:** The central server aggregates the updates received from the clients to produce an updated global model:

$$\omega^{(t+1)} = \text{AGG}(\omega_1^{(t+1)}, \omega_2^{(t+1)}, ..., \omega_C^{(t+1)}). \tag{2.36}$$

5. **Broadcast:** The updated global model $\omega^{(t+1)}$ is broadcast back to the clients, initiating the next training iteration.
6. **Iteration and Convergence:** The above steps are repeated for multiple communication rounds until the model achieves a satisfactory level of accuracy or convergence criteria are met.
7. **Deployment:** The final global model is deployed for real-world use once the model converges. This could mean deploying the model in clinical decision support systems or diagnostic applications in healthcare.

These steps, combined with the architecture outlined in Figure 2.10, highlight the iterative collaboration between the central server and the clients while maintaining data privacy and scalability.

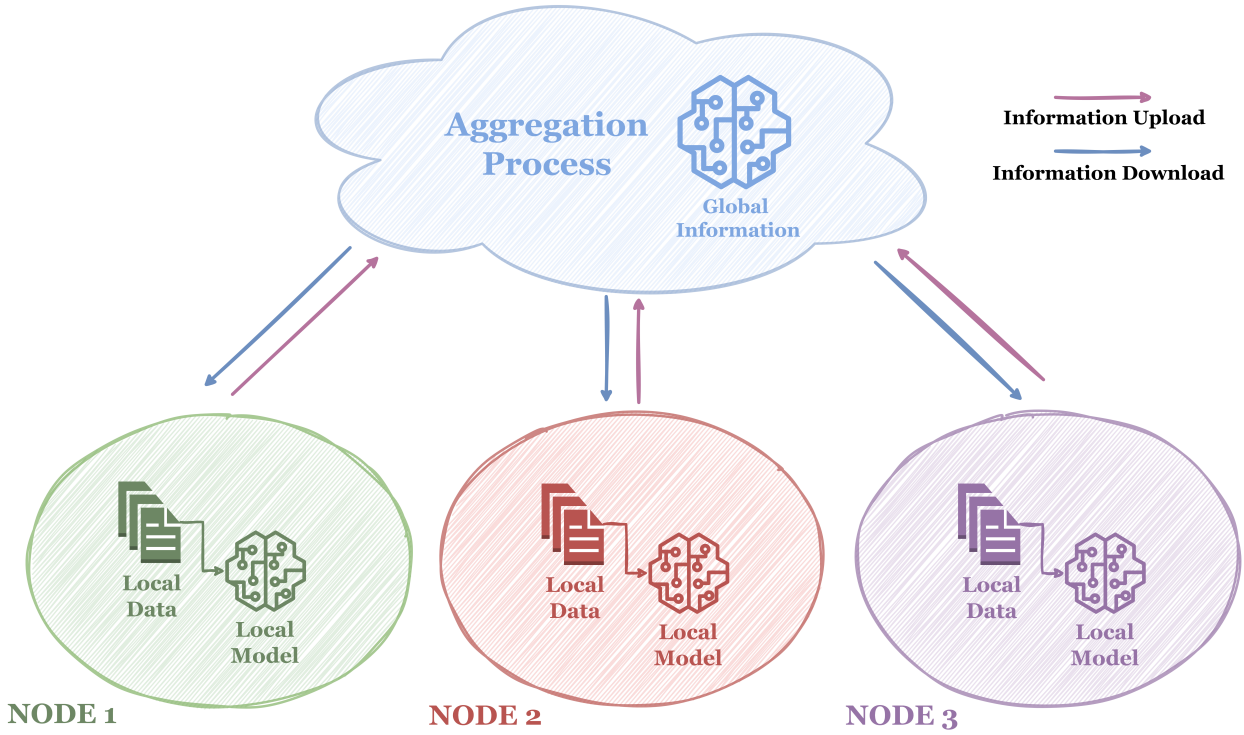


Figure 2.10: FL general architecture. Multiple nodes (e.g., Node 1, Node 2, and Node 3) collaboratively train a global model without sharing their local data. Each node trains a local model using its private dataset and uploads only relevant information—such as model parameters, gradients, or other metadata—to a central aggregation process. The aggregation server combines the updates from all nodes to create an improved global model, which is subsequently downloaded back to the nodes for further training iterations.

Types of Federated Learning

FL can be categorized into three primary paradigms, each designed to address specific data distribution scenarios: Horizontal Federated Learning (HFL), Vertical Federated Learning (VFL), and Federated Transfer Learning (FTL). These paradigms differ based on how the data are partitioned across clients regarding feature, sample, and label space.

1. **Horizontal Federated Learning:** HFL is applicable when datasets across clients share the same feature space but differ in sample space. In this scenario, clients possess datasets with similar attributes (features) but containing records from different individuals (samples). This setup is common when institutions collect the same data type but from different populations. Formally, for any two clients i and j, their datasets

D_i and D_j satisfy [185]:

$$X_i = X_j, Y_i = Y_j, I_i \neq I_j \quad \forall D_i, D_j, \quad i \neq j, \tag{2.37}$$

where X_i and X_j represent the feature spaces of clients i and j, Y_i and Y_j denote the label spaces in a supervised learning scenario, and I_i and I_j represent the sample spaces (indices of data instances).

An example of HFL in healthcare is when multiple hospitals with identical medical data schemas collaborate to train a global model for disease prediction or diagnosis. Each hospital has data on different patients but records the same clinical features (e.g., blood pressure and cholesterol levels). This approach enables the integration of datasets with similar attributes while preserving local data confidentiality [186], [187].

HFL has been widely adopted in medical applications, particularly for disease classification tasks [188]. For instance, Li et al. [189] demonstrated HFL in autism spectrum disorder prediction involving four medical institutions across different locations. Each institution shared the same user features but had different patient samples, allowing them to train a global model collaboratively without sharing sensitive patient data.

2. **Vertical Federated Learning:** VFL Vertical Federated Learning is applicable when clients have datasets that share a subset of overlapping sample space but differ in feature space. This scenario arises when different entities hold complementary information about the same individuals. Formally, for any two clients i and j, their datasets D_i and D_j satisfy [185]:

$$X_i \neq X_j, Y_i \neq Y_j, I_i = I_j \quad \forall D_i, D_j, \quad i \neq j. \tag{2.38}$$

In this context, the clients share the same sample space $I_i = I_j$, meaning they have data on the same individuals, and their feature spaces and label spaces (in a supervised scenario) are different.

A typical example in healthcare is the collaboration between a hospital and an insurance company. The hospital holds clinical data (e.g., treatment history, medical imaging) on patients, while the insurance company possesses financial and claim information for the same individuals. By aligning their datasets on the shared patient identifiers, they can jointly train a global model that benefits from the combined feature set, enhancing predictive performance while maintaining data privacy [187], [190].

[191] proposed a model based on VFL for partitioning medical data. In their approach, features from each user were transformed into latent dimensions using an AE in an FL setting. These latent representations were concatenated and sent to the global model for further training, enabling collaborative learning without sharing raw data.

3. **Federated Transfer Learning:** FTL addresses situations where both the feature spaces and sample spaces are different across clients, with minimal or no overlap. This paradigm is particularly useful when datasets are small or highly heterogeneous and clients have little in common regarding data attributes or samples.

Formally, for any two clients i and j, their datasets D_i and D_j satisfy [185]:

$$X_i \neq X_j, Y_i \neq Y_j, I_i \neq I_j \quad \forall D_i, D_j, \quad i \neq j. \tag{2.39}$$

In this scenario, the clients have different sample spaces $I_i = I_j$, meaning they have data on different individuals, and their feature spaces and label spaces (in a supervised scenario) are different.

FTL combines transfer learning with FL to adapt a pre-trained global model to local datasets with minimal data similarity. It leverages knowledge from a source domain to improve learning in a target domain, facilitating model training when data overlap is limited or nonexistent [192].

In healthcare, FTL can assist in disease diagnosis using data from different patients in multiple hospitals with varying therapeutic programs. For example, hospitals in other countries may have datasets with different features due to variations in medical practices and patient demographics. FTL allows these institutions to collaboratively improve their models by transferring knowledge and enhancing performance without compromising privacy [187].

[193] developed an FTL-based model called FedHealth, which collects data from healthcare institutions and creates personalized models for each client. FedHealth initially learns human activity recognition tasks and then employs transfer learning to extend classification to the diagnosis of Parkinson's disease, demonstrating the potential of FTL in creating global models adaptable to various medical challenges.

In summary, the three paradigms of FL—horizontal and vertical FL and FTL —provide flexible frameworks for collaborative model training under different data distribution scenarios. By selecting the appropriate paradigm, institutions can harness collective intelligence while preserving data privacy and advancing ML applications in sensitive domains like healthcare.

Federated Learning Tools for Implementation and Simulation

Over the years, various FL frameworks have been developed to support the implementation and simulation of FL models. Table 2.4 summarizes some of the most prominent systems available today, highlighting their unique features and areas of application.

[14]Source: FL powered by NVIDIA Clara (Accessed December 11th, 2024

Framework	Key Features	Use Cases
FATE [194]	Modular design with secure computation protocols sand algorithm customization; supports private set intersection and Kubernetes deployment.	Enterprise-level FL deployments, including healthcare, finance, and other data-sensitive domains.
TensorFlow Federated [195]	Python-based interface leveraging TensorFlow; supports FedAvg algorithms for collaborative learning.	Academic research and lightweight FL experiments, particularly in educational or smaller-scale environments.
PySyft [196]	Focused on secure computation with multi-party computation and DP; integrates with PyTorch and TensorFlow.	Privacy-sensitive domains, including healthcare and finance, where data confidentiality is critical.
IBM FL Framework [197]	Enterprise-focused framework with APIs for ML library integration; supports secure MPC.	Enterprise-level applications in healthcare, finance, and collaborative research projects.
NVIDIA Clara[14]	Specialized for healthcare, providing AI-powered imaging and genomics solutions with built-in GPU support; supports CUDA environments.	Medical imaging, diagnostics, and other high-performance healthcare use cases.
FLOWER [198]	Scalable framework for horizontal and vertical FL, supporting a wide range of devices; offers flexible aggregation methods, extensive communication stack, and virtual client engine.	Academic research and large-scale industrial FL deployments, particularly on mobile and edge devices
FedLab [199]	Provides customizable interfaces for FL; supports tensor-based communication, aggregation modules, and hierarchical FL simulation.	Cross-process, standalone, and FL simulations for communication and optimization research.
FedML [200]	Supports diverse FL configurations and paradigms; includes on-device testbeds for mobile and IoT platforms.	Real-world FL implementations requiring robust support for heterogeneous devices and environments.
FedScale [201]	Offers standardized APIs and runtime environments for FL evaluation; provides datasets for realistic FL use cases.	On-device FL evaluation for mobile devices and in-cluster FL simulations.
OpenFL [186]	Open-source FL framework tailored for medical imaging; supports PyTorch and TensorFlow models.	Federated tumor segmentation, respiratory distress prediction, and other collaborative medical applications.
Fed-BioMed [202]	Focused on real-world medical applications; supports multi-center analysis and integration with PySyft.	Multi-center medical data analysis for structural imaging, including MRI datasets.

Table 2.4: Overview of FL frameworks. The table highlights their key features and use cases, particularly on applications in privacy-sensitive domains like healthcare and collaborative research.

2.4.2 Aggregation Techniques

Aggregation is a cornerstone of FL, enabling the integration of locally trained models from multiple clients into a unified global model. The choice of aggregation strategy significantly impacts the quality, convergence, and robustness of the global model while addressing challenges such as non-IID data distributions, resource constraints, and adversarial influences. FedAvg is the most widely used technique among the various methods due to its simplicity, efficiency, and versatility across multiple applications. Given its dominance in FL research and practice, we thoroughly examine FedAvg before exploring other notable techniques.

