APRIL 2024

ARTIFICIAL DATA IN HEALTHCARE: ANALYSIS AND AREAS FOR CONSIDERATION

White paper coordinated by:

Pr. ALLASSONNIÈRE Stéphanie

Dr. FRAYSSE Jean-Louis

Université Paris Cité

BOTdesign

AKNOWLEDGEMENTS

This book was a collective endeavor in its writing. We extend our heartfelt thanks to:

Prof. Stéphanie ALLASSONNIÈRE, Professor of Applied Mathematics and Vice-President fo Valorization and Industrial Partnerships, Paris-Cité University, Associate Director and PR[AI]RIE Chair, Co-founder of Sonio
Ms. Manon DE FALLOIS, Assistant to the Chief of Health Services, CNIL
Dr. Stanley DURRLEMAN, Research Director at Inria and the Paris Brain Institute, Co-founder and CEO of Qairnel (la clinique du Docteur Mémo), PR[AI]RIE Chair
Dr. Fabrice FERRÉ, PhD, Anesthesia and Critical Care Specialist, Toulouse University Hospital
Mr. Marco FIORINI, CEO, Artificial Intelligence & Cancers Association
Dr. Jean-Louis FRAYSSE, Co-founder, BOTdesign
Mr. David GRUSON, Founder, ETHIK-IA
Dr. Jérôme KALIFA, President et Founder of Let it Care
Dr. Yann Maël LE DOUARIN, Medical Advisor at the DGOS and Head of the Health and Digital Transformation Department, Ministry of Labor, Health and Social Affairs
Pr. Marie-France MAMZER, Professor of Medicine and Hospital Practitioner, Paris-Cité University
Mr. Thierry MARCHAL, President, Secretary General, Avicenna Alliance, Chief Technologist Health, EMEA, Ansys, Co-founder of Biomed In Silico France, Member of EMA and eHealth Expert Groups, European Commission
Dr. Jean-Baptiste MASSON, Researcher at the Pasteur Institute, PR[AI]RIE Chair, Co-founder of AVATAR MEDICAL
Dr. Hervé NABARETTE, Deputy Director of Public Affairs, AFM-Téléthon
Dr. Raphaëlle PARKER, Principal Scientist, The Janssen Pharmaceutical Companies of Johnson & Johnson
Prof. Raphaël PORCHER, Professor of Medicine and Hospital Practitioner, Paris-Cité University, PR[AI]RIE Chair
Dr. Camille SCHURTZ, Head of Regulatory Processes and Market Access, Agence de l'Innovation en Santé
Dr. Sylvie TROY, Deputy Medical Director and Real-World Data/Real-World Evidence Lead France, Pfizer
Dr. Vinh-Phuc LUU, Head of Epidemiology and Innovation Division, Artificial Intelligence & Cancers Association
Dr. Sarah ZOHAR, Research Director, Inserm, Head of the "HeKA" Research Team - Inserm, Inria, Paris-Cité University

For this diligent, friendly, and collaborative work.

We also wish to thank France Asso Santé, Dr. Lise Alter (AIS), Ms. Gaëlle Bouvet (BOTdesign), Ms. Corinne Collignon (HAS), Mr. Florent Della Valle (CNIL), Mr. Jérôme Lang (IRIT), Ms. Virginie Lasserre (Janssen), Ms. Floriane Pelon (HAS), Prof. Isabelle Ryl (PR[AI]RIE), Prof. Antoine Tesnières (PariSanté Campus), Mr. Olivier Thuillart (BOTdesign), and Mr. Félicien Vallet (CNIL) for their valuable contributions.

AKNOWLEDGEMENTS

GLOSSARY

AI (Artificial Intelligence): Artificial intelligence is a set of mathematical models and theories, as well as programming techniques, aimed at creating machines capable of mimicking some tasks attributed to human intelligence.

AI Act: The AI Act (Artificial Intelligence Act) is a regulation aimed at regulating and promoting the development and commercialization of AI systems in the European Union.

Artificial Cohorts: Data generated by algorithms or automated processes.

Classification: Statistical method for assigning a group label to unlabeled data.

Clustering: Statistical method that allows partitioning data into homogeneous subgroups.

CNIL (National Commission for Information Technology and Civil Liberties): It assists professionals in their compliance efforts and helps individuals control their personal data and exercise their rights.

Control arm or placebo arm: In a clinical trial, a group of participants who are subjected to the new method under study aimed at preventing, detecting, treating, or controlling the disease.

Data Anonymization: Method rendering it impossible to identify a person from a dataset.

Data Augmentation (DA): In the field of artificial intelligence, the data augmentation process increases the quantity of training data by creating new data from existing data.

Data Protection Impact Assessment (DPIA): Tool for building a GDPR-compliant and privacy-respecting process. It concerns the processing of personal data that may result in a high risk to the rights and freedoms of data subjects.

Data Pseudonymization: Data pseudonymization is a data protection technique that involves preprocessing data in such a way that it is not possible to attribute them to a specific person without additional information. Specifically, it involves replacing real personal identifiers (names, first names, emails, addresses, phone numbers, etc.) with pseudonyms.

Deep Learning: Deep learning is a machine learning process that uses neural networks with multiple layers of neurons. These algorithms have a very large number of parameters, requiring a large amount of data to train.

GLOSSARY

Deterministic: Relating to determinism, the philosophical doctrine according to which all events are linked and determined by the chain of previous events.

Digital Medical Device (DMD): Medical devices are integral to medical care, those that integrate a digital function can generate a large amount of real-life data and pave the way for more personalized medicine.

Digital Twins: A Digital Twin is a virtual replica of a physical object.

EDPB (European Data Protection Board): Independent supervisory authority of the European institutions (e.g., the European Commission) on data protection.

EFPIA (European Federation of Pharmaceutical Industries Associations): Its missions are to promote pharmaceutical research and development in Europe, as well as to create a favorable economic, regulatory, and policy environment to meet health needs and increasing patient expectations.

Electronic Health Record (EHR): A file stored electronically containing information about a patient's health and care throughout their life.

EMA (European Medicines Agency): Contributes to protecting and promoting human and animal health by evaluating and controlling medicines within the European Union (EU) and the European Economic Area.

Extended Reality (XR): Represents all technologies that create computer-generated environments and objects. This includes augmented reality (AR), mixed reality (MR), or virtual reality (VR).

FDA (Food and Drug Administration): American institution responsible for overseeing food and drug safety. It authorizes the marketing of pharmaceutical products in the United States.

Federated Learning: A learning paradigm in which multiple entities collaboratively train an AI model without pooling their respective data.

France 2030: The "France 2030" plan, endowed with 54 billion euros over 5 years, aims to develop industrial competitiveness and future technologies, with half of the funding allocated to emerging actors and half to decarbonization efforts.

GLOSSARY

GAN based (Generative Adversarial Networks): In neural network-based learning, Generative Adversarial Networks, sometimes also called Generative Adversarial Networks, are a class of learning algorithms that generate data through a competition between a generator proposing the most relevant data possible and a discriminator that seeks to separate real data from generated data.

Generalizability: Property of mathematical models allowing them to generalize their results to a class of observations they have not previously encountered during their calibration phase.

GDPR (General Data Protection Regulation): European regulatory text that regulates data processing equally throughout the territory of the European Union (EU).

HAS (Haute Autorité de Santé): HAS promotes good practices and the proper use of care among users. It participates in public information and improves the quality of medical information.

HDLSS Data (High Dimension Low Sample Size): A set of data with a very low number of observations compared to the dimension of its attributes. Example: an abdominal CT imaging database with a cohort of 50 patients, each represented by an image with 100,000 pixels.

Human Oversight or Human Guarantee: Human Oversight or Human Guarantee is a label recognized at the French, European, and even international levels. The principle of Human Guarantee ensures the ethical development of artificial intelligence contributing to health, by establishing points of human oversight throughout their evolution.

ICH (International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use): An organization that brings together regulatory authorities and pharmaceutical industry representatives to standardize regulations for the registration of human-use drugs.

Inference (statistics): Set of mathematical and algorithmic methods used to infer the characteristics of a general group from those of an observed sample.

IRBs (Institutional Review Boards): They are responsible for issuing prior opinions on the validity conditions of any research involving human subjects, in accordance with the criteria defined by Article L 1123-7 of the Public Health Code.

GLOSSARY

ISO 13485 Standard: ISO 13485:2016 specifies requirements for a quality management system when an organization needs to demonstrate its ability to consistently provide medical devices and related services that meet customer requirements and applicable regulatory requirements.

K-Nearest Neighbors: Classification method based on the proximity of an observation to its nearest neighbors, transferring their characteristics to it.

LEEM (Les Entreprises du Médicament): This is a French professional pharmaceutical syndicate and lobby that replaced the Syndicat national de l'industrie pharmaceutique (SNIP) in 2002.

LLM (Large Language Model): Large language models based on estimating very large neural networks that mimic the probabilistic relationships of words between them in sentences or paragraphs.

Machine Learning (ML): Mathematical methods for calibrating models from data observations. These models give computers the ability to learn from data to solve tasks defined by the user without being explicitly programmed.

Marketing Authorization (MA): Permission granted to a holder of exploitation rights for an industrially manufactured drug to market it.

MD (Medical Device): A health product intended by its manufacturer to be used for diagnostic, preventive, monitoring, treatment, or alleviation purposes of a disease or injury.

Metaverse: Virtual space in which one interacts completely immersively, using avatars.

Mixed Reality (MR): Technology that adds virtual elements to what we see and allows physical interaction with these elements. MR is a blend of augmented reality and virtual reality. For its use, you will need to use an Augmented Reality headset, knowing that the user's position is calculated in real time, allowing interaction with virtual elements through gestures or controllers.

MRI (Magnetic Resonance Imaging): Set of techniques used to obtain images from nuclear magnetic resonance.

Multimodal Data: Multimodal data refer to datasets containing multiple modes or sources of data, such as text, audio, video, and images.

Navier-Stokes Equations: Equations of fluid mechanics describing the motion of gases or most liquids.

GLOSSARY

Neural Networks: Also known as artificial neural networks (ANN) or simulated neural networks (SNN), they are a subset of machine learning and are at the core of deep learning algorithms. Their name and structure are inspired by the human brain, mimicking how biological neurons send signals to each other.

NSCLC (Non-Small Cell Lung Cancer): This is the most common type of lung cancer, accounting for 85-90% of all lung cancers.

Omniverse: A set of supposed coexisting universes.

Outlier: An observation that is distant from the rest of the observed population.

Parametric Equations: In mathematics, a parametric equation is an equation with a finite number of unknowns to estimate. In contrast to a non-parametric equation where the number of degrees of freedom is therefore infinite.

Phase 1 (clinical trial): Phase I trials usually correspond to the first administration of a drug to humans. Phase I/II trials are a variant of Phase I trials, allowing for a preliminary evaluation of efficacy at the selected dose or testing drug combinations.

Phase 2 (clinical trial): Phase II trials aim to confirm the preliminary clinical and/or pharmacological activity of the drug at the recommended dose following Phase I. A limited number of patients are included in these trials (typically 40 to 80). Some Phase II trials compare two treatments. The duration of a Phase II trial is generally two to three years, depending on the selected pathology and the number of patients.

Phase 3 (clinical trial): Comparative trials are intended to compare the new drug to a standard treatment to determine its effectiveness. Phase III trials include several hundred to several thousand patients and usually last at least four to five years, depending on the pathology and expected effect.

Phase 4 (clinical trial): After their commercialization, drugs continue to be subject to strict long-term monitoring, known as post-marketing surveillance, to identify any serious and/or unexpected side effects due to their administration. This is referred to as pharmacovigilance. Phase IV trials may also be intended to evaluate this newly approved drug under different administration conditions, such as administration frequency, number of courses, duration of infusion, etc.

PK-PD: Pharmacokinetic and pharmacodynamic modeling (integral part of the drug development process).

Real-World Data (RWD): Real-world data come from multiple sources and reflect the "real" daily lives of patients and physicians: care practice, management, or the impact of disease and treatment on daily life.

GLOSSARY

Segmentation: Delimitation of the boundary of objects in an image.

Singleton: In mathematics, a set consisting of a single element.

SoC (Standard of Care): Treatment recognized by medical experts as the most appropriate for a certain type of disease in a particular context and widely used by healthcare professionals. Also called best practices, standard medical care, best available therapy, and standard therapy.

SVM (Support Vector Machine): Clustering method based on creating boundaries between populations from a small number of pivot observations.

Synthetic Cohorts: Real data derived from the aggregation of previously collected data.

VAE (Variational Auto-Encoders): Mathematical model allowing the encoding of data into a representative space of small dimension as well as the decoding of points from this space, thus generating new observations.

Virtual Cohorts: Data stored, processed, or exchanged in digital form.

Voxel: Tri-dimensional Pixel (Picture Element) (Volume Element).

PREFACE

It is our collective responsibility to use all ethical methods to ensure the health of our current and future fellow citizens. It is our personal ethics to embrace any technology that may accomplish this mission, even if we do not yet understand it or if it threatens our familiar way of working. Some may tell us not to rush; but can we slow down when patients are dying? However, we cannot rush and risk putting these patients in danger. Therefore, we must enthusiastically welcome all constructive criticism and ensure satisfactory responses.

The first quarter of this twenty-first century has seen the emergence and development of many digital methods such as predictive statistics, numerical modeling and simulation, and therefore artificial intelligence. They are widely used to drastically improve the safety of cars, airplanes, and other power plants. Why should it be any different for health?

Just a thousand years ago, Avicenna, a Persian physician and scientist, published "The Canon of Medicine"; he was the first to suggest testing any new treatment on a group of patients while keeping a control group: the first clinical trials. Over the past ten centuries, medicine has cautiously but systematically adopted new technologies to reduce the risks faced by these human guinea pigs and thus accelerate medical innovation. These progressive attitudes have led to major revolutions, often appearing in our countries, such as modern surgery, pharmacy, and vaccines. We are on the brink of a new medical revolution that will quickly lead us to personalized health.

We have an immense amount of data that we can now exploit by developing more relevant algorithms. We have the wisdom to adopt regulations for the protection of personal data and we establish protocols to establish the validity and reliability of this data. Far beyond the experience and memory of the individual physician, it is now possible to compare a new situation with the vastness of collective memory and predict the likely evolution of a patient via data-driven artificial intelligence. The mastery, albeit relative, of the physics, chemistry, and biology of the human body offers us increasingly efficient numerical models to predict the evolution of a pathology via knowledge-driven artificial intelligence. Artificial intelligence combines these approaches, data and knowledge, not to replace doctors and nurses, but to assist them in routine tasks and memory and enable them to devote themselves to developing the best treatment for the patient, the quintessence of human genius.

As explained in this white paper, our researchers are also developing multiple approaches to augment these data and models, anonymize them to respect patients, with the aim of testing any new treatment on potentially infinite cohorts of patients, now virtual, long before traditional clinical trials begin. It would be absurd to deprive ourselves of these so-called 'in silico' clinical trials, in contrast to 'in vivo' tests conducted on cohorts of virtual patients, which may include extreme patients, rare diseases, or minorities (children, pregnant women, ethnic minorities, etc.). In silico clinical trials will thus provide very useful information, more quickly and at lower cost, drastically reducing risks during traditional clinical trials.

