

RESEARCH

Open Access



# Validating methods for inferring co-occurring diseases: a flexible framework for simulating synthetic data

Hannah Marchi<sup>1</sup>, Sophie Schmiegel<sup>1</sup>, Tamara Schamberger<sup>1</sup> and Christiane Fuchs<sup>1,2\*</sup>

## Abstract

**Background** The validation of methods is an integral part of statistical research, defining conditions under which methods yield reliable results. Empirical validation requires a solid data basis to control and manage relevant characteristics like sample size, dimensionality, and underlying dependency structures. Real-world data often fails to meet these requirements, particularly in medical contexts where privacy regulations restrict availability. For this reason, synthetic data is an effective alternative for method validation. However, generating synthetic data is demanding when it must precisely mirror complex dependence structures while simultaneously controlling specific target characteristics.

**Methods** We address the medical context of co-occurring diseases, where symptoms may overlap or conflict. We propose a four-step framework to generate synthetic data for the simulation-based validation of statistical methods. The framework involves: (I) generating patient covariates; (II) connecting this information to predictors for single or joint disease occurrence; (III) transforming predictors into disease probabilities or scores; and (IV) converting these into disease occurrences. Each step offers several alternatives for modeling the overall dependence structure. We apply our framework to a case study of pain-causing diseases which share certain similarities in their clinical presentations, and which can occur either individually or jointly. By employing five combinations of methodological alternatives, we evaluate the approaches' ability to achieve target characteristics and demonstrate their specific strengths and weaknesses.

**Results** Matching the data-generating process with the estimation method allows for the successful recovery of input information, such as coefficients and correlations. Target properties like disease prevalence and associations are achieved to varying degrees depending on the methods used.

**Conclusions** While the proposed theory-driven framework is broadly applicable beyond the specific medical use case, it relies on careful, domain-informed parameter curation to generate meaningful synthetic datasets. Its flexible, adjustable input settings enable researchers to tailor data generation to their precise methodological requirements, providing a controlled basis for simulation-based validation without implying direct clinical inference.

**Keywords** Synthetic data generation, Method validation, Simulation studies, Co-occurring events, Comorbidities

\*Correspondence:  
Christiane Fuchs  
christiane.fuchs@uni-bielefeld.de

<sup>1</sup>Data Science Group, Faculty of Business Administration and Economics, Bielefeld University, Bielefeld, Germany

<sup>2</sup>Institute of Computational Biology, Computational Health Center, Helmholtz Munich, Munich, Germany



© The Author(s) 2026. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## Background

A relevant part of statistics is inferential statistics, where conclusions are drawn from a data sample about an underlying population. An important subarea within this field is the discovery and learning of relationships between observed or unobserved random variables. To study such relationships, a wide range of statistical methods exists, where the suitability of each method depends on the assumptions about the underlying ground truth as well as on the availability and characteristics of the data. It frequently occurs that problems of interest are structured in such a way that existing inference methods cannot be applied directly, which leads to the development of new approaches. For both newly developed and already established methods, method validation is a crucial step of the model building process [1] and model evaluation [2]. By validation, we refer to the process of assessing how well a statistical procedure performs for its intended purpose, i. e. whether the method produces accurate, reliable, and meaningful results under the conditions in which it will be used [3].

In some cases, such validation is achieved through theoretical considerations, which typically rely on assumptions and often involve asymptotic arguments. Frequently, however, validation is carried out computationally, e. g. by means of a simulation study. In the best case, such a study is based on controllable and idealized data in the following general sense: First, a ground truth exists and is known. Second, sources of noise can be influenced and, in particular, minimized. Third, the sample size can be increased arbitrarily.

Real-world data generally differs from such requirements: ground truth such as true effect sizes or dependency structures is often unknown, variability is hardly controlled, and sample sizes are limited by practical constraints. Such limitations can be overcome by using synthetic data for method validation [4]: being generated in silico under well-defined assumptions, synthetic datasets enable the systematic evaluation of statistical methods, for example by assessing the bias of estimates, robustness to violations of assumptions, calibration of models, or sensitivity of results to different dependency structures. Consequently, synthetic data serve as a flexible and powerful tool for validating statistical methods, especially in settings where real-world datasets are unavailable, incomplete, or insufficient for rigorous performance assessment. [4] present a tutorial on planning, executing and reporting simulation studies for method validation.