Federated Averaging

FedAvg, initially introduced by [182], is the most widely employed method in FL. FedAvg operates through an iterative process where multiple nodes train a global model collaboratively. Each node holds a portion of the data and trains the model locally. The key steps involved in FL are as follows:

1. **Initialization:** A single model architecture is designed for all participating nodes. While the initial weights do not need to be identical across nodes, the model architecture (such as the number of layers) must remain consistent to ensure compatibility during aggregation. Each participating node then initializes and trains this shared model architecture using its local dataset.

2. **Local Training:** Each of the C, represented as $c = 1, 2, ..., C$, uses its local dataset D_c to perform training on their local models. These training results are denoted as ω_c for the c-th node. The local training process can involve multiple epochs and use standard optimization methods such as stochastic gradient descent.

3. **Model Aggregation:** Following local training, each node sends its updated parameters ω_c to a central server. The central server performs weighted averaging to combine these local models and update a global model ω. The weighted average is calculated by considering the number of local data samples N_c at each node. The aggregated global model ω is computed as follows:

$$\omega = \sum_{c=1}^{C} \frac{N_c}{N_{total}} \omega_c, \tag{2.40}$$

   where $N_{total} = \sum_{c=1}^{C} N_c$ is the total number of samples across all nodes.

4. **Iterative Training Process:** The model distribution, local training, and aggregation process is repeated iteratively. The global model is updated and sent back to the nodes to refine it using their local datasets at each iteration. This iterative process continues until the global model reaches a desired level of accuracy or a predetermined number of iterations is completed.

5. **Stopping Criteria and Convergence:** The training process can be terminated after a fixed number of iterations or when the global model achieves convergence, where further updates do not yield significant improvements in model performance. This is

typically monitored using a validation set or performance metrics calculated during training.

The core of the FedAvg algorithm lies in the aggregation step, which ensures that the global model captures the distributed knowledge from all participating nodes without requiring access to their local data. This formulation guarantees that nodes with larger datasets (i.e., higher N_c) contribute more to the global model, thereby enhancing the overall generalization of the model. Using weighted averaging mitigates the risk of overfitting to individual nodes and ensures a balanced contribution from nodes with varying data quantities. Figure 2.11 depicts the explained process.

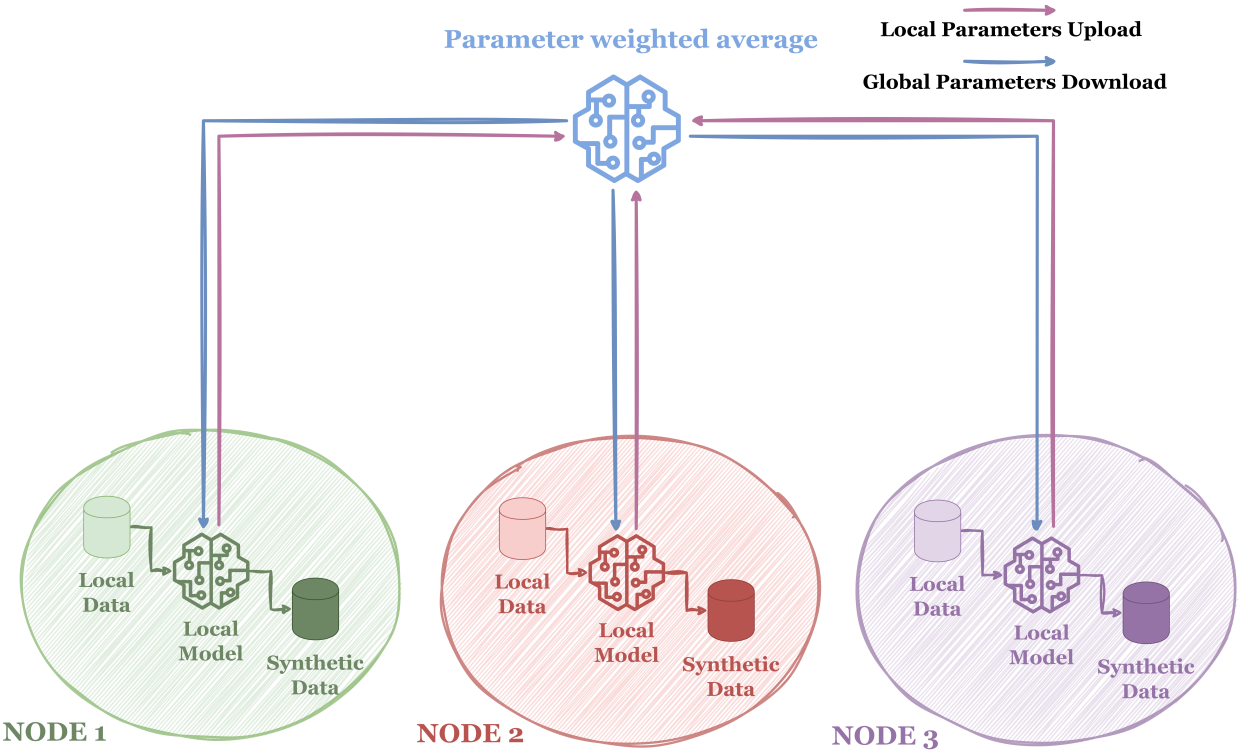


Figure 2.11: FedAvg process in FL. Each node trains a local model based on its data and shares the model parameters with a central server. The server averages the parameters over several rounds to create a global model, which is then distributed back to the nodes.

Despite its popularity, FedAvg exhibits several limitations, particularly in challenging FL scenarios:

- **Non-IID Data Distributions:** FedAvg assumes that data distributions across clients are similar (IID). However, in real-world scenarios, data are often non-IID, leading to biased local updates that hinder the convergence and generalization of the global model. This challenge is particularly pronounced in healthcare, where patient demographics and medical practices vary across institutions.
- **Communication Efficiency:** While FedAvg reduces communication overhead compared to fully synchronous methods, the iterative exchange of model parameters between

clients and the server can still be bandwidth-intensive, especially in environments with limited connectivity or resource-constrained devices.

- **Client Heterogeneity:** FedAvg struggles with client hardware capabilities and resource availability variations. Clients with limited computational power or smaller datasets may underperform, leading to an imbalanced contribution to the global model.
- **Model Divergence:** In scenarios with extreme heterogeneity, local models may diverge significantly from the global objective, resulting in suboptimal convergence or the failure to achieve a satisfactory global model.

Other Aggregation Techniques

While FedAvg remains the foundational aggregation method, several advanced techniques have been developed to address its limitations and expand the applicability of FL.

FedProx [203] was designed to extend the capabilities of FedAvg in scenarios with highly heterogeneous data distributions and diverse resource capacities. One significant limitation of FedAvg is its tendency to diverge when the local objectives of the clients differ substantially from the global objective. To mitigate this, FedProx introduces a proximal term into the local objective function, expressed as:

$$\frac{\mu}{2}||\omega_c^t - \omega_{global}^t||^2, \tag{2.41}$$

where $\mu \geq 0$ acts as a penalty parameter regulating the strength of the constraint. This term penalizes deviations between local and global models, ensuring that local updates remain aligned with the global goal. FedProx is particularly useful in healthcare collaborations involving institutions with diverse patient populations and data distributions. For example, FedProx can stabilize the convergence of the global model in multi-hospital studies by aligning locally trained models. However, the effectiveness of FedProx relies on careful tuning of μ. Improper tuning can either overly constrain local updates, reduce their flexibility, or fail to stabilize the optimization process. Additionally, the proximal term increases the computational complexity of local training, which can be a challenge for resource-limited devices.

Federated Stochastic Gradient Descent (FedSGD) [182] is a method tailored for environments where computational and communication resources are constrained. Unlike traditional approaches that emphasize accuracy, FedSGD focuses on optimizing communication efficiency by reducing the frequency and size of gradient updates. This makes it particularly suitable for edge devices or Internet of Things (IoT) sensors in healthcare applications, such as wearable health monitors or remote patient management systems. FedSGD allows collaborative training across large-scale, distributed networks without overburdening individual devices by minimizing communication overhead.

Federated Matched Averaging (FedMA) [204] introduces a novel layer-wise aggregation approach for neural networks. Instead of averaging parameters directly, FedMA aligns similar layers across different client models before aggregation, ensuring structural consistency in feature extraction. This alignment is particularly advantageous in complex medical imaging tasks, such as tumor segmentation or radiological diagnostics, where variations in

model architectures across clients can hinder the effectiveness of traditional aggregation techniques. However, the computational demand of FedMA for matching layers and reliance on homogeneous network architectures make it less suitable for large-scale deployments with highly diverse client environments.

Scaffold [205] addresses client drift, a common issue in FL caused by heterogeneity in local datasets. By introducing control variates that adjust local updates, Scaffold ensures alignment with the global objective, accelerating convergence and improving consistency in the global model. This technique is particularly effective in healthcare scenarios involving multi-institutional collaborations, such as genomic studies, where data distributions across institutions vary significantly. However, the need for additional computation to maintain and apply control variates can increase the overall system complexity.

Federated Normalized Averaging (FedNova) [206] redefines the aggregation process by normalizing client updates based on their level of participation. This approach promotes fairness in contributions to the global model, ensuring that institutions with smaller datasets or underrepresented populations are not disproportionately weighted. FedNova is particularly valuable in healthcare contexts where data availability and quality vary widely across institutions, such as rural versus urban hospitals. By balancing contributions, FedNova enhances the inclusiveness and robustness of the global model.

Aggregation in FL encompasses various techniques designed to address the diverse challenges of heterogeneous data and computational environments. While FedAvg remains the dominant method due to its simplicity and effectiveness, advanced techniques such as FedProx, FedSDG, FedMA, Scaffold, and FedNova offer specialized solutions for complex scenarios. These methods enhance the applicability of FL in healthcare by improving model robustness, fairness, and efficiency. However, this discussion has focused on the most widely used techniques; numerous other aggregation methods exist, providing ample opportunities for further research and development.

2.4.3 Federated Learning Based Models for Healthcare Applications

Healthcare data are inherently distributed across multiple institutions and are often characterized by sensitive and diverse datasets. FL has emerged as a promising solution to leverage such data while ensuring privacy and compliance with strict regulations like HIPAA. Recent studies demonstrate the extensive adoption of FL across various healthcare tasks, primarily in classification, segmentation, anomaly detection, and regression problems. According to [207], approximately 70 of 89 reviewed studies from 2015 to 2023 applied FL to classification tasks, with segmentation addressed in 7 studies, while the remainder focused on other specialized applications such as tensor factorization, feature selection, and anomaly detection.

1. **Classification Tasks:** Classification remains the most explored domain in FL-based healthcare applications. Studies span a wide spectrum of diseases and use cases, showcasing the potential of FL to enhance diagnostic capabilities while addressing privacy concerns.

- **Cancer Detection:** FL has been instrumental in early cancer detection across various types, including breast, lung, and skin cancers. For example, [208] utilized FL with transfer learning models like ResNet and DenseNet for chest X-ray analysis to predict Pneumonia and other abnormalities. These models achieved high classification accuracy while preserving data confidentiality.
- **COVID-19 Diagnosis:** During the COVID-19 pandemic, FL frameworks supported collaborative model training for detecting COVID-19 using chest X-rays and CT scans. Studies like [209] integrated advanced aggregation techniques and secure communication protocols, achieving competitive diagnostic performance even under non-IID data distributions.
- **Neurological and Mental Health Disorders:** Personalized FL approaches have shown promise in detecting epileptic seizures [210] and assessing mental health conditions such as major depressive disorder [211]. These studies emphasize the role of FL in creating robust and generalizable models while addressing challenges like data heterogeneity and imbalance.
- **Multi-Disease Detection:** FL frameworks have been developed to tackle the complexity of multi-disease classification, integrating diverse datasets to identify multiple conditions simultaneously. For example, [212] designed FL systems for diagnosing bacterial and viral Pneumonia alongside COVID-19, achieving significant improvements in accuracy through advanced aggregation techniques like Scaffold and FedBN.