Therefore, we thank the authors of this work for this remarkable synthesis and implore the authorities to give it their full attention, in order to keep a door wide open while maintaining a critical mindset to challenge these approaches. Thus, our fellow citizens will be able to fully and quickly benefit from new treatments, and our healthcare industry will enjoy regulation allowing them to test their medical innovations quickly and effectively without having to export themselves to regions more open to these digital methods.

We are in a global race where we must act swiftly, but cautiously.

Thierry Marchal
President, Secretary-General, Avicenna Alliance
Chief Technologist Health, EMEA, Ansys
Co-founder of Biomed In Silico France
Member of the Expert Groups EMA and eHealth, European Commission

CONTENTS

CONTEXT

01

Research and development in the service of health are accelerating, marked by a rapid and continuous increase in discoveries, technological advancements, and therapies. The challenge is to build a development plan that demonstrates the value of new technologies and their contribution compared to the existing arsenal, thereby allowing real innovations to be detected and made available.

Several factors contribute to this acceleration:

- Collaboration and data sharing on a global scale, facilitating the rapid circulation of information and ideas;
- Significant increase in computing power;
- Use of artificial intelligence (AI);
- Funding and investments within the framework of France 2030 in particular;
- Special access pathways to reimbursement aimed at speeding up the availability of health technologies while ensuring the establishment of evidence necessary for assessing their value and contribution to the care arsenal. HAS (Haute Autorité de Santé) is conducting work, in particular, to define acceptable methodological conditions for early access to medicines;
- Research and clinical trials evolving to become more efficient.

Furthermore, our healthcare system must face significant structural challenges (aging population, increase in chronic diseases, inequalities in access to care, etc.).

In its publication "Clinical Trials 2030" in March 2022, the professional organization of pharmaceutical companies (LEEM) highlights the challenges, perspectives, and challenges to ensure that France remains competitive globally in the market launch of new molecules, which can be extended to medical devices (MD). The growing use of personalized medicine, the difficulty in recruiting and retaining patients in studies, the increasing complexity of trial methodology, increased competition with other countries, and the cost of implementing these clinical studies raise questions about adapting structures and systems so that France continues to participate in the development of tomorrow's molecules.[1] [2]

1 "Clinical trials provide initial access to innovation for patients, particularly those with serious conditions such as cancer (42% of initiated trials), as well as autoimmune diseases (18%) and central nervous system disorders (13%)." Source: LEEM.

2 While the randomized double-blind trial remains the gold standard to demonstrate the efficacy of a drug and should be favored, the HAS introduces the possibility of integrating less consolidated data provided they allow comparison with available treatments. Indeed, only comparison enables an assessment of the added value of a new treatment. The objective is to enable reimbursement access for immature products while maintaining an acceptable level of quality standards. The new evaluation doctrine of the CT thus opens up to indirect comparison data of good methodological quality or data from control groups, provided they are explained and justified in advance by the manufacturer.

To address these challenges, the rise of digitalization and AI allows us to envisage, in the near future, the implementation of new protocol schemes and clinical trials that will complement the methodological arsenal by using, for example, real-world data or artificial patient arms. Indeed, by leveraging real medical data collected in the context of care, prevention, or research, Machine Learning (ML) tools enable the generation of virtual or artificial data from real medical data. Once validated, these tools could potentially compensate for the difficulties in recruiting real patients into control arms by using artificially generated patients.

This is referred to as augmented cohorts, which consist, on one hand, of real patients and, on the other hand, of artificial patients. The data of these artificial patients are generated from scratch by Machine Learning models based on real data from the cohort, derived from patients recruited for the respective study. These augmented cohorts are distinct from so-called synthetic cohorts (such as synthetic control arms), which are composed of "recycled" real data from patients included in previous cohorts. These cohorts must be reliable, representative of the source patients, and the latter must not be re-identifiable.

So in this document, we will distinguish between (1) synthetic cohorts, which involve data collected previously, reused, and collated to create a new cohort, and (2) artificial cohorts, which increase the number of real patients recruited for the study to be conducted.

Artificial data can offer other benefits for research in the healthcare field by allowing experiments to be conducted, such as the development of new analysis methods or training and/or testing of AI models, without using sensitive data as they do not belong to real patients.

Thus, artificial data could contribute to facilitating the development of models to:

- Assist healthcare professionals in diagnosing;
- Help personalize treatments (therapeutic decision-making);
- Identify patient populations sensitive to a given treatment;
- Support medical education (simulations for learning and training);

- To simulate the spread of epidemics and test prevention and control strategies;
- To test the robustness and security of computer systems and digital applications (simulate potential attacks and security vulnerabilities);
- To implement clinical trials with artificial patients to evaluate the effectiveness and safety of new molecules or medical devices.

For ethical reasons related to the confidentiality of sensitive health data, and thus the protection of privacy, especially by the General Data Protection Regulation (GDPR) within the European Union (EU), access to patient data necessary for the development of these models is very complex and limited. Artificial data, generated from real data in highly secure environments, but not originating from individuals, could be easily made available to researchers on more accessible platforms.

The reasoned use of artificial data in the healthcare domain would offer valuable benefits while limiting certain risks associated with the use of real data. Their "routine use" requires rigorous validation by experts (healthcare professionals, mathematicians, patients) to ensure their reliability, i.e., their ability to faithfully reproduce real data. Their use indeed implies mastering their specific risks and defining the possible scopes of use.

Currently, there are no recommendations from agencies or regulatory authorities defining the acceptability criteria for artificial patient cohorts for the evaluation of health products or medical devices, as the very concept of using artificial patient cohorts for this purpose is not universally agreed upon. Experiences of using such cohorts in studies submitted to agencies are, at this stage, limited.

By reading this white paper, we aim to provide a better understanding of the process of creating artificial patients, the various possible fields of application, and the limitations of these patient cohorts, but above all, to describe the validation processes required to ensure the reliability and representativeness of these patients of a new kind. This validity is essential before considering that this data could be used by health authorities for the market authorization of health products, for their evaluation in reimbursement processes, but also to provide guarantees to users of these healthcare technologies.

Innovation only has value if it translates into benefits for users, and we hope that this white paper will contribute to that.

Happy reading.

The implementation of artificial patient cohorts through generative AI confirms the effectiveness of collaboration between public and private teams in the fields of research, medicine, ethics, regulators, patient associations, and industry. These agile and structured collaborations accelerate the validation and deployment of highly innovative technologies within the framework of France 2030.

Prof. Stéphanie Allassonière
Professor of Applied Mathematics and
Vice President for Valorization and Industrial Partnerships,
University of Paris-Cité,
Deputy Director and PR[AI]RIE Chair,
Co-founder of Sonio

SCIENTIFIC JUSTIFICATIONS

We now invite you to delve into the heart of the subject.

Data augmentation (DA) is the art of increasing the size of a dataset of interest by creating annotated artificial data with the same characteristics as the original population, without reproducing the real data identically. This is what we call the "generalizability" of models.

This method also allows for the rebalancing of the number of samples per class by oversampling minority classes. This helps address cases where a subpopulation is not adequately represented in a study, such as vulnerable populations (pregnant women, children, patients with rare diseases, etc.).

In addition to data collected in real-life situations, where recruitment is difficult, digitally generating so-called artificial data can be a lever. These artificial data are information or datasets created numerically to simulate characteristics and structures similar to those of real data.

They are generated using algorithms implementing mathematical models. There are two types of models.

The first are called mechanistic models. These mathematical models consist of concatenating and interacting known equations from mechanics, chemistry, or other disciplines that describe a phenomenon. These equations are deterministic in that there is no randomness. This is the case with the Navier-Stokes equations, nonlinear partial differential equations that describe the motion of Newtonian fluids. These equations generate outputs that are mostly physical quantities characterizing the studied phenomenon. In the case of clinical trials, such methods are employed by Nova Discovery, for example, which, through detailed research in the literature of the equations involved in a pharmacological process, is able to mimic the results of a clinical trial. However, this may involve many equations with many parameters for which it is important to find pivotal values in the literature.

The other mathematical models used are based on statistical models. Their goal is to propose equations mimicking patterns observed in real data or representing the probability distribution of these data in a mathematical space. These approaches can use mechanistic equations, such as PK-PD equations, i.e., pharmacokinetic (Pharmaco-Kinetic) and pharmacodynamic (Pharmaco-Dynamic) modeling, but add randomness allowing for great flexibility, especially to take into account population and individual scales in the same model. These are then called mixed-effects models. They are preferred when there is no equation governing the studied phenomenon (for example, the height of children between 0 and 20 years old is not generated by any equation, yet it is very well described in health records).

These approaches, based on so-called Generative Artificial Intelligence models, enable resampling: after a calibration phase of these models (learning), the user can generate new data, different from the training data, but reproducing the observed and captured population characteristics. These models can involve a more or less complex level of explicit or implicit equations (for example, using a neural network). Such methods are employed by Quinten, for example.

From these models, we are able to describe a phenomenon based on the sample of a population that is observed. They provide another perspective of use: allowing the creation of new observations not seen in the initial sample but retaining the same characteristics. These models are called generative. The creation of additional data is called data augmentation.

There are more generally several techniques to generate artificial data. They can be divided into three levels.

The use of AI in the field of healthcare presents particularly complex challenges inherent to the types of data it relies on (sensitive, sparse, heterogeneous, often limited quantities, etc.). However, research efforts in recent years that aim to address these specific challenges now allow us to envision dizzying potential, particularly in the field of clinical studies, offering promises of studies that are not only cheaper but also fairer, more inclusive, and more effective. The use of artificial data for the development of new drugs is a particularly bold and sensitive ambition that requires a common and sustained effort from diverse and complementary stakeholders to progress in the most efficient manner possible.

Dr. Raphaëlle Parker
Principal Scientist,
The Janssen Pharmaceutical Companies of Johnson & Johnson

I. A single patient's data producing one or more similar data points

An example is simple transformations applied to images, such as adding noise (random modifications of pixel grayscale levels), blurring (convolution), zooming, or deformations like translations and/or rotations. Then, the label of the original image is assigned to the created images.

Although these augmentation techniques have proven very useful, they remain highly dependent on data and thus limited. Some transformations may not be informative or even introduce biases. For instance, consider a digit representing a 6 that becomes a 9 when rotated too much, or a 4 that might resemble a 9 if the shape is "blurred" too much (see figure below). Evaluating the relevance of augmented data becomes increasingly challenging with the complexity of the original data and may require expert intervention to assess the degree of relevance of the proposed transformations. This evaluation can also be technically challenging, as expertise faces limitations (think of acceptable deformations of a healthy liver that are difficult to characterize). Another health-related example would be a nodule that could be mistaken for a mixed structure such as a hemorrhagic cyst if it were too noisy.

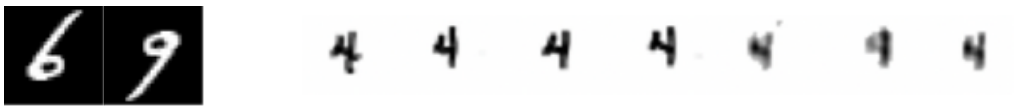


Fig. 1. Examples of error-prone modifications.

These techniques are highly favored because they are simple and very effective for imaging. However, they are more challenging to apply to other types of data, especially tabular data or genetic data.

II. A patient's data producing an artificially modified data by utilizing their identified neighbors within a population: noise added through averaging the neighbors.

This technique was brought to light by the company Octopize, which proposed a method called "avatarization."

Note: Initially intended for anonymizing data, this method could also be considered for creating artificial data. In this case, it would involve producing not just one data point per observation, but a small number for each. However, generating more data could potentially undermine the anonymizing aspect of the produced cohort. This question remains open to date.

The principle is to represent the data in a mathematical space where local averages can be computed. Quantitative tabular data are already naturally in such a space. For qualitative data (i.e., categorical data resulting from multiple-choice options) or more complex data (such as images), it is necessary to correctly define the notion of local average. Indeed, averaging two images considered as arrays of numbers makes no sense, as it would produce a blurred image with twice the elements present (think of the example of averaging two faces producing an image with four eyes, two noses, and two mouths). For images, it is necessary, for example, to resort to notions of deformable models [69] generalizing the notion of average to shapes contained within images. Once this notion of local average is mathematically defined, each data point will then be modified by the influence of k of its nearest neighbors. The number of neighbors used (k) defines the level of noise.

Besides the fact that these methods may require a complex mathematical model, they are sometimes challenging in the case of small cohorts where each individual is a "singleton," i.e., unique compared to others. Moreover, outliers from larger populations (data points isolated from the rest of the observations) may also disappear due to this phenomenon, which tends to concentrate the population around "average" data points.

To address this issue, oversampling methods, aiming to increase the number of samples of minority classes and to account for their learning difficulty, have been developed. The synthetic minority oversampling technique (SMOTE), introduced for the first time by the team of D. K. Iakovidis [6], involves interpolating data points belonging to minority classes in their feature space.

This approach has been extended in other works where the authors proposed to oversample (produce new, unobserved samples) near the decision boundary (the area where subgroups are separated by the algorithm) using various classification algorithms (either the k-Nearest Neighbor algorithm [7] or a Support Vector Machine [8]), thus focusing on samples that are potentially misclassified to refine the method. Other methods have also been proposed [9], [10], however, these are difficult to adapt to high-dimensional data [11], [12].

Note: The anonymization solution proposed by Octopize was evaluated by the CNIL services, which concluded that there were no obstacles to demonstrating compliance with European criteria for data anonymization provided adequate parameterization is in place. Additionally, an update of the European guidelines on anonymization is currently being rewritten by the "Article 29" working group.

III. A population of patients characterized globally by identifying its features and replicated by creating new individuals sharing them (including digital twins).

i. Mechanistic

The use of properties from real systems, known as mechanistic, which include physical, chemical, and biological knowledge, allows for the prediction of organ properties and facilitates medical decision-making. Cardiac simulations [21-27] represent a major achievement in this field. These equations are often parametric, and the choice of parameters greatly influences the quality of the model. As these parameters often represent physical quantities, it becomes possible to calibrate these models by comparing their results to reality, using observed organ data, as well as data from the literature. Once calibrated, these models provide a twin of an observation. Often referred to as a digital twin, it also applies in the random framework presented below. It provides a digital representation of the organ, the patient, or a phenomenon (such as emergency department flow).

Many simulated mechanistic twins are suitable for surgical applications involving various organs such as the liver or the brain, and can also include the physical properties of surgical materials used in the surgical interventions themselves. Extensions of these simulations using mixed reality (XR) are rapidly emerging, both in surgical planning procedures with virtual reality and in the operating room with augmented reality.

The major initiatives of large technology companies could extend to medical applications. The Metaverse [38] alludes to medical applications through partnerships and commercial announcements (mainly regarding surgical operations), and in the shorter term, most likely, the Omniverse [39], which merges approximate physical simulations, advanced rendering, and numerous initiatives from the Monai [40] to provide a suite of medical applications.

ii. Statistical learning models without Deep Learning

1) ABC methods: data simulation according to a model gradually calibrated by comparison to reality

Approximate Bayesian Computation (ABC) methods, also known as likelihood-free methods, have emerged over the past thirty years as the most satisfying approach to problems with incalculable likelihoods.

They offer an almost automatic solution to the difficulties encountered with models that are complex but can be simulated from. They were first proposed in population genetics by Tavaré et al. (1997), who introduced ABC methods as a technique to bypass the calculation of the likelihood function through simulation from the corresponding distribution of the generative statistical model. However, these methods suffered to some extent from calibration difficulties that made them rather unstable in their implementation and therefore not widely used. Various improvements and extensions to the original ABC algorithm have made these methods more robust and are still at the forefront of new research.

The principle is simple: a parametric generative model is constructed to mimic a phenomenon, meaning that parameters are simulated a priori and then observations are generated from them. The artificial data closest to the observed (real) data point to a set of better initial parameters than others. This is the calibration part. Then, the model is able to generate data according to this configuration.

2) Inferences based on simulations and damped inferences

Recently, ABC methods have been enhanced by advances in Machine Learning or automatic learning. Thus, inferences based on simulated data and damped inferences accelerate the analysis and generation of data.

The fundamental principle of these methods is the use of simulated data to create and optimize the procedure for generating virtual data. It requires having a generative model of data, even if it is approximate a priori, which will be optimized during the algorithm's learning process.

Extensions of these approaches will be fueled by recent progress in simulation-based inferences. Although they rely on similar principles to ABC approaches, they aim to facilitate the inference procedure by training neural networks on simulations so they can be run directly and more quickly on experimental data.

3) Atlas estimation methods

The specificity of atlas estimation methods lies in their focus on a premise: a population can be represented by a mean element around which data varies. Based on this assumption, populations are described at two levels: a population level, characterized by the mean and interindividual variance, and an individual level, characterized by each individual's specific variation relative to this mean. The means and variances (which can be complex quantities such as images, shapes, or regulatory networks for the means, and variations in shapes or structures for the variances) are unknown quantities that need to be estimated from observations. It no longer involves comparing a priori simulations to real data, but rather maximizing the likelihood of observations that will point to an optimal parameter (or set of parameters). This is the learning or calibration part of the model. Once this step is completed, the models can generate artificial data.

These techniques have been used extensively to estimate characteristic elements of populations as well as their "normal" variability, for data ranging from shapes and images[3] to longitudinal[4] processes.

---

3 www.deformetrica.org
4 https://disease-progression-modelling.github.io/pages/main.html

This has allowed the generation of large cohorts of artificial patients with various modalities (here's an example[5] where one million subjects were generated, each with cognitive scores, cortical and hippocampal atrophy, and metabolic dynamics). These advancements are at the core of the technology for early detection of memory-related diseases utilized by www.docteurmemo.fr [73].

Disease progression models generate artificial data, and it's now demonstrated that they enhance the statistical power of clinical trials when combined with observed real-world data.

Dr. Stanley Durrleman
Director of Research at Inria and the Paris Brain Institute,
Co-founder and CEO of Qairnel (Dr. Memo's clinic),
PR[AI]RIE Chair

[5] https://project.inria.fr/digitalbrain

iii. Deep Learning techniques

1) GAN based

The recent improvement in the performance of generative models such as Generative Adversarial Networks (GANs) [13] or Variational Autoencoders (VAEs) [14], [15] has made them very appealing for data analysis. GANs have already been widely used in various fields [16], [17], [18], [19], [20], including medicine [21]. For instance, GANs have been applied to magnetic resonance imaging (MRI) [22], [23], computed tomography (CT) [24], [25], radiography [26], [27], [28], positron emission tomography (PET) [29], mass spectrometry [30], dermoscopy [31], or mammography [32], [33], yielding promising results.

However, most of these studies focused on a fairly large training set (over 1,000 training samples) or on data of relatively low dimensionality, while in everyday medical applications, it remains very challenging to gather such large cohorts of labeled patients. Therefore, to date, the case of high-dimensional data combined with a very small sample size remains largely unexplored.

2) VAE (Variational Auto-Encoders) based

Compared to GANs, VAEs had attracted less interest for data augmentation and had been primarily used for speech applications [34], [35], [36]. The use of these generative models on medical data for classification tasks [37], [38] or segmentation [39], [40], [41] is growing. When not well controlled, these models can produce blurry and imprecise samples. This undesirable effect was even more pronounced when trained with a small number of samples, making them very challenging to use for data augmentation in the context of High Dimension Low Sample Size (HDLSS). Recent work demonstrates that VAEs can be used for data augmentation reliably, even in the HDLSS data context, provided that some modeling of the latent space is provided and the way data are generated is modified.

The ORIGA platform by BOTdesign harnesses the power of VAEs in partnership with Professor Stéphanie Allassonnière. Existing and ongoing publications support the relevance of using this technology in generating artificial data applied to healthcare.

3) Denoising diffusion

Diffusion models from a noise model use stochastic differential equations (SDE) and an iterative process of adding and removing noise to generate new images. The model essentially transforms the task of image generation into a denoising task through a diffusion process. Although medical applications are more recent [17-21], these models exhibit the same level of effectiveness as general image generation.

4) The future of using these data generation models

The large-scale digitization of patient records has paved the way for the creation of efficient virtual patient records. Various approaches have been used to synthesize these reports or Electronic Health Records (EHR) using primarily GANs [41-44], and more recently, diffusion models [45,46] and VAEs [47,48]. There is active debate [49,50] on the relevant characteristics to explore to ensure the medical utility of these generated data. When these data are generated in the context of longitudinal studies, many approaches [44,51-54] (largely based on the same three main architectures) have emerged in the past decade. They need to be related to statistical approaches aimed at modeling longitudinal data [55-57] to establish their accuracy. Beyond these initiatives, recent breakthroughs in text generation models [58-61], also known as Large Language Models (LLM), are expected to play an important role by providing artificial data with nearly weekly updates on the properties of these models and fine-tuning possibilities.

Most current generative models have limitations in that they mainly generate single types of clinical data for a patient (imaging, medical reports, more generally all EHR data). There are still few comprehensive approaches to multimodal synthesis that couple these different types of data, which would pose enormous challenges for evaluating their relevant properties. However, recent technical advancements adapted for medical analysis now integrate multimodal data. Med-PALM [62] is a multimodal generative model that codes and analyzes biomedical data, including clinical language, imaging, and genomics.

The JNF-VAE model [90] utilizes VAEs and normalizing flows to propose a representation of multimodal data as a single population and concurrently representations of each modality conditionally on any other combination of the others. Many initiatives [63-66] now incorporate multiple modalities into the data analysis process. It is likely that similar to the foundational model [67], these initiatives will soon dominate with the usual challenges associated with the propagation of unpredictable errors, privacy protection anomalies, and lack of data analysis.

Different methods can therefore be used to create artificial data. The typology of the initial data, the pursued objective, and the expertise of the teams will guide the choice of technology to use for data augmentation.

Artificial cohorts will provide the opportunity to extend the domain of simulation-based inferences to the medical field. Thus, methods developed on simulations can be tested and improved on real medical data.

Dr. Jean-Baptiste Masson
Researcher at the Pasteur Institute
PR[AI]RIE Chair
Co-founder of AVATAR MEDICAL

03

USE OF ARTIFICIAL PATIENT COHORTS

Toward regulatory framework for the use of in silico trials and virtual cohorts as clinical evidence

Dr. Sarah Zohar
Research Director, Inserm
Head of the Research Team "HeKA",
Inserm, Inria, Paris-Cité University

In the previous chapter, we described the possibility of augmenting imaging, tabular, and genetic data in a cross-sectional, longitudinal, or multimodal manner, which will be used for the various aforementioned applications.

We will address each of these three themes successively.

I. Using artificial data to accelerate clinical research and the market launch of healthcare products (drugs and Medical Devices (MD))

Clinical studies are scientific investigations aimed at documenting the effectiveness and safety of a drug or medical device.

Clinical studies are essential for evaluating a new product by health authorities, whether it's for market access or reimbursement, if applicable, but also to inform users. The possibility of conducting studies in France represents an opportunity at both the national and individual levels, with the experimental arm presumed to be as or more effective than the control arm. The control arm should systematically represent the "Standard of Care" (SoC), meaning the care administered in routine practice according to current guidelines or Good Clinical Practices (GCP); if the medical need is unmet, it may involve a placebo arm. These studies are regulated, subject to approvals, and adhere to ethical standards to avoid causing harm to patients.

The inclusion of a sufficient number of patients in the trial is one of the factors that ensures statistical power and the robustness of the results, although it does not guarantee the clinical relevance of these results. The expected effect size is also a key element in a protocol: the larger it is, the fewer patients are needed. Approximately 80% of clinical trials fail to recruit the required number of patients within expected timelines, and 55% are prematurely terminated due to patient recruitment issues.[6] Recruitment and retention difficulties in trials are particularly acute for various reasons.

In clinical research, the generation of artificial data could help with the relevant design of clinical trials and create artificial patient cohorts that could strengthen the control arms of phase 3 studies and/or patients included in phase 2.

[6] Desai M. Recruitment and retention of participants in clinical studies: Critical issues and challenges. Perspect Clin Res. 2020 Apr-Jun;11(2):51-53. doi: 10.4103/picr.PICR_6_20. Epub 2020 May 6. PMID: 32670827; PMCID: PMC7342339.

In clinical research, the generation of artificial data could assist in the relevant design of clinical trials and establish artificial patient cohorts that could strengthen the control arms of phase 3 studies and/or the patients included in phase 2.

Currently, trials involving synthetic patients (not yet artificial – see Glossary) have already been authorized by regulatory authorities (FDA or Food and Drug Administration, EMA or European Medicines Agency) for the implementation of clinical trials.

For example, Alecensa® (alectinib),[7] developed by Roche, indicated as monotherapy in the treatment of non-small cell lung cancer (NSCLC). This drug received conditional approval from the EMA in 2017. By establishing a synthetic active control arm (67 patients receiving ceritinib), the company was able to provide the EMA with evidence of efficacy compared to a standard treatment. The level of evidence was recognized as robust and allowed for the market release of alectinib 18 months earlier than if waiting for data from a "traditional" clinical study. The role of the synthetic control arm in obtaining conditional marketing authorization needs to be clarified.
The use of artificial data could play a similar role to this synthetic data, with the advantage of being temporally equivalent to the treated arm.

It is also conceivable that artificial data could one day complement active arms in phase 3, increasing the statistical power of comparative tests between two arms and the patients. In phase 4, artificial data could also help make weak signals more "visible."

Below, we have listed situations that seem to be use cases where artificial data could help overcome certain challenges. This non-exhaustive list already demonstrates the various issues that these artificial patients could address.

7 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7218288/#CIT0009

i. Uses

1) Facilitate the establishment of studies (facilitate patient recruitment)

In many situations, recruiting a sufficient number of patients necessary to conduct a large-scale study (phase 3) presents a real challenge. An obvious example is trials in the field of rare diseases, as patients are, by definition, rare. However, this situation is becoming more common with precision medicine practices, which aim to treat a very specific population of patients characterized by a set of biomarkers and thus defined by very strict inclusion and exclusion criteria. This recruitment difficulty is also encountered in pediatric trials, which always require very long recruitment phases.

Furthermore, generating artificial patients could also enrich a cohort with underrepresented patients, thereby increasing the diversity of these cohorts. This would involve populations that are difficult to access (certain ethnicities) or vulnerable (pregnant women, elderly individuals / polypharmacy / immunocompromised individuals / remission from cancer, etc.). However, the inclusion of these "vulnerable" populations can only occur under certain conditions described in Articles L.1121-5 and following of the Public Health Code.

2) Accelerating the market entry of new treatments

By reducing the number of patients to recruit in an initial study, if the robustness of these new methodologies is established, generating artificial patients would reduce the cost and duration of studies. The funds saved could then be reinvested in the research of new innovative treatments, and the time saved would make these treatments accessible to patients more quickly.

3) Facilitating inclusions in a trial (especially when difficulties in inclusion in the control arm are known).

To demonstrate the effectiveness or superiority of a treatment, it is necessary to show a difference in effect compared to a reference strategy (depending on the case, absence of treatment, reference treatment). Therefore, establishing a control arm in trials is essential. This always raises delicate ethical questions regarding the allowed concomitant treatments to avoid biasing the analyses.

In some studies, "escape" arms are planned to allow patients who are suffering too much to be switched to treatment. This poses a risk for the study because if the number of patients reaching the end of this placebo period is insufficient, the primary endpoint cannot be evaluated with the necessary statistical power.

Subject to validation of these methods, the prospect of replacing real patients on placebo with artificial patients could resolve these dilemmas. It could also allow for comparing the evolution of the treated arm with an untreated arm over a longer period, which could be very informative.

Moreover, if patients, like the physicians accepting these studies, no longer had this apprehension of potentially receiving a non-active treatment, recruitment would be facilitated de facto.

However, it should be noted that patients knowing they are definitely receiving treatment may potentially have a more significant placebo effect. For this reason, it will probably be preferable not to completely substitute the placebo arm with an arm entirely composed of artificial patients.

Virtual data is still in its infancy but could alter the organization and development of innovation in the healthcare industries by shortening the development time of new molecules, especially in rare diseases.

Dr. Sylvie Troy
Associate Medical Director and Real-World Data/Real-World Evidence Lead, France,
Pfizer

ii. Points to consider for the use of artificial data in clinical research

1) Defining what an artificial patient should consist of

Throughout this white paper, we discuss artificial patients, but what constitutes an artificial patient remains to be defined, and will likely depend on the situation.
It will be necessary, in consultation with relevant experts, to determine the parameters to generate (potentially conditioned by the maximum number of dimensions that can be handled by the model used), the format of these parameters, the number of time points required...
The biggest challenge will undoubtedly lie in generating potential adverse events.

2) Data relevance

The first requirement, whose importance is evident, is to ensure that the datasets used to train the model that will enable the generation of artificial patients accurately represent the population to be augmented.
These datasets must represent:

a) The studied pathology

As mentioned, precision medicine increasingly requires focusing on subcategories of patients within a pathology. Therefore, it will be necessary to ensure that real patients correspond to these specific populations.

b) The specific characteristics of the population to be augmented

Similarly, if the aim is to enrich the diversity of a cohort, representative datasets of the physiological specifics of these populations will need to be identified (ethnicities, immunocompromised patients, pregnant women, children/adolescents, elderly individuals, etc.). This will pose a significant challenge as these data may be scarce and, additionally, they will always need to represent the studied pathology.