2. **Segmentation Tasks:** FL has also been applied to segmentation tasks, particularly in medical imaging, where delineating structures like tumors and lesions is crucial for diagnosis and treatment planning.
   - **Tumor Segmentation:** FL models have been employed for brain and prostate tumor segmentation, utilizing advanced architectures to align feature extraction across institutions. For instance, [213] combined curriculum learning with FL to improve segmentation outcomes.
   - **Skin Lesions and Pneumothorax:** Studies have extended FL to segment skin lesions and detect pneumothorax in chest X-rays [214]. Aggregation methods like FedAvg and FedProx have improved segmentation accuracy under challenging data scenarios.
3. **Specialized Applications:** Beyond classification and segmentation, FL has been employed in various specialized healthcare applications:
   - **MRI Reconstruction:** FL models for MRI reconstruction have enhanced imaging efficiency and accuracy, reducing computational costs while maintaining privacy [215].
   - **Prognostic Models:** In SA and lifespan prediction, FL has facilitated collaborations among healthcare institutions to develop CoxPH [216], [217] tailored to diverse demographics.

- **Cancer Prediction and Staging:** Self-supervised FL frameworks, such as SSL-FL-BT [218], have demonstrated their efficacy in histopathological cancer detection, offering new avenues for early diagnosis.

This review highlights the extensive application of FL in healthcare, emphasizing its potential in leveraging distributed and heterogeneous datasets. While adopting FL is widespread in classification tasks, segmentation, integration, and specialized use cases are gaining momentum. Integrating novel aggregation methods and privacy-preserving mechanisms ensures the adaptability of FL to various healthcare scenarios. Further exploration of underrepresented tasks and methodological enhancements will drive future innovations in FL-based healthcare systems.

2.4.4 Challenges and Future Trends

Implementing FL frameworks is fraught with challenges stemming from the unique nature of healthcare data, the constraints of decentralized networks, and the diversity of participating institutions and devices. Addressing these challenges is critical to optimizing FL systems and ensuring their practical adoption. This section outlines key challenges in implementing FL in healthcare, alongside potential future trends that may drive innovation and adoption in this field.

Challenges in Federated Learning for Healthcare

Figure 2.12 visually outlines these key challenges, highlighting the fundamental barriers that hinder the effective implementation of FL in healthcare environments.

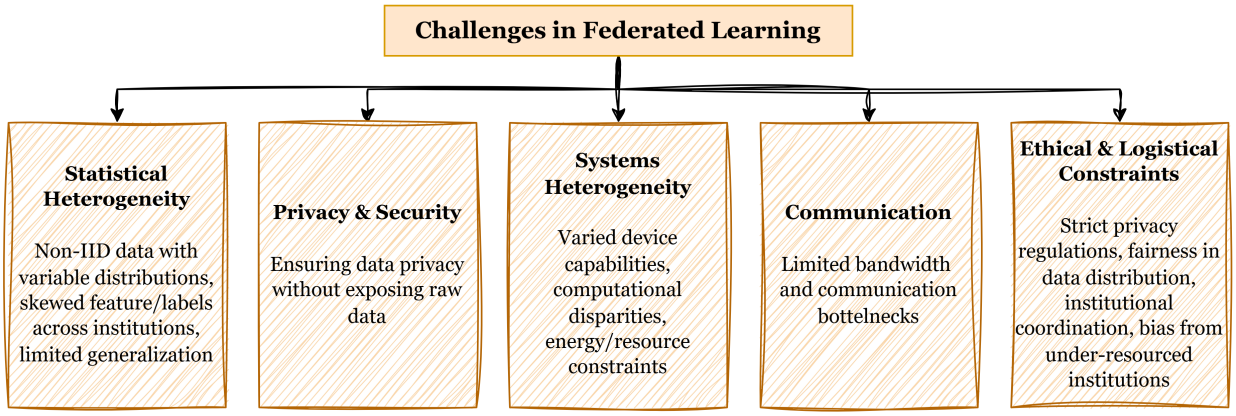


Figure 2.12: Challenges in Federated Learning for Healthcare. The figure highlights key challenges such as statistical heterogeneity, privacy and security concerns, systems heterogeneity, communication bottlenecks, and ethical and logistical constraints.

→ ***Statistical Heterogeneity***
Healthcare data are inherently non-IID, with significant variability across institutions, demographics, and geographies. Hospitals and clinics collect data under different conditions, leading to datasets with skewed distributions of features and labels. For instance, some

institutions may focus on a specific patient demographic or disease type, resulting in localized models that may not generalize well across broader populations.

Traditional FL algorithms like FedAvg struggle with this heterogeneity, often failing to converge or to produce suboptimal global models [182]. Enhancements like FedProx introduce mechanisms to mitigate divergence by aligning local and global objectives [203]. Despite these advances, the issue of statistical heterogeneity remains a persistent barrier to achieving effective FL in healthcare, particularly in diverse clinical settings.

→ ***Privacy and Security***
FL provides a significant advantage by allowing models to be trained collaboratively without exposing raw data. However, ensuring robust privacy and security in this setting is challenging and out of the scope of this thesis. Techniques such as DP and cryptographic methods like Secure Multi-Party Computation (MPC) and Homomorphic Encryption (HE) are critical for protecting sensitive medical data during the learning process [219], [220]. Nevertheless, these techniques can increase computational overhead and reduce efficiency, particularly for resource-constrained devices. Further exploration of these methods remains an active area of research and innovation.

→ ***Systems Heterogeneity***
Another significant challenge is the diversity of hardware and software capabilities among clients in an FL network. Devices participating in FL may range from high-performance hospital servers to low-power wearable devices. This variation introduces disparities in computational power, memory capacity, and energy availability.

Approaches to address these challenges include:

- **Asynchronous Communication [221]:** Allowing devices to upload updates at different times to accommodate varying capabilities.
- **Active Device Sampling [222]:** Selecting clients based on their resources and availability.
- **Fault Tolerance [223]:** Managing device failures through redundancy or selective updates.

However, these solutions come with trade-offs in terms of scalability and complexity, which require further refinement.

→ ***Communication Bottlenecks***
Communication efficiency is a critical bottleneck in FL, particularly in cross-device settings where network bandwidth may be limited or unreliable. Frequent model updates between clients and servers can overwhelm communication channels, especially in large-scale deployments.

Solutions include:

- **Local Updates [203]:** Increasing the number of local training iterations to reduce communication frequency.

- **Compression Schemes [224]:** Reducing the size of updates using techniques like gradient compression and model pruning.
- **Decentralized Training [225]:** Implementing peer-to-peer communication frameworks to minimize dependence on central servers.

Despite these strategies, balancing communication costs with model accuracy remains an ongoing challenge in FL research.

→ ***Ethical and Logistical Constraints***
Beyond technical challenges, FL must navigate ethical and logistical barriers in healthcare. Privacy regulations such as GDPR and HIPAA impose strict constraints on data usage, and ethical considerations prevent the unrestricted pooling of sensitive patient information. Logistical challenges include coordinating multiple institutions with varying policies and infrastructures and addressing fairness issues where under-resourced institutions may contribute less data, potentially leading to biased global models.

Future Trends in Federated Learning for Healthcare

- **Personalization and Multi-Model Approaches:** To address statistical heterogeneity and improve model performance, future FL systems may shift toward personalized models tailored to individual clients. Multi-model frameworks, where each client retains a customized model while contributing to a global meta-model, could enhance local performance and generalizability.
- **Advanced Aggregation Techniques:** Innovations in aggregation algorithms, such as FedMA and FedNova, will continue to play a crucial role in improving convergence and fairness in FL. These methods can accommodate diverse data distributions and client capabilities, making them particularly suitable for healthcare applications.
- **Decentralized and Hierarchical Architectures:** Implementing decentralized or hierarchical architectures could mitigate communication bottlenecks and improve scalability by enabling direct client communication or using intermediate aggregators, reducing reliance on central servers.
- **Integration with Edge and IoT Devices:** The increasing prevalence of wearable and IoT devices in healthcare presents opportunities for FL. Lightweight algorithms optimized for edge computing will enable these devices to participate in FL networks, facilitating real-time analytics and personalized healthcare solutions.
- **FL for Medical Tasks such as SDG and SA:** Applying FL to SDG and SA offers a promising research direction. FL-enabled SDG allows privacy-preserving synthetic data generation, fostering collaboration without sharing sensitive patient data. FL enhances predictive modeling in SA by aggregating survival patterns from distributed datasets, improving accuracy and generalizability.
- **Regulatory Compliance and Ethical AI:** Future FL frameworks must integrate mechanisms to ensure compliance with evolving privacy regulations and ethical standards. Transparent and auditable processes and verifiability techniques like zero-knowledge

proofs will be essential for building trust in FL systems.

While FL offers a huge potential in healthcare, its widespread adoption requires addressing several technical and ethical challenges. FL can achieve its promise of enabling collaborative, privacy-preserving AI in healthcare by advancing statistical heterogeneity, systems and communication efficiency, and personalized modeling. Future research and development will play a pivotal role in overcoming these barriers and unlocking the full potential of FL in this critical domain.

Chapter 3

Survival Analysis

3.1 Introduction

SA remains a cornerstone of modern healthcare analytics, enabling clinicians and researchers to extract meaningful insights from complex temporal data. As explored in the State-of-the-Art review in Section 2.2, its integration with AI has revolutionized the field, introducing models capable of addressing challenges posed by high-dimensional and non-linear data. These advancements are essential for personalized medicine, healthcare optimization, and improving patient outcomes. However, significant gaps remain in the current methodologies, demanding further research and development.

Traditional SA models, such as the KM estimator and CoxPH model, provide interpretable and robust tools for analyzing survival data. Nevertheless, they are limited by their assumptions of linear relationships and proportional hazards, making them less effective in real-world settings characterized by complex, non-linear relationships and high-dimensional covariates. ML and DL approaches, such as DeepSurv and DeepHit, have addressed some limitations, introducing models capable of capturing intricate patterns in survival data. Despite their revolutionary potential, these models are not without shortcomings. For instance, DeepSurv retains the assumption of proportional hazards. At the same time, DeepHit, one of the few models designed for CR, often encounters challenges in convergence and operates exclusively in the discrete-time domain.

This section presents the methodology behind two novel approaches: Survival Analysis Variational Autoencoder (SAVAE) and Competing Risks-SAVAE (CR-SAVAE). SAVAE is a deep GM designed to address the limitations of both traditional and ML-based SA methods. It leverages VAEs to model non-linear relationships, high-dimensional data, and censored observations, providing a robust and flexible tool for SA. CR-SAVAE extends SAVAE to tackle CIFs in CR scenarios, addressing the unique challenges of multiple CR in survival data. Using various datasets, both models are evaluated against established benchmarks, including CoxPH, DeepSurv, and DeepHit. These contributions address the gaps identified in the State-of-the-Art review, offering robust tools for advancing SA in personalized medicine and beyond.

The development and evaluation of these models are detailed in the following publications:

- **SAVAE:** A. Apellániz P., Parras J., and Zazo S., *Leveraging the variational Bayes autoencoder for survival analysis,'* in Scientific Reports, vol. 14, 24567, 2024, doi: 10.1038/s41598-024-76047-z
- **CR-SAVAE:** A. Apellániz P., Parras J., and Zazo S., *CR-SAVAE: A Parametric Method for Survival Analysis with Competing Risks,'* in the 32nd European Signal Processing Conference (EUSIPCO), Lyon, France, 2024, pp. 1526-1530, doi: 10.23919/EUSIPCO63174.2024.10715431.

3.2 SAVAE: Variational Autoencoders for Survival Analysis

3.2.1 Methodology of the Proposed Model (SAVAE)

We recall the notation for an SA problem introduced in Chapter 2 to define and explain the proposed model while maintaining consistency with that notation. In this context, a dataset D consists of N observations, each represented as a triplet $D = (x_i, t_i, d_i)_{i=1}^N$, where:

- $x_i = (x_i^1, ..., x_i^C ov)$ is a Cov-dimensional vector of covariates associated with the i-th individual.
- t_i denotes the observed time-to-event.
- $d_i \in \{0, 1\}$ is the censoring indicator, where $d_i = 1$ signifies that the event of interest was observed at time t_i (uncensored data), and $d_i = 0$ indicates that the event was not observed up to t_i (censored data).