The creation of an artificial patient cohort from real patient data, gathered within the field of anesthesia, confirms the potential of artificial intelligence in healthcare processes and clinical research. As healthcare professionals and scientists, we are working towards validating these technologies for routine use, to make them available to our colleagues and the most vulnerable patients. The collaboration between mathematical, statistical, and medical domains to develop reliable, secure, and ethical artificial intelligence tools foreshadows the advent of an increasingly personalized and humane medicine.

Dr. Fabrice Ferré
PhD, Anesthesia and Critical Care Specialist,
Toulouse University Hospital

c) The period during which the trial is conducted

In the long term, there can be a natural evolution of certain diseases (examples include the decrease in the frequency of infections vs the increase in allergic conditions, unexplained decrease in the frequency of certain forms of spondylarthritis, etc.), and in the short term, we can face particular situations, such as the COVID pandemic to give a pertinent example (infectious context, but also impact on mental health).

3) The compliance of the source data with the latest state of scientific knowledge

As mentioned in the previous point, the timing and environment during which the trial takes place can impact the effectiveness of the evaluated treatment (either overestimated or underestimated). Similarly, the concomitant treatments tolerated in some studies vary over time. The databases used to generate artificial patients should closely resemble the context in which the trial is conducted.

During the evaluation of brodalumab, an abnormally high number of suicides were observed during phase 3. Biologically, the reason why inhibition of the IL-17 receptor might increase suicidal thoughts and actions was very puzzling. One hypothesis suggested was that, concurrently with this study, the United States was going through the subprime mortgage crisis. However, it was not possible to formally demonstrate that this explained this severe adverse effect. This example is extreme and could not have been anticipated, but it illustrates the importance of having a comparator arm closely aligned with the context in which the trial takes place.

4) Eligibility criteria

As mentioned in the previous points, the real data used to generate synthetic data must be representative of the studied population. One of the challenges in this regard will be to ensure that these patients meet the eligibility criteria for the studies: severity of the disease, previous treatments, specific biomarkers, etc. Some of these criteria may be difficult to identify in the databases used to generate artificial data.

5) The quality of the reference database

It will be necessary to calibrate the quantity of data required and sufficient to capture weak signals within the population to be augmented. It may be challenging to find a sufficient database that allows for representativeness of interindividual variability. These preliminary analyses should guide the optimal choice between a synthetic (and therefore retrospective) arm and an augmented arm in case of sufficient representativeness of the target population.

These points of consideration will be further elaborated in the chapter on evidence and methodology for the validation of healthcare products.

The reasoned creation of artificial data and patients can in the future contribute to reducing the number of patients required in the comparator arm of clinical trials and to accelerating these trials, particularly in rare diseases. It should contribute to encouraging rather than discouraging the sharing and pooling of real data. Following the ongoing scientific maturation on the subject, regulatory agencies will need to organize useful workshops and symposiums, and then establish recommendations and scientific opinions for therapeutic developers.

Dr. Hervé Nabarette
Deputy Director of Public Affairs,
AFM-Téléthon

II. Use of artificial data for training artificial intelligence algorithms (calibration)

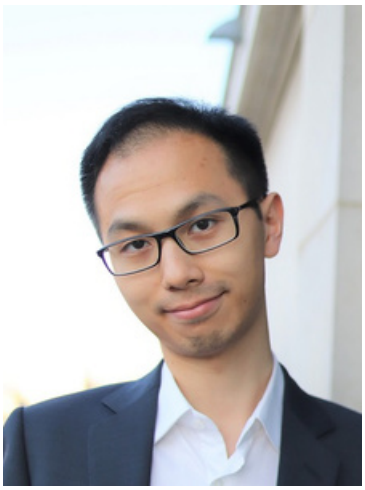
Calibration, also known as estimation, of mathematical models using Machine Learning or Deep Learning requires a sufficient quantity of reliable and high-quality data, which implies that they must be sufficiently representative of the study population. This often entails multicentric data that are difficult to mobilize, leading to time-consuming and resource-intensive processes, especially when it comes to health data that cannot easily be used outside their collection facility.

To obtain multicentric data, the emergence of Federated Learning holds much promise. It allows models to be trained without centralizing data on a single server and thus keeping them locally stored. However, it is still a subject of research and its deployment is not yet a short-term certainty. Indeed, Federated Learning can pose significant security risks to models learned, such as the risk of data poisoning[8]. The use of artificial data could create local, artificial, non-identifiable databases that could be aggregated to form a multicentric artificial database, which would represent a significant time and cost-saving, enabling a quicker availability of new decision support tools for healthcare professionals and patients.

Another scenario where such data could save time and expertise is if, for example, to train a radiology image recognition algorithm, 10,000 images need to be collected and annotated (by several experienced radiologists). The use of artificial data would allow for the collection of only a portion of this database, which would then be augmented. For instance, only 2,000 images would need to be collected and annotated, which could then be augmented to reach the required 10,000. The collection of such data can be lengthy and costly, especially due to challenging contractual arrangements and poorly regulated intellectual property sharing clauses for such uses. Similarly, expert annotation incurs costs for the development of these new devices. The reduction in data collection and annotation (2,000 instead of 10,000) demonstrates a new interest in data augmentation methods.

The diagnostic industry is also affected by artificial data. For instance, GANs have been successfully used by American researchers to improve the prediction of type 2 diabetes diagnosis.[9]

8 https://cybersecurity.springeropen.com/articles/10.1186/s42400-021-00105-6
9 https://pubmed.ncbi.nlm.nih.gov/37873778/



Virtual patient cohorts are an emblematic example of how AI can innovate in generating evidence and complement classical research, even in situations where patients and data on their illness are rare.

Dr. Vinh-Phuc Luu
Head of Epidemiology and Innovation Division,
Artificial Intelligence & Cancers Association

04

DATA PROTECTION AND GENERATION OF ARTIFICIAL PATIENT COHORTS

The algorithms used to generate artificial patient cohorts rely on data from "real patients" collected as part of patient care or previous research. When this data contains personal information, the General Data Protection Regulation (GDPR)[10] and Law No. 78-17 of January 6, 1978, as amended (known as the "Data Protection Act"), will apply, provided that these processing activities fall within the material and territorial scope of the regulation and the law.

Furthermore, depending on the context of use of these systems, other sector-specific provisions from national law (such as the Public Health Code and the Penal Code) and European law (notably regulations on clinical trials [11] and medical devices[12]) may also apply.

Finally, the regulatory landscape will soon be supplemented by the provisions of the European regulation on artificial intelligence, known as the "AI Act," once it comes into effect, particularly those applicable to high-risk artificial intelligence systems. This will include systems integrated into certain medical devices. Before being placed on the market in the European Union or put into service, the compliance of these systems must be assessed to demonstrate that they meet certain mandatory requirements (such as data quality, documentation and traceability, transparency, human oversight, accuracy, cybersecurity, and robustness). This assessment must be repeated in the event of a substantial modification to the system or its purpose. Additionally, providers of high-risk artificial intelligence systems must implement quality and risk management systems to ensure compliance with the new requirements and to minimize risks for users and individuals concerned, even after the product has been placed on the market.

However, this chapter will only address issues related to the regulations applicable to the processing of personal data.

10 Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation).

11 Regulation (EU) 536/2014 of the European Parliament and of the Council of 16 April 2014 on clinical trials on medicinal products for human use, and repealing Directive 2001/20/EC.

12 This includes notably Regulation (EU) 2017/745 of the European Parliament and of the Council of 5 April 2017 on medical devices and Regulation (EU) 2017/746 of the European Parliament and of the Council of 5 April 2017 on in vitro diagnostic medical devices. It should be noted that certain specific national provisions are also provided for in the Public Health Code (articles L.1125-1 et seq.).

Virtual cohorts are one of the drivers to accelerate the development of technological innovations. This white paper opens avenues to create the conditions for their validation.

Corinne Collignon
Head of Digital Health Mission Service,
Haute Autorité de Santé

I. Qualification of training and generated data

To identify the applicable legal framework, it is necessary to qualify the training data, the artificial intelligence model, and the generated data.

i. The concept of processing personal data

On one hand, "processing" refers to any operation performed on personal data, such as recording, structuring, storing, adapting, altering, retrieving, consulting, using, disclosing, or making available. Thus, training an artificial intelligence model constitutes processing under Article 4 of the GDPR.

On the other hand, a "personal data" is, according to the same provision, "any information relating to an identified or identifiable natural person." This person "can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier, or one or more specific elements specific to his or her physical, physiological, genetic, mental, economic, cultural, or social identity." This data can be directly identifying (linked to the person's first and last name) or indirectly identifying (linked to an order number or an alphanumeric code, in which case it is pseudonymized [15] data).

15 As defined in Article 4 of the GDPR, pseudonymization is "the processing of personal data in such a way that the data can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and subject to technical and organizational measures to ensure that the personal data are not attributed to an identified or identifiable natural person."

The GDPR provides special protection for certain categories of data mentioned in Article 9, also known as "sensitive data." This includes health data[16], defined in Article 4 of the GDPR as "*personal data relating to the physical or mental health of a natural person, including the provision of health care services, which reveal information about the health status of that person.*" The processing of this data is generally prohibited, unless one of the exceptions provided by the GDPR or the "information and freedom law" (French Data Protection Act) can be invoked, such as obtaining explicit consent from the data subject or the necessity of processing the data for health care or scientific research purposes. This legal framework is justified by the fact that processing these particular categories of data could pose significant risks to the freedoms and rights of the data subjects.

ii. The concept of anonymization

Anonymization is a process of handling personal data to produce information that cannot, by '*reasonable means*,' be linked to identified or identifiable individuals. Data protection legislation remains applicable concerning the anonymization process, but no longer applies to data resulting from this process once their anonymous nature has been demonstrated.

As the anonymization process reduces the information present in the original dataset, it should be designed to preserve valuable information as much as possible while accepting to mitigate other aspects. To be useful, this process can vary greatly depending on the nature of the original dataset and business needs.

In practice, two families of anonymization techniques can be combined for this purpose:

- Random perturbation (or randomization) involves modifying attributes in a dataset in such a way that they are sufficiently uncertain to alter accuracy. This uncertainty weakens the link between a piece of data and the individual to whom it relates, while preserving the overall statistical properties of the dataset. This family of techniques helps protect the dataset from the risk of inference. For example: adding random noise to collected data (± two centimeters to patients' height); permuting attribute values among patients (exchanging 'age' data between two randomly chosen patients);

16 To learn more, see the practical guide published on the CNIL website.

- Generalization involves modifying the scale or magnitude of attributes in a dataset to make them common to a larger group of people. It helps make the information less specific to individuals while preserving its utility. This family of techniques prevents the individualization of a dataset and limits possible correlations of the dataset with others. Aggregating data (counting, averaging, etc.) and techniques that give the dataset properties like k-anonymity, l-diversity, and t-closeness are examples of generalization methods.

However, solely employing these techniques does not conclusively ensure that a dataset is anonymous. For instance, artificial intelligence systems typically consist of statistical models trained on real data, which may include pseudonymous health data. These models contain information from the data they are fed and could potentially identify individuals in the training databases. Thus, they are susceptible to attacks such as reconstruction attacks or attribute inference attacks[17], which can lead to identifying individuals. Therefore, it cannot be assumed by default that an artificial intelligence model is anonymous.

In 2014[18], European data protection authorities outlined three criteria that, when met together, indicate that the data in question is anonymous:

1. **Uniqueness**: It should not be possible to isolate an individual in the dataset.
2. **Correlation**: It should not be possible to link the data with another separate dataset concerning the same individual.
3. **Inference**: It should not be possible to deduce, with near certainty, new information about an individual.

The anonymization process should aim to produce a dataset that meets all three of these criteria. This compliance should be documented and demonstrable.

However, if any of the criteria are not satisfied, a more thorough examination should be conducted to determine if the GDPR and the "information and freedoms" law apply.

[17] Check out this recent publication from the National Institute of Standards and Technology (NIST): https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-2e2023.pdf

[18] For more information, refer to the guidelines. These are currently being updated, but the three criteria outlined in this document remain valid.

Thus, if these three criteria are not fully met, the data controller wishing to anonymize a dataset must demonstrate, through a thorough assessment of the risks of identification, that the probability of identifying a person using means that are reasonably likely to be used, by themselves or by any other person, is negligible.[19]

In conclusion, to identify the applicable legal framework in the case of cohorts composed of virtual patients generated through artificial intelligence techniques, three questions must be distinguished:

- That of the training data: these will often consist of personal data. However, in some cases, considering the pursued purpose, it is conceivable that the data required for training the AI model contain no identifying data and are of a coarse enough granularity to demonstrate that they are anonymous.
- That of the data of the virtual patients generated by AI models: an analysis, based on the three criteria mentioned above or on an ad hoc basis, should be conducted to demonstrate that the data of the virtual patients thus constituted do not replicate data from real patients.
- That of the AI model itself, which, if trained on personal data, may not be anonymous and should undergo the same analysis.

In the event that personal data is involved, the following legal framework will apply,[20] in addition to any sector-specific regulations mentioned in the introduction to this chapter.

II. Ensuring compliance with key "information technology and freedoms" principles at all stages of the algorithm's lifecycle.

The development and deployment phases of an artificial intelligence system constitute separate processing of personal data. For instance, training a model aimed at producing virtual patient data is part of the development phase, while its use for generating virtual patient cohorts in a hospital context occurs during a deployment phase. As described earlier, none of these phases may constitute processing of personal data, just one of them, or both. Often, and upon initial analysis, concerning the formation of virtual patient cohorts, data processing will be implemented for the model generation training, and the processing related to the model's usage and virtual patient data creation will be considered anonymous, provided the aforementioned conditions are met.

19 To learn more, see the practical guide published on the CNIL website.

20 To learn more, see the practical guide published on the CNIL website.

The compliance of each of these processes must be documented by their data controller, i.e., the entity that determines, alone or jointly, the purposes - the objectives - and the means - the way to achieve the objectives - of the processing: the algorithm provider[21] *during its development and its user during its deployment, except in the case of non-commercial*[22] *personal use. Adhering to these principles, which may sometimes raise new questions, will help generate reliable and quality data. Their consideration should be anticipated as early as possible, following a privacy-by-design*[23] *approach. To do so, algorithm providers can be supported, notably by the National Commission on Informatics and Liberty (CNIL), which has already published several practical guides*[24] *and an action plan on artificial intelligence.*[25]

i. The principle of lawfulness of processing.

Like any processing of personal data, the creation and use of a database for the training of artificial intelligence systems must be based on one of the six legal bases provided for in Article 6 of the GDPR to be legally implemented. Among the legal grounds that may be invoked are consent, legal obligation, performance of a contract, performance of a task carried out in the public interest, vital interests, and legitimate interests.[26]

The processing of personal data carried out in the development and deployment of the algorithm must therefore be based in all circumstances on one of the legal bases mentioned above. It should be noted that the choice of the legal basis selected, which is subject to specific conditions, has consequences for the exercise of the rights of the data subjects.