In SA, the probability density function $p(t|x)$, hazard function $h(t|x)$, and the survival function $S(t)$ are related through the equation:

$$p(t|x) = h(t|x)S(t|x). \tag{3.1}$$

This equation highlights that the probability density of the event occurring at a specific time t, given the covariates x, can be expressed as the product of the instantaneous hazard at time t and the probability of surviving beyond t, thus providing a foundational relation in SA.

This section introduces SAVAE, the VAE-based model for SA. To facilitate understanding, a detailed explanation of the original VAE architecture can be found in Appendix A.

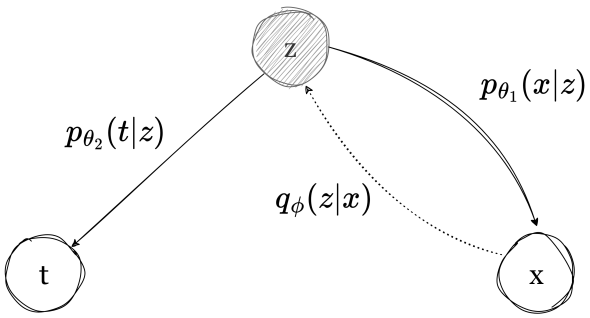


Figure 3.1: SAVAE Bayesian model. The shadowed circle refers to the latent variable, and the white circles refer to the observables. Note that the probabilities $p_{\theta_1}(x|z)$ and $p_{\theta_2}(t|z)$ denote the GMs, and $q_\phi(z|x)$ denotes the variational approximation to the posterior, since the true posterior $p_\theta(z|x)$ is unknown.

The interest lies in using VAEs to obtain the predictive distribution of time-to-event given covariates. The proposed SAVAE approach, depicted in Figure 3.1, extends the Vanilla VAE. SAVAE includes a continuous latent variable z, two vectors (an observable covariate vector x and the time-to-event t), and GMs $p_{\theta_1}(x|z)$ and $p_{\theta_2}(t|z)$, assuming conditional independence,

which is a characteristic inherent to VAEs and their ability to model the joint distribution of variables effectively. This means that knowing z, the components of the vector x and t can be generated independently. A single variational distribution estimates the variational posterior $p(z|x)$ to define the predictive distribution based on covariates. While it is possible to include the effect of time ($p(z|t,x)$), this approach focuses on using only covariates to obtain the latent space, as the time t can be unknown to predict survival times for test patients and could be censored. SAVAE combines VAEs and SA, offering a flexible framework for modeling complex event data.

Goal

To achieve the main objective, which is to obtain the predictive distribution for the time to event, variational methods will be used in the following way [226]:

$$p\left(t^*|x^*, \{x_i, t_i\}_{i=1}^N\right) = \int p\left(t^*|z, \{x_i, t_i\}_{i=1}^N\right) p\left(z|x^*, \{x_i, t_i\}_{i=1}^N\right) dz, \tag{3.2}$$

where x^* represents the covariates of a particular patient, and its survival time distribution $p\left(t^*|z, \{x_i, t_i\}_{i=1}^N\right)$ needs to be estimated.

ELBO derivation

Considering our main objective and the use of VAE as the architecture on which we base our approach, the Evidence Lower BOund (ELBO) development described in Appendix A can be extended to apply to our case. SAVAE assumes that the two GMs $p_{\theta_1}(x|z)$ and $p_{\theta_2}(t|z)$ are conditionally independent. This implies that if z is known, generating x or t is possible. Furthermore, due to the VAE architecture, it is assumed that each component of the covariate vector x is also conditionally independent given z. Therefore,

$$p(x,t,z) = p_{\theta_1}(x|z)p_{\theta_2}(t|z)p(z) = p_\theta(x,t|z)p(z). \tag{3.3}$$

It also assumes that the distribution families of $p_{\theta_1}(x|z)$ and $p_{\theta_2}(t|z)$ are known, but not the parameters θ_1 and θ_2.

Taking into account these assumptions, the ELBO can be computed in a similar way to the Vanilla VAE. First, the conditional likelihood of a set of points $\{x_i, t_i\}_{i=1}^N$ can be expressed as follows:

$$\log p_\theta(x_1, x_2, ..., x_N, t_1, t_2, ..., t_N|z) = \sum_{i=1}^N \log p_\theta(x_i, t_i|z) = \sum_{i=1}^N \left(\log p_{\theta_2}(t_i|z) + \sum_{c=1}^{Cov} \log p_{\theta_1}(x_i^c|z)\right), \tag{3.4}$$

where the expected conditional likelihood can be expressed as:

$$\begin{aligned} &\mathbb{E}_z\left[p_\theta(x,t|z)\right] \\ &= \int p_\theta(x,t|z)p(z)dz \\ &= \int \frac{p_\theta(x,t,z)}{p(z)}p(z)dz \\ &= \int p_\theta(x,t,z)dz \\ = p_\theta(x,t) &= \int p_\theta(x,t,z)\frac{q_\phi(z|x)}{q_\phi(z|x)}dz \\ &= \mathbb{E}_{q_\phi(z|x)}\left[\frac{p_\theta(x,t,z)}{q_\phi(z|x)}\right]. \end{aligned} \tag{3.5}$$

As the interest lies in computing the log-likelihood:

$$\log p_\theta(x,t) = \log\left[\mathbb{E}_{q_\phi(z|x)}\left[\frac{p_\theta(x,t,z)}{q_\phi(z|x)}\right]\right] \geq \mathbb{E}_{q_\phi(z|x)}\left[\log\frac{p_\theta(x,t,z)}{q_\phi(z|x)}\right], \tag{3.6}$$

where the inequality comes from applying the inequality of Jensen. Then, this could be rearranged as:

$$\begin{aligned} &\mathbb{E}_{q_\phi(z|x)}\left[\log\left(\frac{p_\theta(x,t,z)}{q_\phi(z|x)}\right)\right] \\ &= \int q_\phi(z|x)\log\frac{p_{\theta_1}(x|z)p_{\theta_2}(t|z)p(z)}{q_\phi(z|x)}dz \\ &= -\int q_\phi(z|x)\log\frac{q_\phi(z|x)}{p(z)}dz + \int q_\phi(z|x)\left(\log p_{\theta_1}(x|z) + \log p_{\theta_2}(t|z)\right)dz \\ &= -D_{\mathbb{KL}}(q_\phi(z|x)||p(z)) + \mathbb{E}_{q_\phi(z|x)}\left[\log p_{\theta_1}(x|z) + \log p_{\theta_2}(t|z)\right] \\ &= \mathcal{L}(x,\theta_1,\theta_2,\phi). \end{aligned} \tag{3.7}$$

After computing this ELBO, it can be seen that it is similar to the Vanilla VAE (Equation (A.7)). The only difference lies in the reconstruction term, expressed differently to explicitly distinguish between the covariates and the time-to-event. By using Equation (3.4) and the reparameterization trick, the ELBO estimator is obtained, explicitly accounting for each dimension of the covariates vector:

$$\begin{aligned} &\hat{\mathcal{L}}(x,\theta_1,\theta_2,\phi) \\ &= \frac{1}{N}\sum_{i=1}^{N}\left(-D_{\mathbb{KL}}(q_\phi(z|x_i)||p(z)) + \log p_{\theta_2}(t_i|g_\phi(x_i,\epsilon_i)) + \sum_{c=1}^{Cov}\log p_{\theta_1}(x_i^c|g_\phi(x_i,\epsilon_i))\right). \end{aligned} \tag{3.8}$$

Three Deep Neural Networks (DNNs) have been used in implementation, as specified in Figure 3.2. Note that the decoder DNNs output the parameters of each distribution.

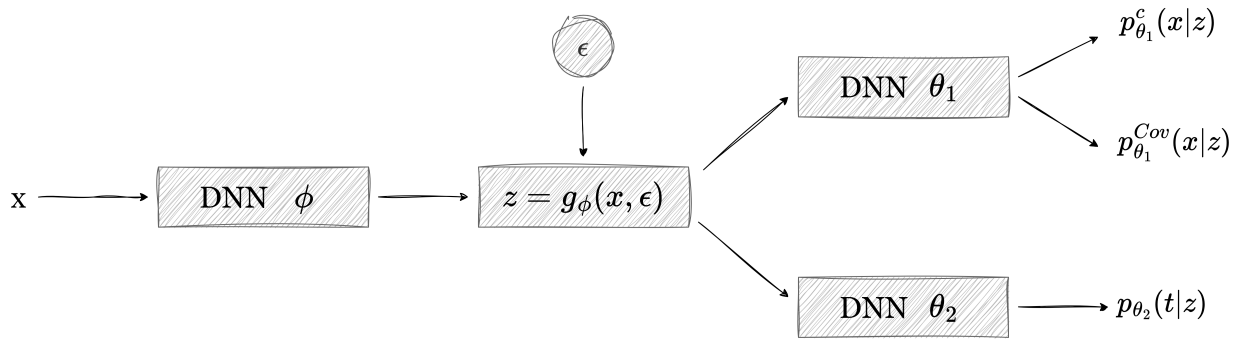


Figure 3.2: SAVAE implementation using DNNs. One acts as an encoder with the covariates vector as input. The other two act as decoders, one for the covariates and the other for the time.

Divergence computation

SAVAE assumes that $q_\phi(z|x)$ follows a multidimensional Gaussian distribution defined by a vector of means μ, where each element is μ_j and by a diagonal covariance matrix $\mathbf{C}$, where the main diagonal consists of variances σ_j^2. It can be stated that:

$$-D_{\mathbb{KL}}(q_\phi(z|x)||p(z)) = \frac{1}{2}\sum_{j=1}^{J}(1 + \log(\sigma_j^2) - \mu_j^2 - \sigma_j^2), \tag{3.9}$$

where J is the dimension of the latent space z [10]. This means the $D_{\mathbb{KL}}$ from the ELBO Equation (3.8) can be calculated analytically.

Time modeling

One significant challenge in handling survival data is censorship, which occurs when a patient has not yet experienced the event of interest. In such cases, the survival time remains unknown, resulting in partial or incomplete observations. Consequently, SA models must employ techniques capable of reliably accommodating censored and uncensored observations to estimate relevant parameters.

In our case, to account for censoring in survival data, we start from the time t reconstruction term from Equation (3.8) for a single patient:

$$\hat{\mathcal{L}}_{time}(x_i, \theta_2, \phi) = \log p_{\theta_2}(t_i|g_\phi(x_i, \epsilon_i)). \tag{3.10}$$

Taking into account the censoring indicator d_i:

$$d_i = \begin{cases} 0 & \text{if censored} \\ 1 & \text{if event experienced} \end{cases}, \tag{3.11}$$

we could just use the information given by uncensored patients. However, we would waste information since we know that the censored patients have not experienced the event until

time t_i. Hence, considering Equation (3.1) and following [227], we model the time pdf as:

$$p_{\theta_2}(t_i|g_\phi(x_i,\epsilon_i)) = h(t_i|g_\phi(x_i,\epsilon_i))^{d_i} S(t_i|g_\phi(x_i,\epsilon_i)). \tag{3.12}$$

Therefore, the hazard function term is only considered when the event has been experienced, and the data are not censored. This way, SAVAE incorporates information from censored observations, providing consistent parameter estimates.

Regarding the distribution chosen for the time event, we have followed several publications such as [228], where the Weibull distribution model is used. This distribution is two-parameter, with positive support, that is, $p(t) = 0, \forall t < 0$. The two scalar parameters of the distribution are λ and α, where $\lambda > 0$ controls the scale and $\alpha > 0$ controls the shape as follows:

$$\begin{cases} p(t;\alpha,\lambda) = \frac{\alpha}{\lambda}\left(\frac{t}{\lambda}\right)^{\alpha-1} \exp\left(-\left(\frac{t}{\lambda}\right)^{\alpha}\right) \\ S(t;\alpha,\lambda) = \exp\left(-\left(\frac{t}{\lambda}\right)^{\alpha}\right) \\ h(t;\alpha,\lambda) = \frac{p(t;\alpha,\lambda)}{S(t;\alpha,\lambda)} = \frac{\alpha}{\lambda}\left(\frac{t}{\lambda}\right)^{\alpha-1} \end{cases} . \tag{3.13}$$

Although the Weibull distribution is our primary choice for modeling time-to-event data in SAVAE, it is crucial to highlight that other distributions are feasible as long as their hazard functions and CDFs can be analytically calculated. This versatility distinguishes SAVAE from other models. For example, the exponential distribution, a particular case of Weibull with $\alpha = 1$, can represent constant hazard functions. Integrating alternative distributions, such as the exponential, into SAVAE is straightforward and only requires adjusting the terms in eq. (3.12). The ability of SAVAE to predict the distribution parameters for each patient facilitates the calculation of various statistics, such as means, medians, and percentiles, providing flexibility beyond the models customized to a single distribution.

Marginal log-likelihood computation

Assigning distribution models to patient covariates in the reconstruction term is essential in SAVAE. This choice enables control over the resulting output variable distribution, but it also implies that the model approximates the chosen distribution even if the actual distribution differs. The third component of the ELBO (3.8) depends on the log-likelihood of the data, which for some representative distributions is:

- **Gaussian distribution:** Suitable for real-numbered variables ($x_i^c \in (-\infty, +\infty)$), it has parameters $\mu \in (-\infty, +\infty)$ and $\sigma \in (0, +\infty)$, known for its symmetric nature. Its log-likelihood function is:

$$\log(p(x_i^c;\mu,\sigma)) = -\log(\sigma\sqrt{2\pi}) - \frac{1}{2}\left(\frac{x_i^c - \mu}{\sigma}\right)^2. \tag{3.14}$$

- **Bernoulli distribution:** Applied to binary variables ($x_i^c \in \{0,1\}$), it has a single parameter $\beta \in [0,1]$, representing the probability of $x_i^c = 1$.Its log-likelihood function is:

$$\log(p(x_i^c;\beta)) = x_i^c \log(\beta) + (1 - x_i^c)\log(1-\beta). \tag{3.15}$$

- **Categorical distribution:** Models discrete variables with κ possible values. We can think of x_i^c as a categorical scalar random variable with κ different values. Each possible outcome is assigned a probability θ_κ (note that $\sum_{\kappa=1}^{K} \theta_\kappa = 1$). The log-likelihood function can be computed based on the Probability Mass Function (PMF) following the expression:

$$\log(p(x_i^c|\theta_1, \theta_2, ..., \theta_\kappa)) = \log\left(\prod_{\kappa=1}^{K} \theta_\kappa^{\mathbb{I}(x_i^c=\kappa)}\right), \tag{3.16}$$

where the indicator function means:

$$\mathbb{I}(x_i^c = \kappa) = \begin{cases} 1 & x_i^c = \kappa \\ 0 & x_i^c \neq \kappa \end{cases}. \tag{3.17}$$

Recall that other desired distributions can be implemented in SAVAE if their log-likelihood is differentiable.

Handling Missing Data

SAVAE has the particular capability of managing incomplete data without propagating the influence of missing values during training. To achieve this, we generate a binary mask matrix $Mask \in 0, 1^{Cov}$, where Cov is the number of covariates. The mask has 0s in positions corresponding to missing values and 1s elsewhere. While missing values are imputed to avoid issues during NN training, the mask ensures these values do not contribute to the reconstruction loss. Specifically, the log-likelihood terms corresponding to missing values are set to zero during the loss calculation:

$$\log p_{\theta_1}(x_i|z) = \sum_{c=1}^{Cov} Mask_c \cdot \log p_{\theta_1}(x_i^c|z), \tag{3.18}$$

where $Mask_c$ is the mask for the c-th covariate of observation x_i, and $\log p_{\theta_1}(x_i^c|z)$ is the reconstruction term for that covariate. This approach ensures that the model does not propagate invalid information from missing values while utilizing the available information effectively.

By leveraging this masking strategy, SAVAE becomes particularly well-suited for incomplete datasets, addressing a common limitation in SA, as explored in models like HI-VAE [123].

3.2.2 Overview of State-of-the-Art Comparative Models

To evaluate the performance of the proposed SAVAE model, we compare it against three widely recognized models in SA: CoxPH, DeepSurv, and DeepHit. These models were selected for their prominence in the field, representing a spectrum of methodologies ranging from classical statistical approaches to advanced DL frameworks. Each offers distinct strengths while presenting limitations our proposed model aims to address. By benchmarking against these models, we demonstrate the ability of SAVAE to overcome key challenges in SA, such as handling non-linear relationships, high-dimensional data, and complex interactions between covariates.

CoxPH

The CoxPH model [45], a seminal work in SA, is a semi-parametric method that estimates the relationship between covariates and the hazard function while assuming proportional hazards. Its interpretability, robustness, and simplicity make it a cornerstone in SA. However, as discussed previously in Section 2.2.2, its reliance on linearity in covariate effects and the proportional hazards assumption limits its applicability in scenarios involving high-dimensional, non-linear, or time-varying covariate relationships.

The model is trained by maximizing the partial likelihood function, simplifying the estimation process by avoiding explicit specification of the baseline hazard $h_0(t)$. The negative log partial likelihood, often used as the cost function, is given by:

$$\mathcal{L}_{CoxPH}(\beta) = -\sum_{i=1}^{N} d_i \Big[\beta^T x_i - \log \sum_{j \in Risk(t_i) e^{\beta^T x_j}} \Big], \tag{3.19}$$

where $Risk(t_i)$ is the risk set, i.e., the set of individuals still under observation at time t_i.

This formulation allows the model to leverage the observed event times while handling censored data effectively.

DeepSurv: Extending CoxPH with Non-Linear Modeling

DeepSurv [59] represents a significant evolution in SA by integrating DL to overcome the limitations of CoxPH. Specifically, DeepSurv addresses the restrictive assumption of linearity by employing a feed-forward NN to model complex, non-linear interactions between covariates and survival outcomes.

In DeepSurv, the hazard function is expressed as:

$$h(t|x) = h_0(t) \cdot e^{h_\theta(x)}, \tag{3.20}$$

where $h_\theta(x)$ is the log-risk function parameterized by the weights θ of the neural network. The network replaces the linear log-risk function from CoxPH with a non-linear mapping, enabling the model to capture intricate relationships without requiring manual feature engineering.

The training objective minimizes the regularized negative log partial likelihood:

$$\mathcal{L}_{DS}(\theta) = -\frac{1}{N_{d=1}} \sum_{i:d_i=1} \left(h_\theta(x_i) - \log \sum_{j \in Risk(t_i)} e^{h_\theta(x_j)} \right) + \lambda \|\theta\|_2^2, \tag{3.21}$$

where $N_{d=1}$ is the number of events observed in the dataset, $Risk(t_i)$ denotes the risk set for the i-th individual, and $\lambda\|\theta\|_2^2$ is the L_2-regularization term to prevent overfitting.

DeepSurv leverages modern DL techniques, including dropout, adaptive optimizers like Adam, and advanced activation functions like scaled exponential linear units to enhance training stability and predictive performance.

One notable feature of DeepSurv is its treatment recommendation system, which estimates the differential risk associated with different treatment options. For a given patient with

covariates x, the treatment-specific hazard function is:

$$\lambda(t; x|\tau = i) = \lambda_0(t) \cdot e^{h_i(x)}, \tag{3.22}$$

where $h_i(x)$ represents the log-risk for treatment i. The recommender function compares the risks of two treatments i and j:

$$rec_{ij}(x) = h_i(x) - h_j(x). \tag{3.23}$$

Positive values indicate a preference for treatment j, while negative values suggest treatment i is more favorable.

DeepHit: A Probabilistic Approach to Survival Analysis

DeepHit [64] offers a paradigm shift in SA by directly modeling the distribution of survival times using a DNN. Unlike CoxPH and DeepSurv, which focus on hazard functions, DeepHit estimates the joint probability distribution $P(t|x)$, enabling it to capture complex, non-linear, and time-dependent relationships between covariates and survival outcomes.

The architecture of DeepHit comprises a shared sub-network that extracts latent features from covariates x and outputs a probability distribution $y = [y_s]$, where y_s represents the estimated probability of the event occurring at time s. The survival function is then derived as follows:

$$S(t|x) = \prod_{s=1}^{t} (1 - y_s), \tag{3.24}$$

where $S(t|x)$ gives the probability of surviving beyond time t. This approach allows DeepHit to operate in the discrete-time domain, facilitating the modeling of complex datasets without relying on parametric assumptions.

The training objective of DeepHit is a composite loss function $\mathcal{L}_{DH}$, combining two terms:

1. **Likelihood Loss ($\mathcal{L}_1$)**: This term accounts for observed events and censored data:

$$\mathcal{L}_1 = -\sum_{i=1}^{N} \Big[\log(y^{(i)}_{s^{(i)}}) \cdot 1(d_i = 1) + \log(S(s^{(i)}|x_i)) \cdot 1(d_i = 0) \Big]. \tag{3.25}$$

2. **Ranking Loss ($\mathcal{L}_2$)**: This term ensures that patients are ranked correctly based on their predicted risk:

$$\mathcal{L}_2 = \sum_{i \neq j} A_{k,i,j} \cdot \eta\Big(S(t|x_i), S(t|x_j)\Big), \tag{3.26}$$

   where $A_{k,i,j}$ is an indicator for comparable pairs of patients, and η is a convex loss function that penalizes incorrect rankings.

Suitability for Comparative Evaluation

To effectively evaluate the performance of the proposed model SAVAE, it is essential to select comparative benchmarks representing various methodologies in SA. The models chosen for

this study—CoxPH, DeepSurv, and DeepHit—each hold a significant position in the field and bring unique strengths and weaknesses that highlight key aspects of the performance landscape in SA.

The selected models encompass a range of methodological approaches, from classical statistical methods to modern DL architectures. CoxPH is a foundational statistical model that provides a baseline for comparison and illustrates the transition from traditional to ML-based methods. DeepSurv bridges the gap by combining the theoretical underpinnings of CoxPH with the flexibility of deep neural networks, offering insights into the advantages of incorporating non-linear modeling into SA. DeepHit represents a paradigm shift, focusing on directly estimating survival time distributions and relaxing many assumptions inherent in traditional models.

This diversity ensures a comprehensive evaluation of the capabilities of SAVAE across various modeling paradigms. However, each model presents theoretical and practical challenges that emphasize the contributions of SAVAE:

- **CoxPH**: While robust and interpretable, the restrictive assumptions of CoxPH limit its flexibility in real-world applications. The log-risk function assumes a linear relationship between covariates and the log hazard, which limits its ability to capture complex, non-linear interactions. It serves as a benchmark for evaluating the benefits of DL enhancements.
- **DeepSurv**: While DeepSurv significantly improves over CoxPH, it has certain drawbacks. Despite its ability to model non-linear relationships, DeepSurv inherits the proportional hazards assumption from CoxPH. This assumption may limit its flexibility in scenarios where the effects of covariates change over time, reducing its effectiveness in modeling dynamic risk profiles.
- **DeepHit**: While DeepHit is a highly flexible model capable of capturing non-linear and non-proportional hazard relationships, it introduces specific challenges. Its reliance on a discrete-time framework, while simplifying optimization, requires datasets with many observations to maintain time granularity. This constraint can limit its applicability in real-world scenarios with sparse or irregularly spaced data. Additionally, including the ranking loss $\mathcal{L}_2$, which directly trains a surrogate metric like the C-index, introduces potential biases in performance evaluation, as it optimizes for a specific metric rather than a generalizable loss function.