Furthermore, if the development of the algorithm involves reusing data already collected (in the context of medical care or previous research), it will be necessary to ensure

21 The natural or legal person, public authority, agency, or other body that develops or has developed an AI system for placing it on the market or putting it into service under its own name or brand, whether for remuneration or free of charge.

22 To learn more, see the practical guide published on the CNIL website.

23 To learn more, see the practical guide published on the CNIL website.

24 To assist organizations in developing privacy-respecting solutions, the CNIL has published numerous fact sheets, including a self-assessment guide for artificial intelligence systems.

25 For more information, see the action plan.

26 For more information, see the practical guide published by the CNIL regarding the legal bases and specificities related to the use of artificial intelligence.

that the initial database was regularly established and implemented (for example, by ensuring that the initial processing was, if necessary, subject to adequate formalities and that the retention period of the initial database has not expired).

ii. The pursuit of a specific, explicit, and legitimate purpose

Any processing of personal data must be part of a sufficiently determined and specific objective. This purpose must be legitimate in light of the controller's missions. In practice, within the scope of health research, this purpose corresponds to the specific scientific question that the project aims to answer. It is materialized within the main objective as well as, where applicable, the secondary objectives mentioned in the study protocol.

Similarly, the establishment of a database containing personal data for the development of an artificial intelligence system constitutes processing of personal data that must pursue a determined, explicit, and legitimate purpose. In order to assist controllers in defining this objective according to their use cases, particularly considering whether the operational use of the system is identified during the development phase or not, the CNIL has indeed published a practical guide.[27]

Finally, it should be noted that in the case of reuse of already collected data, it is also the responsibility of the data controller to verify that the purpose of the new processing is compatible with that of the initial processing, with the understanding that the reuse of data for scientific research purposes is presumed to be compatible when the processing meets certain guarantees (considering 50 and article 5.1.b) of the GDPR as well as article 4.2° of the "informatique et libertés" law).[28]

[27] To learn more, see the practical guide published on the CNIL website.
[28] To learn more, see the practical guide published on the CNIL website.

A multidisciplinary perspective is essential for understanding the scientific, legal, and ethical challenges of future clinical trials.

Prof. Marie-France Mamzer
Professor of Medicine and Hospital Practitioner,
Paris-Cité University

iii. A processing of accurate and necessary data

The processed data must be relevant, adequate, and limited to what is strictly necessary for the intended purpose. This rule, known as the "principle of minimization," is provided for in Article 5 of the GDPR. In the context of a research project, the processing of each category of data must be scientifically justified (for example, through existing scientific publications) or related to one of the study's objectives or evaluation criteria. Thus, processing patients' social security numbers in a research project aimed at developing an algorithm to generate virtual patient cohorts would not appear to comply with the principle of minimization.

While the use of large quantities of data is central to the development and use of artificial intelligence systems, the principle of minimization is not an obstacle to carrying out these processes, just as it has not hindered the development of health data warehouses authorized by the CNIL since 2017.

Moreover, the data necessary for the development of the algorithm must be accurate and, if necessary, kept up to date by the data controller.

iv. Limited data retention

Personal data must be processed for no longer than is necessary for the purposes for which they were collected, in accordance with the provisions of Article 5 of the GDPR. They are kept in an "active database", meaning they are only retained for as long as necessary to achieve the intended purpose. Subsequently, they must be destroyed, anonymized, or archived in compliance with applicable legal obligations.

The need to define a retention period for data used in a processing operation does not prevent the implementation of artificial intelligence processes (especially in the case of continuous learning), as long as the chosen duration, which may be relatively long, is justified.

It is thus necessary to establish a data retention period for the development of the artificial intelligence system:

- For the development phase: data retention must be planned in advance and monitored over time.

- For maintenance, monitoring, or product improvement purposes: when the data are no longer needed for the daily tasks of those responsible for developing the artificial intelligence system, they should generally be deleted. However, they may be retained for product maintenance and monitoring or improvement if safeguards are implemented (e.g., compartmentalized support, restricted access to authorized personnel only, etc.).

Retention of training data can enable audits and facilitate the measurement of certain biases. In such cases, extended data retention may be justified. This retention should be limited to necessary data and accompanied by enhanced security measures.

v. The implementation of appropriate technical and organizational measures and the conduct of an impact analysis.

The data controller must implement appropriate technical and organizational measures to ensure a level of security adapted to the identified risks (such as unauthorized access, unwanted modification, or data loss).[29] Generally, security measures suitable for processing patients' health data involve using a secure platform, such as those offered by healthcare data hosting providers[30] or established within healthcare data warehouses.[31]

Furthermore, the development of artificial intelligence systems may require, in some cases, the conduct of a Data Protection Impact Assessment (DPIA). This is mandatory if the intended processing is likely to result in a high risk to the rights and freedoms of individuals (Article 35 of the GDPR), which may be the case regarding the establishment of a health database for AI system training.

The European Data Protection Board (EDPB) has identified nine criteria to assist data controllers in determining whether a DPIA is required, such as the collection of sensitive data, the collection of data from vulnerable individuals, large-scale data collection, or innovative usage. Any processing meeting at least two criteria from this list will be presumed subject to the obligation to carry out a DPIA.

29 To learn more, see the practical guide published on the CNIL website.

30 To learn more, see the page dedicated to healthcare data hosts published by the French National Agency for Digital Health (Agence du Numérique en Santé).

31 To learn more, see the reference document regarding healthcare data warehouses by the French Data Protection Authority (CNIL).

In practice, one or more healthcare data processing operations (considered sensitive data) involving patients (deemed vulnerable individuals according to the EDPB guidelines) for the development of an artificial intelligence system to generate virtual patient cohorts (potentially classified as an "innovative use" given current technological knowledge) meet at least two of the nine EDPB criteria. Therefore, a Data Protection Impact Assessment (DPIA), which may cover a set of similar high-risk processing operations, must be conducted by the data controller. This assessment will map and evaluate the risks of the data processing on personal data protection and establish an action plan to mitigate them to an acceptable level.[32]

vi. Transparency, loyalty, and respect for individuals' rights

The individuals whose data is reused (also referred to as "data subjects") must be informed and able to exercise their rights.

1) The methods of information

Several provisions are applicable regarding information. Some are specific to the context in which the data will be processed.

The GDPR requires that individuals be informed in a concise, transparent, understandable, and easily accessible manner in clear and simple terms. Articles 13 (direct collection) and 14 of the GDPR (indirect collection or when research involves reusing data from an existing database) provide the methods of information and the elements to be included in the information notice (identity of the data controller, purposes of processing, recipients of the data, data retention period, etc.).[33] In the case of automated decision-making, individuals must be informed, at the time of data collection and at any time upon request, of the existence of such a decision, the underlying logic, and the significance and envisaged consequences of this decision. This information must be tailored to each category of individuals affected by the processing, particularly in the presence of minors or vulnerable individuals, to ensure maximum transparency.

[32] To learn more, see the practical guide published by the CNIL concerning DPIA and the specificities related to the use of artificial intelligence.

[33] Furthermore, in the presence of automated decision-making, individuals must be informed, at the time of data collection and upon request at any time, about the existence of such a decision, the underlying logic, as well as the significance and envisaged consequences of that decision.

Regarding research, studies, or evaluations in the healthcare field, Article 69 of the "Informatique et Libertés" law provides for individual information of the data subjects in accordance with the GDPR, especially in the case of direct data collection from the data subjects.

Data collected from healthcare or previous studies may be reused without the need for new individual information, especially when the information provided during the initial data collection allows for data reuse and refers to a specific information mechanism that data subjects can refer to before the implementation of each new data processing operation (for example: a website or dedicated page called a "transparency portal" presenting each research project conducted using the data collected during the initial information). This method is strongly recommended by the CNIL as it is a simple way to inform data subjects about all subsequent research projects. However, it requires anticipating the possibility of reuse from the outset of the initial processing by creating a dedicated web page for information.

In cases where new individual information cannot be provided and data reuse has not been anticipated through a transparency portal, three exceptions to providing individual information are provided for by Article 14.5.b) of the GDPR for processing carried out for scientific research purposes. This includes situations where providing individual information:

- Proves impossible;
- Or would require disproportionate efforts from the data controller in relation to the age of the data or the number of data subjects;
- Or would compromise or make impossible the achievement of the processing objectives.

Appropriate measures must then be implemented by the data controller to protect the rights and freedoms of the data subjects. In such cases, the GDPR, guided by the guidelines adopted by the EDPB, provides that appropriate measures must be taken to protect the rights and freedoms of data subjects, notably through the systematic provision of collective information via appropriate communication channels considering the study's context (at minimum, on the data controller's website).[34]

34 To learn more, see the practical guide published on the CNIL website.

Furthermore, in the event that cohorts of virtual participants are generated as part of research involving human subjects, clinical trials, clinical investigations, or performance studies, certain specific provisions from the French Public Health Code and European regulations on clinical trials, medical devices [34] (including in vitro diagnostic devices [35]) would apply regarding the content of information documents and their recipients.

For example, participants in research involving human subjects have the right, in accordance with Article L. 1122-1 of the French Public Health Code, to be informed in an understandable manner about the research methodology, which includes the use of an algorithm to generate virtual cohorts. In assessing the validity conditions of the research, it would then be up to the ethics committee to consider, among other things, the proposed methodology as well as the adequacy, completeness, and intelligibility of the information documents in accordance with the provisions of Article L. 1123-7 of the French Public Health Code. As for patients whose data from treatment or previous research would be reused in the context of a comparator group in research involving human subjects, they should, under current law, be informed in the same manner as other participants, and their consent, when required, should be obtained, in the absence of specific provisions provided by the French Public Health Code and applicable European regulations on clinical trials and medical devices. It is worth noting that work, in which the CNIL is involved, is currently underway on this subject to adapt this legal framework to new research methodologies.

Finally, the provisions of Article L. 4001-3 within the French Public Health Code regulate the use, for a preventive, diagnostic, or therapeutic act, of medical devices involving algorithmic data processing whose learning has been carried out from massive data. In this case, the healthcare professional must inform the person concerned about the use of algorithmic processing and, where applicable, the interpretation that results from it. Additionally, the healthcare professional must also be informed about the data processing (including the data used and the results). Furthermore, the designers of the algorithmic processing must ensure the explainability of its operation for users. The nature of the medical devices concerned and their methods of use must be specified by an order of the Minister responsible for health established after consultation with the High Health Authority and the CNIL. It has not yet been published, so it is not possible to determine whether the device used to generate virtual patient cohorts could be subject to these provisions.

34 Regulation (EU) 2017/745 of the European Parliament and of the Council of 5 April 2017 on medical devices

35 Regulation (EU) 2017/746 of the European Parliament and of the Council of 5 April 2017 on in vitro diagnostic medical devices and repealing Directive 98/79/EC and Commission Decision 2010/227/EU

2) The modalities for exercising the rights of data subjects

The data controller must ensure that individuals whose data is processed have the opportunity to exercise their rights of access (Article 15 of the GDPR), rectification (Article 16 of the GDPR), objection (Article 21 of the GDPR), erasure (Article 17 of the GDPR), restriction (Article 18 of the GDPR), and portability (Article 20 of the GDPR) under the conditions provided by the GDPR and the "Data Protection Act." [36]

Furthermore, Article 22 of the GDPR provides that, as a principle, individuals have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning them or similarly significantly affects them. Some exceptions are provided, particularly when the decision is based on the explicit consent of the individuals. In this case, the data controller must implement appropriate measures to safeguard the rights and freedoms and legitimate interests of the data subjects. In addition to the aforementioned transparency obligations, any individual subject to such a decision may request human intervention, particularly to have their situation reconsidered, to express their own point of view, to obtain an explanation of the decision made, or to challenge the decision. Finally, individual automated decisions cannot be based on health data, except in specific cases (collection of the data subject's consent allowing derogation from the principle of prohibition of processing such data or processing necessary for reasons of public interest in the field of public health). In this case, appropriate measures to safeguard the rights and freedoms and legitimate interests of the data subjects must be implemented.

These rights constitute essential protection for data subjects, allowing them to avoid the consequences of automated systems without having the opportunity to understand and, if necessary, object to data processing concerning them. In practice, these rights apply throughout the lifecycle of the artificial intelligence system.

36 To learn more, see the practical guide published on the CNIL website.

III. The possible formalities applicable in accordance with the "Data Protection Act"

The possible formalities will differ depending on the life cycle and context of using the algorithm to generate virtual cohorts. To date, these questions are still quite recent and have not yet been resolved by the CNIL board.

i. In the context of algorithm development

The processing of health data for the development of an algorithm intended to generate cohorts of virtual patients in a medical context could constitute research, a study, or an evaluation in the field of health falling under the "Data Protection Act." In accordance with the provisions of Article 66 of the "Data Protection Act," scientific research in health requiring the processing of health data must, unless exceptions provided by law, be subject to a declaration of compliance with a standard data processing methodology described in a reference framework published by the CNIL called "reference methodology."[37] Exceptionally, research that does not comply with these frameworks due to particular sensitivity (such as data processing of offenses, for example) must be authorized by the CNIL, after consultation with the competent committee.

In the absence of compliance with a reference methodology, the CNIL may also issue a single decision after consultation with the competent committee, in accordance with the provisions of Article 66 IV of the "Data Protection Act." Through this mechanism, the CNIL can authorize, with a single decision, a significant number of treatments for a specified period, serving the same purpose, involving the same categories of data, and having identical categories of recipients. In order for the competent committee to assess the scientific relevance of the project, these treatments must be implemented based on a standardized methodology established by the data controller. The use of this mechanism may be justified by the need to implement a significant volume of treatments. Therefore, it could be explored in the event that the development of the algorithm involves the implementation of numerous treatments of personal data.

37 To learn more, see the section dedicated to reference methodologies published on the CNIL website.

In any case, and regardless of the formality to be completed with the CNIL, the scientific protocol developed by the data controller should notably:

- Describe the purpose of the processing of personal data, the nature of the data processed, the proportionality of the chosen techniques, the impact and benefits of using this artificial intelligence system;
- Document the methodology (assumptions made about the training data, conducting a bias study) to ensure a reliable, long-term robust processing (to minimize the risk of drift) and demonstrate that the intended processing serves a public interest purpose.

At the end of this development phase, a scientific evaluation of the system should be conducted before it can be deployed.

ii. As part of the deployment of the artificial intelligence system

1) The generation of new cohorts of virtual patients within research projects

In this scenario, the data processing necessary for generating new cohorts within a clinical trial could fall within the scope of the prior formal procedure related to that trial (declaration of compliance with a reference methodology or, in the absence of compliance, submission of an authorization request to the CNIL).

2) The generation of new cohorts of virtual patients within the framework of patient care

In case of using the artificial intelligence system by a healthcare professional or the institution responsible for the patient's care, for diagnostic, care delivery, or treatment purposes, the processing could fall under one of the exceptions to prior formalities provided for in Article 65 of the "Data Protection Act", more specifically, the development of medical diagnostics, or the administration of care or treatment.

I*n any case, regardless of the uses during the deployment phase, it will be necessary to verify if the artificial intelligence system indeed meets the needs for which it was designed. Moreover, as mentioned above, in the event that the initial dataset contains personal data, it is necessary to analyze the status of the artificial intelligence model and the data generated by the algorithm, particularly in light of the European criteria regarding anonymization.*

The development of these artificial intelligence systems comes with emerging data protection challenges. The CNIL aims to support their growth by assisting their providers to deploy them in compliance with the privacy of the individuals involved.

Manon de Fallois
Assistant to the Chief of Health Services,
CNIL

EVIDENCE AND METHODOLOGY FOR VALIDATING HEALTHCARE PRODUCTS USING ARTIFICIAL DATA

Given the exponential growth of data production, the quality of sensors, and the power of analysis, preparing for the use of data to simulate all the usual parameters of clinical research or real-life research is not an option. We must move forward by honestly assessing the potentials and limitations of these technologies brought together to participate in a global race. I am convinced that it is through white papers like this one, coupled with real projects, that we will achieve this.

Marco Fiorini
CEO, Artificial Intelligence & Cancers Association

The abundant literature shows that mathematical models can have many applications in healthcare. This can involve highly technical aspects, such as testing the robustness and security of IT tools, or more operational ones, like assisting surgeons in intervention planning, developing diagnostic or prognostic models to build decision support systems, creating digital twins of healthcare facilities to anticipate patient flows within them. As previously mentioned, artificial data can also have significant applications in terms of healthcare product development, in very early stages of discovering new medications (for example, [75, 76]), as well as for evaluating their efficacy and safety, or personalizing treatment strategies. These latter uses are currently in their infancy, and it is important to consider the evidence and validation methods to be implemented for the use of such virtual data to be acceptable to the various stakeholders. The trust provided by the validation of these methods is necessary if we ever wish to use artificial or partially artificial cohort data for access to new treatments.

It is also important to make the distinction, as proposed by the FDA, between the analytical validation of algorithms and their clinical validation under real-life usage conditions. There is often confusion between analytical validation, which includes validation on data (this may include multiple data sources from different origins), and clinical validation, which may include prospective clinical trials or using RWD (Real World Data) once the algorithm (via an app or other means) is used and generates its own data. In both cases, the study methods and designs differ, as do the validation criteria.

Currently, there are no recommendations from agencies or regulatory authorities defining the acceptability criteria for artificial patient cohorts for evaluating healthcare products or medical devices. References to simulation studies based on mechanistic models can be found in ICH E11 recommendations, in the ICH guideline E11A section on pediatric extrapolation [77], and EMA and EFPIA co-organized a seminar on modeling and simulation[38] over 10 years ago. This topic is still under consideration at the EMA.[39]The methodological guide from HAS[40] for the clinical development of medical devices, updated in 2021, also mentioned in silico studies. However, it concluded that, to date, in silico studies should be considered complementary tools to existing methodologies when there is a pathophysiological model and should be reserved for feasibility studies. Finally, a framework for evaluating the credibility of mechanistic models has recently been developed by a working group coordinated by the EMA, the Modelling and Simulation Working Party, for the use of in silico studies based on mechanistic models [78].

38 European Medicines Agency-European Federation of Pharmaceutical Industries and Associations modelling and simulation workshop, 2011 ; https://www.ema.europa.eu/en/events/european-medicines-agency-european-federation-pharmaceutical-industries-associations-modelling

39 https://www.ema.europa.eu/en/human-regulatory-overview/research-and-development/innovation-medicines

40 https://www.has-sante.fr/upload/docs/application/pdf/2013-11/guide_methodologique_pour_le_developpement_clinique_des_dispositifs_medicaux.pdf

In summary, this type of data is beginning to emerge, and to date, regulatory agencies and authorities have been relatively underutilized regarding such data. Consequently, none of the currently available documents or acts provide specific expectations for the acceptability of methods for generating artificial data and their use in evaluating the efficacy and safety of healthcare products or medical devices, as highlighted in a position paper by the INSILICO WORLD[41] consortium.

On the contrary, there is a wealth of research on the use of synthetic - or external - control arms constructed from observational data, including from agencies such as the French National Authority for Health (HAS) and the Transparency Commission, and even recommendations from the FDA.[42]

EMA and the network of national medicine agency directors unveiled their work plan on artificial intelligence last December, aiming to develop a European strategy. This work plan is intended to better regulate the use of artificial intelligence by national regulators.

I. In-silico studies based on mechanistic models

As described in Chapter 2, the generation of artificial cohort data can be based on mechanistic approaches, relying on the modeling of the considered physical, chemical, and biological systems, or rather on learning methods aimed at reproducing characteristics similar to those of an observed population. For this type of approach, a framework for assessing the credibility of mechanistic models has also been proposed [78].

41 Toward good simulation practice: best practices for the use of computational modelling & simulation in the regulatory process of biomedical products. Version: R6, 06-05-2023. https://insilico.world/sito/wp-content/uploads/2023/06/Position-Paper-GSP-R6.pdf

42 https://www.fda.gov/regulatory-information/search-fda-guidance-documents/considerations-design-and-conduct-externally-controlled-trials-drug-and-biological-products

The principles outlined in these recommendations can also be applied to studies using artificial data. This group of authors proposed a seven-point approach:

1. Description of the research question and context of use.
2. Definition of model acceptability criteria for this research question and context of use.
3. Description of the regulatory impact (for submissions to agencies such as the EMA).
4. Analysis of decision consequences.
5. Description of model and algorithm credibility (verification and validation).
6. Applicability and uncertainty.
7. Decision-making informed by the model.

Several important points for defining the methods and evidence to be implemented can thus be identified: acceptability criteria for a particular question, which includes, among other things, the theoretical relevance of the models used, as well as the quality of their computer implementation (including the quality assurance processes implemented), the validation methods used, aspects concerning the data used to develop and validate the model, and a study of the impact of various sources of uncertainty. The model's credibility elements primarily concern its ability to (re)produce data that closely match what would have been observed if real patients had been included. The detailed aspects in the list below, based on the proposals of Musuamba et al. [78], should thus be considered to validate this type of approach, adapting them if necessary to the context of artificial data and the methods used to generate them:

Detail the data generation models.

- Approaches used to generate artificial data, whether mechanistic models or learning models, as described in Chapter 2.
- Evaluation of the model's suitability in relation to objectives (research question or use case), especially if the artificial patient group aims to improve algorithm development, augment a control group in a trial, or even potentially replace it; these points will be revisited later on.

Detailing the implementation, in connection with ISO 13485 regulations:

- Methods used to verify the code and results of verification actions.
- Aspects related to algorithm transparency (and development and validation methods): description of the methods used, availability of the algorithm, its implementation, etc.
- Quality assurance process and compliance with ISO 13485 standards in particular.
- Elements attesting to the robustness of the algorithm (to data disturbances, adversarial attacks, etc.), both theoretically and practically.

Detailing the validation methods:

- The algorithm's ability to reproduce observations already obtained (probability distributions of identical populations for initial and augmented data), and on which it has not been trained (same probability distribution for a test dataset).
- External validation study of prediction models, in the population of interest and on the object of interest, for example, by taking 50% of the data from a control arm, artificially supplementing it, and then comparing the treated arm with the original control arm and the treated arm with the augmented arm.
- Classical metrics (R2, discrimination, calibration, etc.) or use of more specific methods (for example, [84]).

Detailing the data used:

- For the development of the algorithm generating artificial data.
- For its calibration and validation.
- For testing on external databases.

In 2023, the European Commission published a call for projects in the health cluster of Horizon Europe to propose methods for in silico simulations, modeling, and clinical trials with the aim of being compatible and accepted from a regulatory point of view for small rare disease populations, pediatric populations, etc. Two projects were funded and started in 2024, ERAMET[43] and INVENTS[44]. In the coming years, methods will therefore be available in this context.

43 https://cordis.europa.eu/project/id/101137141
44 https://ecrin.org/projects/invents

The use of artificial patient data is likely to become more prevalent in clinical studies. It is still up to scientists to identify the conditions and limitations of their use and to provide evidence of their reliability and safety for the benefit of patients and the healthcare system.

Prof. Raphaël Porcher
Professor of Medicine and Hospital Practitioner,
Paris-Cité University,
PR[AI]RIE Chair

II. Modeling using a Bayesian approach without generating artificial data

An example of modeling usage, albeit without generating artificial data, for the evaluation of a drug's efficacy involves the development of secukinumab, an anti-IL-17A, for pediatric psoriasis. Although this example did not employ the generation of artificial cohorts, it relies on a predictive approach from a Bayesian meta-analysis, which could be used to generate data in a manner similar to ABC methods or atlas estimation described in Chapter 2. It seemed relevant to develop this example, as it represents one of the few instances of a dossier evaluation based on non-mechanistic modeling by health authorities.

The drug, initially developed for moderate to severe psoriasis in adults, was authorized in Japan (2014), Europe (2015), and the United States (2015). It is also prescribed for ankylosing spondylitis, psoriatic arthritis, and other inflammatory conditions, with over 500,000 patients treated worldwide.

For pediatric development, two randomized trials were conducted:

- One trial in severe psoriasis, with three treatment arms: secukinumab 150 mg, secukinumab 300 mg, active comparator (planned n = 160, actual n = 162 included).
- One trial in moderate psoriasis, with two treatment arms: secukinumab 150 mg, secukinumab 300 mg (planned n = 120, actual n = 84 included), but without an active comparator (or placebo).

Additionally, a modeling study was planned to extrapolate secukinumab's efficacy in moderate pediatric psoriasis using a Bayesian predictive meta-analysis model. The idea is to model the phenomenon using prior information from another population. The Bayesian modeling framework allows for the formal incorporation of this data from another population into a prior probability distribution. The model then combines the observed data in the population of interest with the prior distribution to yield a posterior distribution of the treatment effect. This approach uses predictions from other groups to propose an update of the prior based on what is observed. However, it does not rely on generating artificial data to construct a control group for the trial in pediatric moderate psoriasis.

The extrapolation from adults to children here addressed several key elements outlined in the ICH-E11A recommendations, namely:

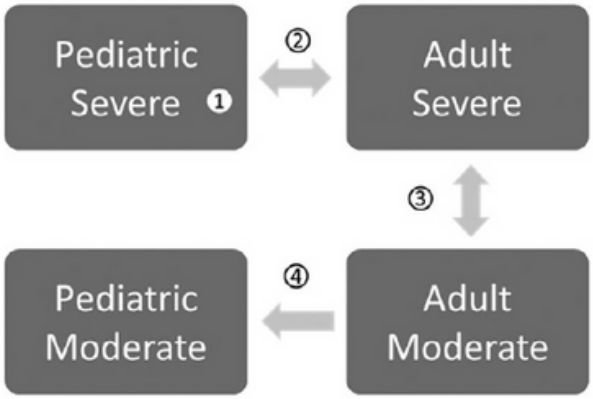
1. Similar pharmacology in children and adults.
2. Similar pathophysiology and disease progression in children and adults.
3. Similar therapeutic response in children and adults.

It is worth noting that the development of new dosage forms for children under six years old when PK-PD rules differ is a real challenge.

It had also been considered that secukinumab, used by a significant number of adults worldwide, had a reassuring safety profile.

The extrapolation from severe pediatric psoriasis to moderate pediatric psoriasis was done using a 4-step model, described in the figure below. Step 4 particularly involved a Bayesian joint model taking into account all data (adult and child). The dossier was finally submitted to the EMA in 2019 and to the FDA in 2020, and authorization for the moderate pediatric forms was obtained both in Europe (EMA, 2020) and in the United States (2021). However, in France, the initial opinion of the Transparency Commission[45] was favorable only for the treatment of psoriasis in children and adolescents failing at least two treatments and in cases of extensive form and/or significant psychosocial impact. The medical service provided was considered insufficient in other forms. This opinion was recently re-evaluated following the results of a post-marketing study, with a favorable opinion for reimbursement in the indication of its marketing authorization and a medical service considered important.[46] It should be noted that in the November 2021 opinion, the Bayesian predictive meta-analysis model allowing for indirect estimation of the effect against placebo was deemed insufficiently robust.

In conclusion, even though in this example no artificial patient data was used and the Transparency Commission did not consider the predictive meta-analysis estimating the effect against placebo, this latter analysis was deemed convincing enough by the EMA and the FDA. Therefore, it is conceivable to consider a use case for artificial cohorts to augment clinical trial data, provided that the methodology employed is deemed sufficiently robust.



Copyright You et al. Clin Pharmacol Ther. 2022 Mar;111(3):697-704.

45 https://www.has-sante.fr/upload/docs/evamed/CT-18841_COSENTYX_enfant_PIC_EI_AvisDef_CT18841&18848.pdf (3 nov. 2021)
46 https://www.has-sante.fr/jcms/p_3426148/fr/cosentyx-secukinumab-psoriasis-en-plaques-chez-l-enfant-et-l-adolescent (15 mars 2023)

It's cruel to have to abandon the development of a promising drug for a rare disease due to the impracticality of a clinical trial with a sufficient number of patients. Artificial data offer a solution by supplementing trials with virtual patients providing the required statistical evidence.

Dr. Jérôme Kalifa
President et Founder of Let it Care

III. Evaluation questions raised by artificial data

Beyond the aspects developed in Section I, the use of artificial data raises specific questions for which practice will identify the necessary prerequisites or appropriate recommendations. Several points can already be considered.

Many methods presented in Chapter 2 have demonstrated their ability to reproduce data with characteristics, in terms of probability distribution, that are very similar to those used for model training. The first question then arises as to what level and quantity of evidence are necessary to consider an approach sufficiently reliable and validated for practical use. Two different approaches can be envisaged. Either the validation of a method once it has proven itself in a certain number of situations (particularly in the targeted application situation, for example the targeted indication, or a related disease), or qualification by a competent body. This could also depend on the types of judgment criteria used, the objectives of using artificial data, and the indication, taking into account, among other things, the prevalence of the disease (rare diseases), its proximity to other already modeled pathologies, the existence of validated physiological models, and the target population (such as an indication targeting a fragile population, for example).

It is thus possible to envisage different levels of requirement depending on the contexts of use. Data augmentation approaches to train image analysis algorithms, for example, or prediction for diagnostic or prognostic applications may require less evidence of the algorithm's ability to generate data similar to those observed than those consisting of augmenting the data from the control group - or even the experimental group - of a clinical trial with artificial data. Indeed, in the former case, algorithm performance measures on a test data set could suffice to quantify the interest of artificial data, especially since independent test data will be necessary, whereas in the latter case, artificial data should represent what could be observed if other participants had been recruited into the study.

The third element concerns the question of dataset size, which includes the minimum size of training datasets necessary, although not sufficient, for an algorithm to be performant, but also the maximum size of an artificial dataset that can reasonably be used for this training. Here too, these maximum sizes could vary depending on the context of use. Methods will need to be proposed to evaluate them reliably and robustly. The representativeness of the database is also a key factor.