By selecting CoxPH, DeepSurv, and DeepHit as comparative benchmarks, we establish a well-rounded framework to assess the contributions of SAVAE to SA. These models represent the state-of-the-art and bring into focus the specific limitations and challenges that SAVAE seeks to overcome, ensuring a rigorous and meaningful evaluation.

3.2.3 Experiments and Results

Evaluation Metrics

The evaluation of the proposed SAVAE model will rely on the metrics previously introduced in Section 2.2.5, namely the C-index and the IBS. These metrics are widely recognized in

SA for assessing model performance regarding discrimination and calibration. In addition to these core metrics, we will also incorporate two supplementary evaluation methods.

To statistically assess the performance of each model based on the global C-index, we propose the Mean Reciprocal Rank (MRR) as the third metric. It measures the effectiveness of a prediction by considering the rank of the first relevant C-index within a list composed of the C-indices obtained from each model. Formally, the Reciprocal Rank (RR) for a set of results for each model is the inverse of the position of the first pertinent result. For example, if the first relevant result is in position 1, its RR is 1; if it is in position 2, the RR is 0.5; if it is in position 3, the RR is approximately 0.33, and so on. Thus, the MRR is the average of the RRs for a set of models:

$$MRR = \frac{1}{Q}\sum_{i=1}^{Q}\frac{1}{rank_i}, \tag{3.27}$$

where Q is the total number of models being compared, and $rank_i$ is the position of the first relevant C-index for the $i-th$ model. Higher MRR values indicate that relevant results appear higher in the list.

Additionally, to add more statistical information on the performance of the models, we performed hypothesis testing to compare the mean C-index and IBS values of our model with those of the state-of-the-art models in multiple folds since we are using a five-fold cross-validation method. Specifically, we formulated a null hypothesis that assumes that the mean performance metrics of the state-of-the-art models are more significant than the mean performance metrics from our model. To assess the validity of this null hypothesis, we used p-values as a statistical measure. We established a significance threshold of 0.05, a common practice in hypothesis testing. When the obtained p-value for each case fell below this threshold, we rejected the null hypothesis. In practical terms, this indicated that our model exhibited superior performance compared to the other models. On the contrary, if the p-value exceeded 0.05, we concluded that there were no statistically significant differences between our model and the others. It is important to note that this approach considered variations in results across different folds, providing a more comprehensive assessment of model performance beyond just the average results. Given the multiple hypothesis tests performed, we acknowledge that the Family-Wise Error Rate (FWER) [229] increases as the number of tests grows, as established in the literature [230], [231]. This increase in FWER raises the risk of Type I errors, where false positives may occur due to the accumulation of multiple tests. To address this issue, we have applied an appropriate method to control this inflation, the Holm adjustment [232]. The results of these adjustments can be seen in Appendix C.1.4, ensuring the robustness of our findings.

Finally, we performed a sensitivity analysis to assess the robustness of our model and understand how variations in the input data influence its predictions. This analysis provides insights into the impact of individual features on the performance of the model and contributes to a better understanding of the decision-making process of the model. Furthermore, we analyzed the computational complexity of our model by comparing its runtime with state-of-the-art models. This analysis considers the time required for training and validating SAVAE across multiple datasets, providing insights into its efficiency relative to other methods. The findings illustrate that while the computational demands in using SAVAE are higher due to its

complex architecture, they remain manageable, making it suitable for practical applications. The detailed results of these analyses can be found in Appendix C.1.1 and Appendix C.1.3.

By combining these metrics, we aim to evaluate the predictive capabilities of SAVAE comprehensively, benchmarking its effectiveness against state-of-the-art models in standard and CR settings.

Implementation Details

The implementation of SAVAE was executed using the PyTorch framework [233]. As defined in Section 3.2.1, three different DNNs were trained, consisting of one encoder and two decoders. These decoders were designed to infer covariates and time parameters, respectively. The Gaussian encoder exhibits a straightforward architecture characterized by a single hidden linear layer featuring a Rectified Linear Unit (ReLU) activation function and an output linear layer with hyperbolic tangent activation. The input to this encoder consists of the covariate vectors from the training dataset, while the output generates a Gaussian latent space. The dimensionality of this latent space has been fixed to 5. The generated latent space is input for both decoders, featuring two linear layers each. The first layer employs a ReLU activation function and incorporates a dropout rate of 20%. However, the final layer of the decoders employs different activation functions based on the specified distribution, thereby tailoring the output to the parameters of the respective covariate distribution. Furthermore, the number of neurons in each hidden layer was also fixed at 50. The training process involved 3,000 epochs with a batch size of 64 samples while incorporating an Early Stop mechanism in case of an insufficient reduction in validation loss.

We conducted an ablation study to understand the behavior of SAVAE better and justify the selection of the defined hyperparameters. This study analyzed how changes in key hyperparameters, such as latent space dimensionality, number of neurons, and dropout rates, affect model performance. We could identify settings that balance performance and computational efficiency by systematically varying these parameters. The results of this ablation study are provided in Appendix C.1.2, offering further insights into the rationale behind our chosen hyperparameter configuration.

We used a five-fold cross-validation technique to evaluate the results while ensuring their robustness against data partitioning. This method was applied to our model and the state-of-the-art models used for performance comparison and result evaluation, including Cox-PH, DeepHit, and DeepSurv. Moreover, due to the inherent sensitivity of VAE architectures to initial conditions, we conducted training using up to 10 different random seeds. Subsequently, the C-index was averaged among the three best-performing seeds. The average performance of the three seeds provides a representative and sufficient evaluation. Lastly, note that the three state-of-the-art models have been implemented using the Pycox package [60] and the different metrics used for validation, C-index, and IBS. The MRR has been calculated manually, while the p-value has been obtained using the SciPy [234] package.

Results for Survival Analysis

This section proceeds with the experimental results achieved for SAVAE and the benchmark models. The data used for these experiments are detailed in Appendix B and include WHAS, Support, GBSG, FLChain, NWTco, Metabric, PBC, STD, and Pneumon. The data and the code can be found at https://github.com/Patricia-A-Apellaniz/savae.

We present a comprehensive assessment of the performance of our proposed model, SAVAE, compared to three well-established state-of-the-art models. Cox-PH, DeepSurv, and DeepHit. Across multiple datasets encompassing a diverse range of medical and clinical scenarios, we conducted extensive experiments to assess the performance of these models. The key focus was evaluating their ability to predict survival outcomes, considering censored and uncensored data points.

As the initial set of results, we focus on comparing the performance and results in terms of the C-index. Table 3.1 provides a comprehensive view of how our model is completely comparable to the state-of-the-art models regarding the average C-index. Additionally, note that all intervals for the minimum and maximum values across various folds overlap, indicating consistent performance across different data subsets. The results displayed in the table reveal that our model consistently achieves a higher MRR compared to others across multiple datasets, showcasing its superiority in many cases regarding the average C-index. However, it is essential to acknowledge that the C-index results among the different models are generally similar, highlighting the competitiveness of our model within the field. Furthermore, it is important to note that the broad intervals are primarily attributed to the limited sample sizes commonly found in medical databases. This characteristic poses challenges when assessing model performance. To address this issue, we employed cross-validation, as previously mentioned, ensuring that the performance of our model is robust and reliable. In summary, while our model demonstrates its strength by outperforming other models in terms of MRR and achieving competitive average C-index scores, the overall similarity in C-index results underscores its robustness and suitability for various medical datasets.

In our validation process, we performed a statistical analysis using p-values to determine whether our model exhibited superior performance in terms of the C-index. To carry out this analysis, we compared the average C-index of our model with the mean C-index values obtained from multiple folds for each state-of-the-art model. The objective was to determine whether the performance of our model was statistically better than the alternative models. We established a significance threshold of 0.05, a common practice in hypothesis testing. Our findings in Table 3.2 reveal several instances in which our model outperformed the state-of-the-art models, as evidenced by p-values below the 0.05 threshold. These results highlight the effectiveness and competitiveness of our proposed approach. This comprehensive analysis, which considers the diverse C-index values in multiple folds, provides a robust evaluation of the performance of the model, extending beyond simple average comparisons.

Our validation through IBS values (Table 3.3 and Table 3.4) yielded conclusions that closely parallel those derived from the C-index analysis. Overall, it is essential to note that IBS results from our model align closely with the state-of-the-art models, demonstrating comparable performance. However, our proposed model consistently demonstrated competitiveness and

Dataset	COXPH		DEEPSURV		DEEPHIT		SAVAE	
	Avg. C-index	(min, max)	Avg. C-index	(min, max)	Avg. C-index	(min, max)	Avg. C-index	(min, max)
WHAS	0.74	(0.66, 0.81)	0.78	(0.57, 0.88)	**0.89**	(0.82, 0.95)	0.74	(0.67, 0.80)
Support	0.58	(0.39, 0.78)	0.57	(0.37, 0.82)	0.55	(0.37, 0.73)	**0.61**	(0.40, 0.86)
GBSG	0.66	(0.61, 0.71)	**0.67**	(0.58, 0.73)	0.66	(0.58, 0.72)	**0.67**	(0.62, 0.72)
FLChain	0.69	(0.50, 0.80)	0.67	(0.55, 0.80)	0.78	(0.73, 0.82)	**0.79**	(0.75, 0.83)
NWTco	0.71	(0.64, 0.79)	0.70	(0.60, 0.79)	**0.72**	(0.66, 0.78)	0.71	(0.63, 0.79)
Metabric	0.59	(0.52, 0.68)	**0.61**	(0.52, 0.69)	0.56	(0.46, 0.64)	**0.61**	(0.53, 0.70)
PBC	**0.81**	(0.64, 0.94)	0.80	(0.65, 0.92)	0.80	(0.62, 0.93)	**0.81**	(0.62, 0.95)
STD	**0.60**	(0.47, 0.72)	**0.60**	(0.49, 0.71)	0.59	(0.50, 0.68)	0.59	(0.46, 0.71)
Pneumon	0.62	(0.54, 0.70)	0.65	(0.49, 0.80)	**0.67**	(0.57, 0.77)	0.65	(0.53, 0.77)
MRR	0.56		0.60		0.62		**0.76**	

Table 3.1: C-index average results across different folds for each state-of-the-art model. Average C-index results across the three best seeds for each fold in SAVAE performance. MRR values are given to rank each model, which attends only to the mean value. **Bold** highlights the best mean. For C-index and MRR, higher is better.

Model	WHAS	Support	GBSG	FLChain	NWTco	Metabric	PBC	STD	Pneumon
CoxPH	0.579	0.058	**0.000**	**0.000**	0.268	**0.003**	0.450	0.887	**0.003**
DeepSurv	1.000	**0.020**	0.149	**0.000**	0.135	0.549	0.280	0.927	0.382
DeepHit	1.000	**0.000**	**0.000**	**0.001**	0.644	**0.000**	0.228	0.727	0.935

Table 3.2: p-values obtained to determine whether the mean of SAVAE is greater than the state-of-the-art folds C-indexes. Bold implies a p-value below our threshold, 0.05. This means that SAVAE is significantly better than the other models.

emerged as the top performer in the various datasets used in our study. This convergence of results across different evaluation metrics reinforces the robustness and effectiveness of our novel approach. While our model maintains a competitive edge within the context of the state-of-the-art models, further solidifying its potential and utility in the field of SA, it also stands out as a top-performing solution.

It is essential to recall that, like DeepSurv and Cox-PH, SAVAE is a parametric model. However, unlike these models, we do not limit ourselves to the exponential distribution to model survival time. Our approach allows for the use of any differentiable distribution. Unlike DeepHit, which trains the model using loss functions, our framework uses likelihood functions, providing considerable flexibility. We specifically assumed the Weibull distribution for these experiments, deriving the shape parameter α and the scale parameter λ for each patient, although any differentiable distribution could have been used. This ability enables us to extract vital statistical information for personalized patient treatments, offering a significant advantage in medical applications.