The fourth element, finally, relates to the clinical validation of the approach if these artificial data are used within the framework of a clinical decision support learning system (diagnostic, therapeutic, preventive...) and therefore related to a digital medical device (DMD) for professional use. In this case, it is important not only to evaluate the algorithm itself but also the entire systems with which it will interact in terms of software, healthcare system, and human. Recently, the HAS[47]published a guide for professionals using DMDs.[48] Although this guide does not specifically address simulated data, it can serve as a first basis for reflection.

To provide answers to these questions, it seems crucial that examples of the use of artificial data in different contexts be published. In this regard, studies demonstrating the ability of artificial data to reproduce real situations are important. Additionally, these applications need to be complemented by more methodological work to evaluate the limitations of using these promising approaches and the milestones to be put in place to allow initial uses. The elements mentioned above can serve as prerequisites for producing and using artificial or augmented cohorts in which we can have confidence, and help the ecosystem to adopt these approaches. Ultimately, this should enable competent authorities and evaluation agencies to establish clear rules for the use and acceptability of these artificial cohorts, particularly for market access or reimbursement applications, knowing that they will have to offer the same level of guarantee as current reference methods. Moreover, artificial data could be taken into account in care pathway modeling, for example. This white paper thus lays the foundation for this endeavor.

It will be important to refer to the brand-new article by The GSP (Good Simulation Practices) Task Force, published on February 24, 2024, by the Avicenna Alliance, proposing:

- A better definition of the scope, actively seeking the broadest possible engagement with all organizations representing stakeholders (universities, industry, patients, regulators, and payers);
- A wider dissemination of work done so far;
- A consensus process through the In Silico World community of practice.

47 48 https://www.has-sante.fr/jcms/p_3363066/fr/dispositifs-medicaux-numeriques-a-usage-professionnel

Increasing cohorts through AI is a tremendous lever to provide solutions in therapeutic areas that are still too often under-addressed (rare or slowly progressing diseases) or for populations that are less evaluated (pediatrics, geriatrics). This white paper initiates reflection on demonstrating the value and conditions of use of these tools for confident use in clinical research. It is an essential step to enable patients to benefit from the effective and safe treatments they need.

Dr. Camille Schurtz
Head of Regulatory Processes and Market Access,
Agence de l'Innovation en Santé
Co-pilot of the AIS/F-CRIN Working Group (The evolution of clinical trial methodologies: new tools, new uses, and conditions of use)

09

ISSUES AND ETHICAL GUARANTEES

This White Paper on artificial data sheds light on a major transformation underway in our healthcare system, which will notably accelerate therapeutic innovation significantly. It also incorporates ethical regulation by design by recommending the implementation of human assurance mechanisms by healthcare professionals and patient representatives, in accordance with the French bioethics law of 2021 and the new European regulation on AI.

David Gruson
Founder,
ETHIK-IA

The use of virtual patient cohorts represents a significant opportunity for advancing medical research and improving the quality of patient care. This essential innovation in learning methods and clinical methodology, however, raises some ethical questions, as well as concerns about robustness as previously discussed, and must be situated within a framework of positive regulation in connection with the use of AI models for creating these artificial patients. It is therefore important to consider several key points.

I. Human oversight of artificial patient cohorts (referring to Article 14 of the AI Act).

Ethically, while the quality and choice of data remain central, the validation of mathematical models and algorithms is equally important. Like any data processing, the implementation of artificial patient cohorts requires human oversight. This principle of human oversight ("Human Guarantee" or "human control") refers to the need not to relinquish all autonomy of action or decision-making in a context of increasingly rapid dissemination of artificial intelligence.

Concretely, experts will need to validate that the created cohorts are in line with the initial patient groups. Organized in the form of human oversight committees, these control measures will help better understand the phase of modeling artificial patient populations and ensure that it is as unbiased and reliable as possible.

This continuous monitoring can occur both during the cohort creation phase and post-creation throughout the use of these databases. For example, experts could supervise and validate collected data, as well as monitor, assist in the calibration, and adjust mathematical models and algorithmic tools used in creating the patients to ensure that the artificial patient cohort is as realistic as possible. Validation of a random sample of patients may be requested to verify that every artificial patient could be a real patient. Finally, it is absolutely imperative that human oversight is conducted on the results of the uses made from these artificial patient cohorts to detect potential errors or unexpected scenarios.

The implementation of these human guarantee principles may include, in addition to clinicians in the relevant discipline, representatives of patients from the cohort or a similar indication to provide complementary validation. Real patients from the augmented cohort could thus be better informed about the statistical treatment of their data, helping to foster public adoption of these technologies and openness of health data for research and innovation.

II. Could this lead to a new essential principle in research ethics involving the human person and a revision of ethical standards?

The ethical stakes involved in the development of these new practices go beyond merely weighing the acceleration of progress at a lower financial cost on one side against the need to protect health data and the rights and freedoms of the individuals concerned on the other, a question that will likely be resolved with the expected advances in anonymization. Once the scientific and technical questions of scientific integrity and reliability are resolved, and thus the equivalence of guaranteed results, the range of possibilities offered in terms of replacing human subjects in interventional research practices appear as formidable tools for protecting individuals. Thus, akin to regulations on animal experimentation, it could quickly become possible to impose an initial reflection step on the scientific, ethical, and societal validity of real patient participation in clinical trials, and even more so the participation of healthy volunteers. This reflection, which is already mandated by international research ethics principles and incorporated into French law for individuals in situations of particular vulnerability such as children, pregnant women, the elderly, or those without social coverage, too often results in the exclusion of these populations from clinical trials targeting the general population, and the difficulty of funding specific trials that are nonetheless essential. This leads to inequitable access to innovation, or at the very least, inequality in the conditions of safety for such access, when offered outside of marketing authorization, at the cost of increased risks and under the responsibility of prescribers.

A broader and fairer policy of replacing human subjects could therefore be implemented in the medium term, whenever technically feasible, with the aim of reducing the number of individuals exposed to risks, even minimal ones, and/or constraints, without prejudice to their access to therapeutic innovation. In the case of phase III trials, the added value of real participant involvement is clearly weighed against participants' lack of interest in the control arm, and thus the difficulties in retaining them in the trial. These reservations, which can at most be manifested by an outright refusal to enter the trial or a retraction upon learning of randomization into the control arm, may be accompanied by a clear demand for access to the active principle being tested, as long as it is promising. The uncertainty of its efficacy, which justifies randomization for scientists, is rather perceived by patients as an unwarranted obstacle to accessing the product, seen as a loss of opportunity. Conversely, replacing certain patients with virtual patients in the arms of tested products would reduce individuals' exposure to risks, while still conducting trials and accessing results, thus either leading to the quicker abandonment of dangerous or unnecessary products or access to validated treatments.

Such a policy of reducing interventions that compromise the integrity of the human body without direct therapeutic benefit to them, or with the aim of reducing the risks of an intervention they specifically need, should apply equally to the educational context or the securing of individual care. Thus, techniques for producing virtual patients, digital twins of real patients, already enable surgeons or other interventional physicians to learn the most invasive procedures without physical risk to patients and while respecting their dignity.

Of course, such changes in practices must be accompanied not only by validated scientific frameworks but also by a renovation of research ethics standards. Moreover, beyond a possible obligation to reflect on reducing the number of individuals exposed to clinical trials without benefit to them, new duties of information need to be considered, especially regarding individuals who will receive validated treatments under these new conditions. Strengthened post-marketing pharmacovigilance obligations could, in addition to human guarantee colleges, help secure and validate these techniques. Finally, a revision of guidelines for research ethics committees (CPP in France or IRB), as well as training for their members, seems necessary. A parallel reflection on the status of virtual cohorts once created, their scope of reusability, and the conditions for this reusability could be envisaged with a triple objective of public interest, parsimony, and consideration of environmental issues. In this regard, it is also likely that virtual patient cohorts for research will further reduce the use of animal experimentation.

Furthermore, it is necessary in the context of retrospective research to have the SIA validated to prove the effectiveness of artificial cohorts before they are used in real life for new research (prospective research). This notably implies that the algorithm development takes place within the framework of a rigorous scientific protocol, evaluated by an ethical committee, and respectful of the privacy of the individuals concerned. This proof of concept must be published for peer review.

Artificial data offers a promising opportunity for our healthcare system and could play an essential role in structuring long-term care pathways. They tangibly embody the convergences between mathematics, medicine, and digital technology.

Dr. Yann-Maël Le Douarin
Medical Advisor at the DGOS and Head of the Health and Digital Transformation Department,
Ministry of Labor, Health and Social Affairs

CONCLUSION

07

This white paper shows us how artificial data could ultimately play an important role in clinical research and the optimization of health algorithms in particular.

This concerns not only diagnostic, prognostic, and therapeutic issues for patients and healthcare professionals, but also a competitiveness and added value issue for France and Europe.

The generation of artificial data is an example of secondary use of health data as described in the report "Bringing together ecosystem actors to unleash the secondary use of health data" written by Mr. Jérôme Marchand-Arvier, Prof. Stéphanie Allassonnière, Mr. Aymeril Hoang, and Dr. Anne-Sophie Jannot.[49]

In this document, we have shown that there are already foundations for opening up the work needed to define and implement validation of these "new genre patients." This includes the establishment of human guarantee boards (to validate the reliability and safety of data produced by artificial intelligence) and of course their methodological and scientific recognition for in silico research.

The Health Innovation Agency as well as the National Research Agency are currently leading working groups on the use of new clinical research methodologies to stimulate France's attractiveness in this strategic area supported by France 2030. Moreover, the interministerial report on "unleashing the secondary use of health data" published by the IGAS in January 2024 as well as the recommendations submitted to the President of the Republic in March by the Artificial Intelligence Commission also help to advance these subjects rapidly.

This white paper shows that all stakeholders are concerned and feel concerned and understand the potential of these new technologies, but also the need for regulation.

We hope to have shed some light on this future topic.

[49] https://sante.gouv.fr/IMG/pdf/rapport_donnees_de_sante.pdf

In its opinion dated April 2, 2024, the National Consultative Ethics Committee for Health and Life Sciences (CCNE) recalls "the imperative need to protect participants in clinical trials." New approaches to medical research, notably artificial patients generated from a small number of real patients, will contribute to this objective.

Dr. Jean-Louis Fraysse
Co-founder of BOTdesign
Member of the Ethics Group of the Digital Health Delegation and of the Reflection Group on these New Clinical Research Methodologies of AIS and F-CRIN.

08

BIBLIOGRAPHICAL REFERENCES

1. Rebuffi, S.-A. et al. Data Augmentation Can Improve Robustness. Preprint at https://doi.org/10.48550/arXiv.2111.05328 (2021).
2.Kingma, D. P. & Welling, M. Auto-Encoding Variational Bayes. (2013).
3.Goodfellow, I. J. et al. Generative Adversarial Networks. Preprint at https://doi.org/10.48550/arXiv.1406.2661 (2014).
4.Sohl-Dickstein, J., Weiss, E. A., Maheswaranathan, N. & Ganguli, S. Deep Unsupervised Learning using Nonequilibrium Thermodynamics. ArXiv150303585 Cond-Mat Q-Bio Stat (2015).
5.Hu, Q., Li, H. & Zhang, J. Domain-Adaptive 3D Medical Image Synthesis: An Efficient Unsupervised Approach. in Medical Image Computing and Computer Assisted Intervention – MICCAI 2022 (eds. Wang, L., Dou, Q., Fletcher, P. T., Speidel, S. & Li, S.) 495–504 (Springer Nature Switzerland, 2022). doi:10.1007/978-3-031-16446-0_47.
6.Diamantis, D. E., Gatoula, P. & Iakovidis, D. K. EndoVAE: Generating Endoscopic Images with a Variational Autoencoder. in 2022 IEEE 14th Image, Video, and Multidimensional Signal Processing Workshop (IVMSP) 1–5 (2022). doi:10.1109/IVMSP54334.2022.9816329.
7. Barbano, R., Arridge, S., Jin, B. & Tanno, R. Chapter 26 - Uncertainty quantification in medical image synthesis. in Biomedical Image Synthesis and Simulation (eds. Burgos, N. & Svoboda, D.) 601–641 (Academic Press, 2022). doi:10.1016/B978-0-12-824349-7.00033-5.
8.Chadebec C, Thibeau-Sutre E, Burgos N, and Allassonnière S. Data Augmentation in High Dimensional Low Sample Size Setting Using a Geometry-Based Variational Autoencoder. IEEE Pattern Analysis and Machine Intelligence..
9.Rhodes, C. The Man With 100,000 Brains: AI's Big Donation to Science. NVIDIA Blog https://blogs.nvidia.com/blog/2022/05/30/ai-brain-images-kcl/ (2022).
10.Nie, D. et al. Medical Image Synthesis with Deep Convolutional Adversarial Networks. IEEE Trans. Biomed. Eng. 65, 2720–2730 (2018).
11.Wolterink, J. M. et al. Deep MR to CT Synthesis using Unpaired Data. ArXiv170801155 Cs (2017).
12.Hu, Y. et al. Freehand Ultrasound Image Simulation with Spatially-Conditioned Generative Adversarial Networks. in vol. 10555 105–115 (2017).
13.Chuquicusma, M. J. M., Hussein, S., Burt, J. & Bagci, U. How to Fool Radiologists with Generative Adversarial Networks? A Visual Turing Test for Lung Cancer Diagnosis. Preprint at https://doi.org/10.48550/arXiv.1710.09762 (2018).
14.Baur, C., Albarqouni, S. & Navab, N. MelanoGANs: High Resolution Skin Lesion Synthesis with GANs. Preprint at https://doi.org/10.48550/arXiv.1804.04338 (2018).
15.Wolterink, J. M., Leiner, T. & Isgum, I. Blood Vessel Geometry Synthesis using Generative Adversarial Networks. Preprint at https://doi.org/10.48550/arXiv.1804.04381 (2018).
16.Skandarani, Y., Jodoin, P.-M. & Lalande, A. GANs for Medical Image Synthesis: An Empirical Study. Preprint at https://doi.org/10.48550/arXiv.2105.05318 (2021).
17.Khader, F. et al. Medical Diffusion: Denoising Diffusion Probabilistic Models for 3D Medical Image Generation. Preprint at https://doi.org/10.48550/arXiv.2211.03364 (2023).

18.Kazerouni, A. et al. Diffusion models in medical imaging: A comprehensive survey. Med. Image Anal. 88, 102846 (2023).
19.Pan, S. et al. 2D medical image synthesis using transformer-based denoising diffusion probabilistic model. Phys. Med. Biol. 68, 105004 (2023).
20.Dorjsembe, Z., Pao, H.-K., Odonchimed, S. & Xiao, F. Conditional Diffusion Models for Semantic 3D Medical Image Synthesis. Preprint at https://doi.org/10.48550/arXiv.2305.18453 (2023).
21.Weber, T., Ingrisch, M., Bischl, B. & Rügamer, D. Cascaded Latent Diffusion Models for High-Resolution Chest X-ray Synthesis. Preprint at https://doi.org/10.48550/arXiv.2303.11224 (2023).
22.Peirlinck, M. et al. Precision medicine in human heart modeling. Biomech. Model. Mechanobiol. 20, 803–831 (2021).
23.Applied Sciences[35] | Free Full-Text | MUSIC: Cardiac Imaging, Modelling and Visualisation Software for Diagnosis and Therapy. https://www.mdpi.com/2076-3417/12/12/6145.
24.Nakatani, Y. et al. Preoperative personalization of atrial fibrillation ablation strategy to prevent esophageal injury: Impact of changes in esophageal position. J. Cardiovasc. Electrophysiol. 33, 908–916 (2022).
25.Banus, J., Lorenzi, M., Camara, O. & Sermesant, M. Biophysics-based statistical learning: Application to heart and brain interactions. Med. Image Anal. 72, 102089 (2021).
26.Levine, S. et al. Dassault Systèmes' Living Heart Project. in Modelling Congenital Heart Disease: Engineering a Patient-specific Therapy (eds. Butera, G., Schievano, S., Biglino, G. & McElhinney, D. B.) 245–259 (Springer International Publishing, 2022). doi:10.1007/978-3-030-88892-3_25.
27.Segars, W. P., Veress, A. I., Sturgeon, G. M. & Samei, E. Incorporation of the Living Heart Model Into the 4-D XCAT Phantom for Cardiac Imaging Research. IEEE Trans. Radiat. Plasma Med. Sci. 3, 54–60 (2019).
28.Kaboudian, A., Cherry, E. M. & Fenton, F. H. Real-time interactive simulations of large-scale systems on personal computers and cell phones: Toward patient-specific heart modeling and other applications. Sci. Adv. 5, eaav6019 (2019).
29.Peterlík, I., Duriez, C. & Cotin, S. Modeling and Real-Time Simulation of a Vascularized Liver Tissue. in Medical Image Computing and Computer-Assisted Intervention – MICCAI 2012 (eds. Ayache, N., Delingette, H., Golland, P. & Mori, K.) 50–57 (Springer, 2012). doi:10.1007/978-3-642-33415-3_7.
30.Plantefève, R., Haouchine, N., Radoux, J.-P. & Cotin, S. Automatic Alignment of Pre and Intraoperative Data Using Anatomical Landmarks for Augmented Laparoscopic Liver Surgery. in Biomedical Simulation (eds. Bello, F. & Cotin, S.) 58–66 (Springer International Publishing, 2014). doi:10.1007/978-3-319-12057-7_7.
31.Mendizabal, A., Márquez-Neila, P. & Cotin, S. Simulation of hyperelastic materials in real-time using deep learning. Med. Image Anal. 59, 101569 (2020).

32.Brunet, J.-N. et al. Physics-Based Deep Neural Network for Augmented Reality During Liver Surgery. in Medical Image Computing and Computer Assisted Intervention – MICCAI 2019 (eds. Shen, D. et al.) 137–145 (Springer International Publishing, 2019). doi:10.1007/978-3-030-32254-0_16.
33.Pulkkinen, A., Werner, B., Martin, E. & Hynynen, K. Numerical simulations of clinical focused ultrasound functional neurosurgery. Phys. Med. Biol. 59, 1679 (2014).
34.Jaroudi, R., Åström, F., Johansson, B. T. & Baravdish, G. Numerical simulations in 3-dimensions of reaction–diffusion models for brain tumour growth. Int. J. Comput. Math. 97, 1151–1169 (2020).
35.Santaniello, S., Gale, J. T. & Sarma, S. V. Systems approaches to optimizing deep brain stimulation therapies in Parkinson's disease. WIREs Syst. Biol. Med. 10, e1421 (2018).
36.Johanns, P. et al.[35] The strength of surgical knots involves a critical interplay between friction and elastoplasticity. Sci. Adv. 9, eadg8861 (2023).
37.Boussès, Y., Brulat-Bouchard, N., Bouchard, P.-O., Abouelleil, H. & Tillier, Y. Theoretical prediction of dental composites yield stress and flexural modulus based on filler volume ratio. Dent. Mater. 36, 97–107 (2020).
38.Digital connection in the metaverse | Meta. Digital connection in the metaverse https://about.meta.com/metaverse/.
39.NVIDIA Omniverse. NVIDIA https://www.nvidia.com/en-us/omniverse/.
40.Cardoso, M. J. et al. MONAI: An open-source framework for deep learning in healthcare. Preprint at https://doi.org/10.48550/arXiv.2211.02701 (2022).
41.Baowaly, M. K., Lin, C.-C., Liu, C.-L. & Chen, K.-T. Synthesizing electronic health records using improved generative adversarial networks. J. Am. Med. Inform. Assoc. JAMIA 26, 228–241 (2018).
42.Ghosheh, G., Li, J. & Zhu, T. A review of Generative Adversarial Networks for Electronic Health Records: applications, evaluation measures and data sources. Preprint at https://doi.org/10.48550/arXiv.2203.07018 (2022).
43.Li, R. et al. Improving an Electronic Health Record–Based Clinical Prediction Model Under Label Deficiency: Network-Based Generative Adversarial Semisupervised Approach. JMIR Med. Inform. 11, e47862 (2023).
44.Li, J., Cairns, B. J., Li, J. & Zhu, T. Generating synthetic mixed-type longitudinal electronic health records for artificial intelligent applications. Npj Digit. Med. 6, 1–18 (2023).
45.Kotelnikov, A., Baranchuk, D., Rubachev, I. & Babenko, A. TabDDPM: Modelling Tabular Data with Diffusion Models. Preprint at http://arxiv.org/abs/2209.15421 (2022).
46.He, H., Zhao, S., Xi, Y. & Ho, J. C. MedDiff: Generating Electronic Health Records using Accelerated Denoising Diffusion Model. Preprint at https://doi.org/10.48550/arXiv.2302.04355 (2023).
47.Nikolentzos, G., Vazirgiannis, M., Xypolopoulos, C., Lingman, M. & Brandt, E. G. Synthetic electronic health records generated with variational graph autoencoders. Npj Digit. Med. 6, 1–12 (2023).

48.Liao, W. et al. Dual autoencoders modeling of electronic health records for adverse drug event preventability prediction. Intell.-Based Med. 6, 100077 (2022).
49.Muller, E., Zheng, X. & Hayes, J. Evaluation of the Synthetic Electronic Health Records. Preprint at https://doi.org/10.48550/arXiv.2210.08655 (2022).
50.Yan, C. et al. A Multifaceted Benchmarking of Synthetic Electronic Health Record Generation Models. Nat. Commun. 13, 7609 (2022).
51.Kuo, N. I.-H., Jorm, L. & Barbieri, S. Synthetic Health-related Longitudinal Data with Mixed-type Variables Generated using Diffusion Models. Preprint at https://doi.org/10.48550/arXiv.2303.12281 (2023).
52.Biswal, S. et al. EVA: Generating Longitudinal Electronic Health Records Using Conditional Variational Autoencoders. Preprint at https://doi.org/10.48550/arXiv.2012.10020 (2020).
53.Lee, D. et al. Generating sequential electronic health records using dual adversarial autoencoder. J. Am. Med. Inform. Assoc. 27, 1411–1419 (2020).
54.Theodorou, B., Xiao, C. & Sun, J. Synthesize High-dimensional Longitudinal Electronic Health Records via Hierarchical Autoregressive Language Model. Preprint at https://doi.org/10.48550/arXiv.2304.02169 (2023).
55.Koval, I. et al. Forecasting individual progression trajectories in Huntington disease enables more powered clinical trials. Sci. Rep. 12, 18928 (2022).
56.Sauty, B. & Durrleman, S. Riemannian Metric Learning for Progression Modeling of Longitudinal Datasets. in 2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI) 1–5 (2022). doi:10.1109/ISBI52829.2022.9761641.
57.Schiratti, J.-B., Allassonnière, S., Colliot, O. & Durrleman, S. Mixed-effects model for the spatiotemporal analysis of longitudinal manifold-valued data. in 5th MICCAI Workshop on Mathematical Foundations of Computational Anatomy (2015).
58.Shahriar, S. & Hayawi, K. Let's have a chat! A Conversation with ChatGPT: Technology, Applications, and Limitations. Preprint at https://doi.org/10.48550/arXiv.2302.13817 (2023).
59.Liu, Y. et al. Summary of ChatGPT-Related Research and Perspective Towards the Future of Large Language Models. Meta-Radiol. 1, 100017 (2023).
60.OpenAI. GPT-4 Technical Report. Preprint at https://doi.org/10.48550/arXiv.2303.08774 (2023).
61.Touvron, H. et al. LLaMA: Open and Efficient Foundation Language Models. Preprint at https://doi.org/10.48550/arXiv.2302.13971 (2023).
62.Tu, T. et al. Towards Generalist Biomedical AI. Preprint at https://doi.org/10.48550/arXiv.2307.14334 (2023).
63.Kline, A. et al. Multimodal machine learning in precision health: A scoping review. Npj Digit. Med. 5, 1–14 (2022).
64.Behrad, F. & Saniee Abadeh, M. An overview of deep learning methods for multimodal medical data mining. Expert Syst. Appl. 200, 117006 (2022).
65.Qiu, J. et al. Multimodal Representation Learning of Cardiovascular Magnetic Resonance Imaging. Preprint at https://doi.org/10.48550/arXiv.2304.07675 (2023).
66.Belyaeva, A. et al. Multimodal LLMs for health grounded in individual-specific data. Preprint at https://doi.org/10.48550/arXiv.2307.09018 (2023).

67.Bommasani, R. et al. On the Opportunities and Risks of Foundation Models. ArXiv210807258 Cs (2021).
68.Cranmer, K., Brehmer, J. & Louppe, G. The frontier of simulation-based inference. Proc. Natl. Acad. Sci. 117, 30055–30062 (2020).
69.Mirza Faisal Beg, Michael I. Miller, Alain Trouvé, Laurent Younes.Computing Large Deformation Metric Mappings via Geodesic Flows of Diffeomorphisms International Journal of Computer Vision 61(2):139-157
70.Jean-Michel Marin, Pierre Pudlo, Christian P. Robert, Robin J. Ryder. Approximate Bayesian computational methods. Stat Comput (2012) 22:1167–1180
71.Jesús Murga-Moreno, Sònia Casillas, Antonio Barbadilla, Lawrence Uricchio and David Enard. An efficient and robust ABC approach to infer the rate and strength of adaptation.
72.https://www.biorxiv.org/content/biorxiv/early/2023/09/28/2023.08.29.555322.full.pdf
73.Louis Raynal, Jean-Michel Marin, Pierre Pudlo, Mathieu Ribatet, Christian P Robert, Arnaud Estoup. ABC random forests for Bayesian parameter inference. Bioinformatics, Volume 35, Issue 10, May 2019, Pages 1720–1728,
74. Etienne Maheux, Igor Koval, Juliette Ortholand, Colin Birkenbihl, Damiano Archetti, Vincent Bouteloup, Stéphane Epelbaum, Carole Dufouil, Martin Hofmann-Apitius & Stanley Durrleman. Forecasting individual progression trajectories in Alzheimer's disease. Nature Communications 14, 761 (2023)
75. Liu G, Catacutan DB, Rathod K, Swanson K, Jin W, Mohammed JC, et al. Deep learning-guided discovery of an antibiotic targeting Acinetobacter baumannii. Nat Chem Biol. 2023;19: 1342-50.
76. Wong F, Zheng EJ, Valeri JA, Donghia NM, Anahtar MN, Omori S, et al. Discovery of a structural class of antibiotics with explainable deep learning. Nature. 2023.
77. European Medicines Agency. ICH guideline E11A on pediatric extrapolation. 2022.
78. Musuamba FT, Skottheim Rusten I, Lesage R, Russo G, Bursi R, Emili L, et al. Scientific and regulatory evaluation of mechanistic in silico drug and disease models in drug development: Building model credibility. CPT Pharmacometrics Syst Pharmacol. 2021;10: 804-25.
79. Azizi Z, Zheng C, Mosquera L, Pilote L, El Emam K, Collaborators G-F. Can synthetic data be a proxy for real clinical trial data? A validation study. BMJ Open. 2021;11: e043497.
80. Thorlund K, Dron L, Park JJH, Mills EJ. Synthetic and External Controls in Clinical Trials - A Primer for Researchers. Clin Epidemiol. 2020;12: 457-67.
81. Lambert J, Lengline E, Porcher R, Thiebaut R, Zohar S, Chevret S. Enriching single-arm clinical trials with external controls: possibilities and pitfalls. Blood Adv. 2023;7: 5680-90.
82. Bakker E, Plueschke K, Jonker CJ, Kurz X, Starokozhko V, Mol PGM. Contribution of Real-World Evidence in European Medicines Agency's Regulatory Decision Making. Clin Pharmacol Ther. 2023;113: 135-51.

83. Vanier A, Fernandez J, Kelley S, Alter L, Semenzato P, Alberti C, et al. Rapid access to innovative medicinal products while ensuring relevant health technology assessment. Position of the French National Authority for Health. BMJ Evid Based Med. 2024;29: 1-5.
84. Jacob E, Perrillat-Mercerot A, Palgen JL, L'Hostis A, Ceres N, Boissel JP, et al. Empirical methods for the validation of time-to-event mathematical models taking into account uncertainty and variability: application to EGFR + lung adenocarcinoma. BMC Bioinformatics. 2023;24: 331.
85. You R, Weber S, Bieth B, Vandemeulebroecke M. Innovative Pediatric Development for Secukinumab in Psoriasis: Faster Patient Access, Reduction of Patients on Control. Clin Pharmacol Ther. 2022;111: 697-704.
86. Neal ML, Trister AD, Cloke T, Sodt R, Ahn S, Baldock AL, et al. Discriminating survival outcomes in patients with glioblastoma using a simulation-based, patient-specific response metric. PLoS One. 2013;8: e51951.
87. Switchenko JM, Heeke AL, Pan TC, Read WL. The use of a predictive statistical model to make a virtual control arm for a clinical trial. PLoS One. 2019;14: e0221336.
88. Guillaudeux M, Rousseau O, Petot J, Bennis Z, Dein CA, Goronflot T, et al. Patient-centric synthetic data generation, no reason to risk re-identification in biomedical data analysis. NPJ Digit Med. 2023;6: 37.
89. Creemers JHA, Ankan A, Roes KCB, Schroder G, Mehra N, Figdor CG, et al. In silico cancer immunotherapy trials uncover the consequences of therapy-specific response patterns for clinical trial design and outcome. Nat Commun. 2023;14: 2348.
90.Senellart A, Chadebec C, Allassonnière S. Improving Multimodal Joint Variational Autoencoders through Normalizing Flows and Correlation Analysis. https://arxiv.org/abs/2305.11832
91. Tavaré, S., Balding, D., Griffith, R., Donnelly, P.: Inferring coalescence times from DNA sequence data. Genetics 145(2), 505–518 (1997)

THANK YOU

For more information:

stephanie.allassonniere@u-paris.fr
jlfraysse@botdesign.net