Dataset	COXPH		DEEPSURV		DEEPHIT		SAVAE	
	Avg. IBS	(min, max)	Avg. IBS	(min, max)	Avg. IBS	(min, max)	Avg. IBS	(min, max)
WHAS	0.171	(0.109, 0.279)	0.134	(0.067, 0.260)	**0.120**	(0.067, 0.175)	0.159	(0.114, 0.205)
Support	0.208	(0.074, 0.374)	**0.205**	(0.057, 0.363)	0.219	(0.086, 0.370)	0.208	(0.063, 0.385)
GBSG	0.182	(0.142, 0.223)	0.179	(0.137, 0.228)	0.208	(0.168, 0.248)	**0.179**	(0.139, 0.222)
FLChain	0.137	(0.089, 0.185)	0.142	(0.088, 0.186)	0.121	(0.098, 0.145)	**0.102**	(0.078, 0.124)
NWTco	**0.107**	(0.080, 0.138)	0.109	(0.082, 0.149)	0.111	(0.083, 0.147)	0.127	(0.101, 0.152)
Metabric	0.186	(0.137, 0.233)	0.191	(0.143, 0.244)	0.214	(0.153, 0.275)	**0.180**	(0.127, 0.236)
PBC	0.147	(0.043, 0.281)	0.146	(0.046, 0.268)	0.195	(0.087, 0.340)	**0.138**	(0.034, 0.267)
STD	0.210	(0.121, 0.302)	0.212	(0.123, 0.305)	0.224	(0.142, 0.315)	**0.209**	(0.121, 0.307)
Pneumon	**0.016**	(0.004, 0.031)	0.017	(0.004, 0.034)	**0.016**	(0.004, 0.031)	0.021	(0.007, 0.037)
MRR	0.55		0.55		0.47		**0.71**	

Table 3.3: IBS average results across different folds for each state-of-the-art model. Average IBS results in the three best seeds for each fold in SAVAE performance. MRR values are given to rank each model. **Bold** highlights the best mean. For IBS, lower is better, and for MRR, higher is better

Model	WHAS	Support	GBSG	FLChain	NWTco	Metabric	PBC	STD	Pneumon
CoxPH	1.000	0.470	0.998	1.000	**0.000**	0.995	0.888	0.575	**0.000**
DeepSurv	**0.000**	0.341	0.561	1.000	**0.000**	1.000	0.868	0.746	**0.000**
DeepHit	**0.000**	0.950	1.000	1.000	**0.000**	1.000	1.000	0.995	**0.000**

Table 3.4: p-values obtained to determine whether the mean of SAVAE is greater than the state-of-the-art folds IBS values. Bold iImplies a p-value below our threshold, 0.05. This means that SAVAE is significantly better than the other models.

3.2.4 Conclusions

In this section, we have successfully described an SA model (SAVAE), which stands out for its ability to avoid assumptions that can limit performance in real-world scenarios. It is a model based on VAEs in charge of estimating continuous or discrete survival times, first, modeling complex non-linear relations among covariates due to the use of highly expressive DNNs, and second, taking advantage of a combination of loss functions that capture the censoring inherent to survival data. Our model demonstrates efficiency compared to various state-of-the-art models, namely Cox-PH, DeepSurv, and DeepHit, because of its freedom from linearity and proportional hazard assumptions. In contrast to DeepHit, which directly learns the C-Index metric, we train using standard likelihood techniques. Note that our approach is more flexible, as it allows using many different distributions to model the data. The performance is competitive, as it performs well in C-Index and IBS, instilling confidence in its capabilities.

Furthermore, the adaptability of our model is a notable strength. While we have assumed specific distributions for both survival times and covariates in our experiments, the versatility of SAVAE extends to accommodating any other parametric distribution, as long as their CDF and hazard function are differentiable, making it a versatile tool. Notably, our model can efficiently handle censoring to mitigate bias, introducing a novel improvement in results. However, it is essential to acknowledge that the reliance of the model on specific parametric distributions could pose limitations. The model may perform suboptimally if the chosen distribution does not align well with the underlying data distribution. This is a known challenge in parametric SA models, and further research could explore more flexible non-parametric or semi-parametric approaches to address this limitation [235].

This work raises several attractive lines for the future. Since the parameters estimated by SAVAE are subject to statistical uncertainty, we propose future work using Monte Carlo sampling from the latent space to derive Confidence Intervals (CIs) for survival predictions, providing more robust patient-wise survival curves with associated margins of error. An additional advantage lies in the architecture of our model, where time and covariates are reconstructed from latent space information. This feature opens opportunities for its utility to be expanded to various tasks that have been developed using VAEs, including clustering [236], imputation of missing data [237], and data augmentation [238] by the generation of synthetic patients, achieved in Section 4.2. Thus, this tool has great potential and can be exploited in future work to have different functionalities even in the world of Federated Learning [239], [240] (addressed in Section 5.4).

In summary, SAVAE emerges as a versatile and robust SA model, surpassing state-of-the-art methods while offering extensibility to a broader range of healthcare applications. It presents a compelling solution for healthcare professionals seeking enhanced performance and adaptability in SA tasks.

3.3 CR-SAVAE: Survival Analysis with Competing Risks

3.3.1 Methodology of the Proposed Model (CR-SAVAE)

CR refers to scenarios in SA where multiple mutually exclusive events can preclude the occurrence of the event of primary interest, as explained in Section 2.2.4. In this context, estimating the CIF for each risk is essential. The CIF, defined as:

$$CIF_k(t|x) = P(T \leq t, r = k|x), \tag{3.28}$$

represents the probability of experiencing an event of type k by time t, considering CR. Unlike single-risk SA, where $S(t|x)$ monotonically decreases to zero, the sum of all CIFs for CR does not necessarily equal one as $t \to \infty$. This distinction arises because CIFs focus on the incidence of each CR rather than the overall survival probability.

We extend the SAVAE architecture to the CR setting to handle these complexities, creating the CR-SAVAE model. CR-SAVAE integrates the CIF into its modeling framework, leveraging the Total Probability Theorem to estimate the survival function:

$$\begin{aligned} S(t|x) =& 1 - F(t|x) = 1 - P(T \leq t|x) \\ =& 1 - \sum_{k=1}^{R} CIF_k(t) \\ =& 1 - \sum_{k=1}^{R} P(T \leq t, r = k|x) \\ =& 1 - \sum_{k=1}^{R} P(T \leq t|r = k, x) \cdot P(r = k|x). \end{aligned} \tag{3.29}$$

To achieve this, CR-SAVAE uses a modified time reconstruction loss incorporating censoring. Specifically, for each patient, the model evaluates the likelihood of the observed time t_i based on whether the sample is censored ($r_i = 0$) or uncensored ($r_i \neq 0$). This is captured in:

$$\log p(t_i|z_i) = \mathbb{I}(r_i = 1) \log p(t_i|z_i) + \mathbb{I}(r_i = 0) \log S(t_i|z_i), \tag{3.30}$$

where $\mathbb{I}$ is the indicator function. For uncensored samples, the log-likelihood of t_i is computed directly, while for censored patients, the model considers the survival probability $S(t_i|z_i)$ at time t_i. This loss formulation ensures that the model appropriately handles both observed events and censoring, a critical requirement in SA.

In addition to this time reconstruction loss, Equation (3.30) is integrated with the CIF estimation to model CR accurately. Specifically:

1. $P(T \leq t|r = k, x)$ is estimated using separate decoders for each risk, trained with Equation (3.30).
2. The probability of risk $P(r = k|x)$ is jointly estimated with the latent variable z using the SAVAE architecture.

Information about censored data is incorporated by exploiting, and no event occurred before the censoring time t. This is reflected in Equation (3.31) for censored patients.

$$\log\left(S(t_i|x_i)\right) = \log\left(1 - \sum_{k=1}^{R} CIF_k(t_i)\right) \tag{3.31}$$

The CIF for each risk (Equation (3.32)) is then estimated by combining the conditional survival and risk probability.

$$CIF_k(t) = P(T \leq t|r = k, x) \cdot P(r = k|x) \tag{3.32}$$

To estimate the risk probability $P(r = k|x)$, CR-SAVAE uses a neural network classifier that inputs the latent representation z of the SAVAE architecture. The classifier outputs $p_\theta(r|z)$, representing the probability of each risk given the latent representation of the patient. The classifier assumes a categorical distribution and is trained using the log-likelihood of this distribution, considering only uncensored patients (Equation (3.33)).

$$\frac{1}{N}\sum_{i=1}^{N} \mathbb{I}(r_i \neq 0) \log p(r_i|z_i) \tag{3.33}$$

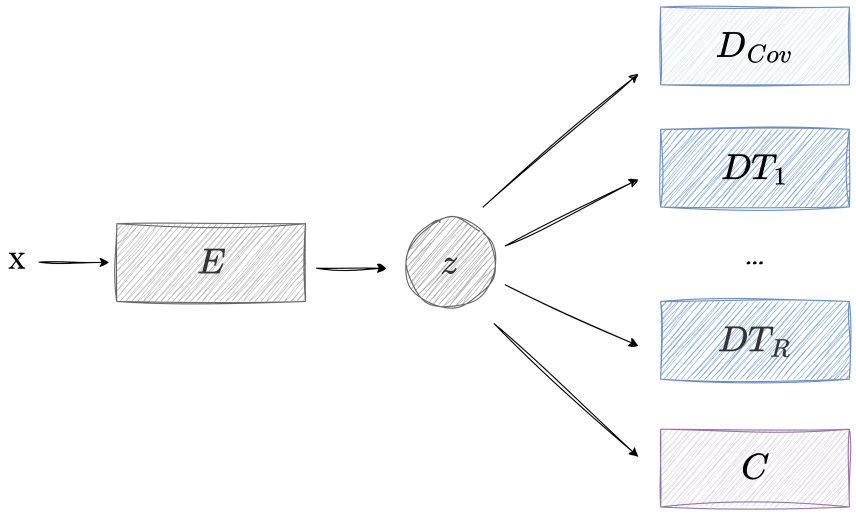


Figure 3.3: Schema comparing SAVAE and CR-SAVAE. Rectangles represent NNs, and shadowed elements are common to SAVAE and CR-SAVAE. E means Encoder, z is the latent space, D_{Cov} is the covariate decoder, DT_k is the time decoder for risk k estimation $P(T \leq t|r = k, x)$, and C is the classifier estimating $P(r = k|x)$. Note that SAVAE is CR-SAVAE with a single risk and without a classifier.

Figure 3.3 visually compares the architectures of SAVAE and CR-SAVAE, highlighting the additional components of CR-SAVAE to handle CR. Notably, SAVAE becomes a special case of CR-SAVAE in the single-risk setting, requiring only a single decoder for time and omitting the risk classifier.

Hence, the total loss for CR-SAVAE is:

$$\mathcal{L}_{CRS} = \\ -\frac{1}{N}\sum_{i=1}^{N}\Bigg[-D_{\mathbb{KL}}(q(z|x_i)||p(z)) + \sum_{c}\log p(x_i^c|z_i) + \sum_{k=1}^{R}\Big(\mathbb{I}(r_i=k)\log p(t_i|r_i=k,z_i)\Big) \\ +\mathbb{I}(r_i=0)\log\left(1-\sum_{k=1}^{R}CIF_k(t_i)\right) + \mathbb{I}(r_i\neq 0)\log p_\theta(r_i|z_i)\Bigg] \tag{3.34}$$

where we note that the differences to SAVAE lie in time estimation, as the KL term and the covariates reconstruction loss are identical to SAVAE loss in Equation (3.8). Finally, note that according to Equation (3.28), once we have estimated $P(T \leq t|r=k,x)$ and $P(r=k|x)$, we can estimate the CIF as in Equation (3.32), which will be needed to calculate the performance metrics.

3.3.2 Overview of State-of-the-Art Comparative Models

In the context of CR SA, we benchmark our proposed CR-SAVAE model against DeepHit [64], a prominent DL approach that has set the standard for CR modeling. DeepHit was selected for comparison due to its widespread recognition, ability to handle CR settings effectively, and demonstrated superiority over classical methods such as the FG sub-distribution hazard model. However, despite its strengths, DeepHit also presents significant limitations that our CR-SAVAE model aims to address.

DeepHit: A Deep Learning Framework for Competing Risks

DeepHit leverages DNNs to directly model the probability of CR over time without relying on the assumption of proportional hazards. The method estimates the conditional probabilities $P(T \leq t, r=k|x)$, where T is the time-to-event, r is the event type, and x represents the covariates. These conditional probabilities are used to construct the CIF for each CR k:

$$CIF_k(t|x) = P(T \leq t, r=k|x). \tag{3.35}$$

The architecture of DeepHit consists of a shared sub-network to capture global features across all risks and separate risk-specific sub-networks to learn unique features for each event type. The output is a discrete probability distribution over time points for each CR, from which the CIF is calculated by summing the probabilities up to the desired time.

The loss function employed by DeepHit in the CR setting is the same as that of its single-risk variant.

While DeepHit represents a major advancement in CR SA, it faces the following limitations:

1. **Nonparametric CIFs:** DeepHit provides numerical estimates of CIFs without analytical expressions, hindering precise statistical computation and interpretation. This limitation makes it challenging to derive additional statistical measures or insights directly from the model outputs.

2. **High Variability in CIF Curves:** The estimated CIFs can vary significantly between patients, reducing the reliability of the model in clinical applications requiring consistent decision-making predictions.
3. **Discrete-Time Representation:** As in the single-risk setting, DeepHit operates exclusively in the discrete-time domain. This approach can lead to information loss or inaccuracies when working with continuous-time data, which is common in medical datasets.
4. **Reliance on C-index Optimization:** By focusing on a surrogate metric like the C-index in its loss function, DeepHit prioritizes ranking performance over directly optimizing the likelihood of the observed data. This can lead to a mismatch between the optimization objective and the ultimate goal of accurate survival prediction.
5. **Bias in CR Estimation:** DeepHit estimates CIFs directly, which is crucial for CR SA. However, the lack of a parametric framework limits its flexibility in producing robust predictions. Other models, such as the one in [241], focus on survival functions instead of CIFs, which can introduce bias in CR settings [242].

DeepHit is one of the few DL models explicitly designed for CR SA, making it a natural choice for comparison. Its ability to model complex, non-linear relationships between covariates and event outcomes highlights its strength as a baseline. However, the limitations outlined above, particularly the lack of a parametric framework and its reliance on discrete-time representation, underscore the need for more flexible and robust approaches.

By comparing CR-SAVAE with DeepHit, we aim to demonstrate the unique contributions of our model. CR-SAVAE directly estimates parametric CIFs, supports both continuous and discrete time-to-event data, and avoids the pitfalls associated with surrogate metrics like the C-index. This comparison highlights the potential of CR-SAVAE to address key challenges in CR SA, advancing the state of the art in survival modeling.

3.3.3 Experiments and Results

Evaluation Metrics

In these experiments, conducted in the CR setting, we evaluated the performance of CR-SAVAE using the time-dependent C-index and the IBS, both adapted to CR as previously defined in Section 2.2.5. These metrics assess the ability of the model to predict survival outcomes, accounting for multiple CR accurately. These metrics, tailored for the CR setting, enable a robust comparison of CR-SAVAE and DeepHit, capturing their predictive performance in scenarios where patients are subject to multiple CR.

Implementation Details

We used 5-fold cross-validation for every dataset, with 80% of the data to train the algorithms and the other 20% to obtain the results in the article. For CR-SAVAE, we modeled the time using a Weibull distribution. This means the decoder outputs the scale parameter $\lambda > 0$ and the shape parameter $\alpha > 0$ each time. The log-likelihood and survival function for the

Weibull distribution are:

$$\begin{aligned} \log p(t) &= \log(\alpha) - \log(\lambda) + (\alpha - 1)\log\left(\frac{t}{\lambda}\right) - \left(\frac{t}{\lambda}\right)^{\alpha} \\ \log S(t) &= -\left(\frac{t}{\lambda}\right)^{\alpha} \end{aligned} \tag{3.36}$$

CR-SAVAE is a parametric method where the time distribution can be changed to any other distribution as long as the log-likelihood and the survival function are differentiable. Although we have chosen the Weibull distribution because it does not assume proportional hazards, other distributions like Log-Normal, Inverse Gaussian, and more can be used. This highlights the advantage of CR-SAVAE: being fully parametric, we can change the time distribution to the one desired and obtain parameter estimates for each patient.

As a concrete example with the Weibull time distribution, after training CR-SAVAE, we can input the covariates of a patient x_i and obtain (α_i, λ_i) for each of the k risks, as well as θ_k, the probability of each risk for that concrete patient. Then, we can compute the analytical, personalized CIF as $CIF_k(t) = (1 - e^{(-t/\lambda_i)^{\alpha_i}})\theta_k$. Since CR-SAVAE estimates the distribution parameters, we can obtain the CIF curve and any statistic that depends on these parameters, yielding personalized predictions for each patient.

Regarding the VAE parameters, we use a latent dimension of 5 and hidden layers of 64 neurons. We trained using minibatches of 32 patients and opted for the Adam optimizer with a learning rate of $1e-3$. We train using a maximum of 2000 iterations, but if the validation loss does not improve in 30 epochs, we early stop the training.

Results

The datasets used for these experiments are detailed in Appendix B and include Melanoma, MGUS2, and EBMT. A more detailed explanation of the datasets can be found in Appendix B. Additionally, the code is available in https://github.com/Patricia-A-Apellaniz/cr-savae.

The results of CR-SAVAE compared to DeepHit can be seen in Table 3.5, where we report the average metric across folds plus the standard deviation. For each triplet of the dataset, risk, and metric, we ran an unequal variance T-test [243] to check whether the means of DeepHit and our model were significantly different, and we found no significant evidence of this (i.e., p-values higher than 0.01). This means that CR-SAVAE is a viable alternative to DeepHit in the CR SA setting since it provides similar metrics and is also a parametric method, such as [241]. Since DeepHit is non-parametric, it may give higher metrics results than a parametric alternative, as highlighted in [64]. Parametric methods are considered to provide good results when they match the performance of DeepHit, as shown in [241]. Thus, our proposed CR-SAVAE offers promising results that match the performance of a non-parametric method such as DeepHit, with all the advantages of being parametric: it facilitates hypothesis testing, CI estimation, and other statistical analyses. DeepHit lacks this interpretability.

Risk	Method	C-index	IBS
1	**CR-SAVAE**	0.6318 ± 0.0731	0.2415 ± 0.0291
1	**DeepHit**	0.6660 ± 0.0303	0.2929 ± 0.0198
2	**CR-SAVAE**	0.6320 ± 0.0963	0.3220 ± 0.1432
2	**DeepHit**	0.6848 ± 0.1243	0.1876 ± 0.0469

(a) Melanoma dataset

Risk	Method	C-index	IBS
1	**CR-SAVAE**	0.6369 ± 0.0446	0.3403 ± 0.0972
1	**DeepHit**	0.5919 ± 0.0478	0.2686 ± 0.0617
2	**CR-SAVAE**	0.5672 ± 0.0427	0.3505 ± 0.0270
2	**DeepHit**	0.6359 ± 0.0174	0.1865 ± 0.0823

(b) MGUS2 dataset

Risk	Method	C-index	IBS
1	**CR-SAVAE**	0.5645 ± 0.0333	0.2459 ± 0.0231
1	**DeepHit**	0.5700 ± 0.0290	0.2570 ± 0.0257
2	**CR-SAVAE**	0.5737 ± 0.0612	0.2592 ± 0.0184
2	**DeepHit**	0.5248 ± 0.0263	0.2489 ± 0.024
3	**CR-SAVAE**	0.5867 ± 0.0960	0.1987 ± 0.0135
3	**DeepHit**	0.7106 ± 0.1197	0.1647 ± 0.0155
4	**CR-SAVAE**	0.5729 ± 0.0567	0.2006 ± 0.0176
4	**DeepHit**	0.7108 ± 0.1246	0.1811 ± 0.0229
5	**CR-SAVAE**	0.7411 ± 0.1111	0.2000 ± 0.0092
5	**DeepHit**	0.6007 ± 0.0796	0.1649 ± 0.0219
6	**CR-SAVAE**	0.5519 ± 0.0214	0.2363 ± 0.0200
6	**DeepHit**	0.5946 ± 0.0408	0.2683 ± 0.0208

(c) EBMT dataset

Table 3.5: Datasets results in SA with CR. Results obtained comparing DeepHit and CR-SAVAE. Data is average ± standard deviation. A two-sided unequal variance T-Test [243] comparing the means of every metric and the risk yields a p-value higher than 0.01, which means that neither of the two methods tested is significantly better than the other.

3.3.4 Conclusions

CR-SAVAE, a parametric CR SA model, demonstrates performance comparable to DeepHit, a widely used non-parametric model, on metrics like the C-index and iBS. However, the parametric nature of CR-SAVAE offers significant advantages. CR-SAVAE supports rigorous statistical analysis, including hypothesis testing, CI estimation, and other procedures essential for research and decision-making. Moreover, our model accurately estimates the CIF and robustly handles both continuous and discrete time data. This flexibility expands its potential in real-world healthcare settings.

Furthermore, its parametric structure also offers computational advantages compared to non-parametric approaches. Future research aims to leverage the VAE architecture of CR-SAVAE for broader healthcare applications. Potential directions include using the latent space for patient clustering and synthesized patient data generation (addressed in Section 4.2), further expanding the utility of the model beyond traditional SA. Furthermore, investigating explainability using standard techniques such as SAGE [244] would significantly improve the interpretability of the model, a critical priority in artificial intelligence in healthcare.

3.4 Chapter Conclusions

This chapter introduced two novel models, SAVAE and its extension CR-SAVAE, designed to address the limitations of existing SA techniques. Rooted in DL advances and grounded in a generative framework based on VAEs, these models offer significant flexibility and robustness for predicting time-to-event outcomes in medical and other domains.

SAVAE brings a generative perspective to SA, modeling time-to-event data with high precision and flexibility. Unlike traditional models, SAVAE avoids proportional hazard assumptions and leverages neural networks to model complex, non-linear relationships between covariates and survival times. This approach ensures adaptability across various distributions, supporting both continuous and discrete time-to-event modeling. Furthermore, SAVAE is trained using standard likelihood techniques rather than surrogate metrics like the C-index, making it a versatile and theoretically sound tool for survival predictions.

Extending this innovation to the CR setting, CR-SAVAE introduces critical capabilities for handling multiple potential outcomes that compete for occurrence. By directly estimating the CIF, CR-SAVAE empowers clinicians with patient-specific survival parameters, facilitating precise and personalized predictions. Unlike state-of-the-art models, CR-SAVAE handles both continuous and discrete survival times efficiently, avoiding the limitations of discretization seen in models like DeepHit. Moreover, it provides a parametric and flexible framework that can model diverse survival times and covariate distributions without relying on proportional hazard assumptions.

The performance of both SAVAE and CR-SAVAE has been rigorously evaluated using real-world datasets encompassing diverse covariate types, including clinical and genomic data. Comparative experiments with traditional models like CoxPH and DL approaches, such as DeepSurv and DeepHit, demonstrate that SAVAE and CR-SAVAE achieve competitive results regarding the C-index and IBS. Importantly, these models provide interpretable and robust predictions while overcoming critical gaps in current SA methodologies.

In conclusion, SAVAE and CR-SAVAE contribute to advancing the field of SA by offering generative, flexible, and accurate tools tailored to contemporary challenges in survival prediction. These models bridge gaps in single-risk and competing-risk settings and lay the groundwork for future research in DL-based SA, particularly in domains demanding personalized and precise SA.