The use of synthetic data for statistical method validation is relevant across many applied disciplines. In medical research in particular, it has gained increasing importance, for example if datasets contain only few observations due to rare diseases or small populations, or due to measurement cost. Moreover, medical data

may lack the required quality for method validation as a result of insufficient documentation, substantial amounts of missing values or measurement error, caused by time constraints in clinical routine, or inconsistent measurement intervals [5]. Although data imputation can be a viable solution against missing values, model misspecification or incompatibility can introduce bias and misestimate uncertainty [6, 7]. In addition, privacy and data protection issues present major challenges in medical data, such that it is often publicly inaccessible or highly anonymized, consequently lacking relevant information.

Due to the substantial need for synthetic data in medical research, several approaches to generate such data have been presented. Specifically, synthetic data has been generated for specific diseases, such as diabetes [8], liver diseases [9] and depression [10]. Beyond disease-specific applications, recent research increasingly focuses on evaluating whether synthetic health data preserve relevant population characteristics and protect patient privacy while maintaining analytical utility. In particular, [11] address the challenge of limited patient data availability for machine learning in medicine and propose evaluation metrics to assess both data utility and privacy risk. They emphasize that synthetic patient data should retain key population-level structures while simultaneously ensuring privacy protection. [12] further introduce a novel metric to measure the trade-off between utility and privacy. Extending this perspective beyond purely statistical fidelity, [13] develop a pipeline to make synthetic patient trajectories clinically consistent. More broadly, [14] provide a structured overview of synthetic data use in healthcare and identify multiple application domains, including e. g. simulation and prediction research, algorithm testing, epidemiology/public health research, education and training and linking data. These use cases highlight the heterogeneous objectives of synthetic data generation, ranging from empirical realism to controlled methodological experimentation and model evaluation. Additional comprehensive reviews of synthetic data generation and evaluation in the health care domain can be found in [15, 16].

When considering specific applications, the requirements for synthetic data become more concrete, extending beyond the general demands of knowledge of the ground truth, controllability of noise, and sample size scalability. In many contexts, it is not expected that synthetic data is medically correct in every detail in the sense that medical conclusions can be drawn from it; rather, the data should serve the purpose of the specific simulation study, namely, the validation of a statistical method for a particular research question [4]. If a research question addresses the relationship between clinical patient characteristics and a diagnosis, one would choose a statistical method capable of analyzing such relationships.

Accordingly, this relationship should be represented in the synthetic data and be controllable during data generation, ranging from none to moderate to strong association. In this way, the method can be validated and subsequently applied to real data with the confidence that it can be relied upon.

In this work, we consider a medical context that involves specific dependencies within the data: namely, the holistic examination of patients in whom multiple diseases may occur simultaneously and overall form a joint clinical picture. Thus, while the aforementioned studies aim to produce medical correct synthetic data for specific diseases, we seek to generate synthetic data that mirrors the interdependency between disease profiles, connected to patient information. In particular, we do not aim for our generated data to be medically accurate or capable of replicating complex medical relationships in a way that allows valid clinical conclusions to be drawn. Instead, our approach is intentionally generic. It is designed to allow users to model the specific requirements of their individual statistical use cases, particularly the overarching characteristics of a target population.

The motivation for this research focus arises from the fact that, in reality, patients frequently suffer from several co-occurring diseases (of varying severity). These diseases present in similar, different or even antagonistic symptoms. This is particularly relevant as, in practice, establishing one diagnosis may lead to the premature exclusion of an alternative one. Such failure to recognize a co-occurring disease can have serious consequences. Examples for known co-occurring diseases with similar symptoms are diabetes mellitus and cardiovascular diseases [17], depression and anxiety disorders [18] or the fibromyalgia syndrome (FMS) and irritable bowel syndrome [19].

Our motivation for examining co-occurring diseases arose from a project in which we were concerned with the diagnosis of patients suffering from chronic pain. Such patients are often affected by multiple co-occurring diseases. Thus, we analyzed patient data with the aim to investigate whether the predictive accuracy of a disease of interest could be improved by joint consideration of comorbidities. We compared the prediction performance of two competing statistical methods: single-label classification and multi-label classification [20]. Our analysis provided answers to the specific research question, namely, that the inclusion of additional diagnoses did not yield substantial knowledge gain in this particular case. However, based on the real data, it was impossible to investigate under which conditions (in terms of the true underlying effect sizes, as well as data quality and sample size) one classification approach would have outperformed the other. We consider such an assessment to be highly relevant and therefore aim to conduct a simulation

study using synthetic data. The present work lays the foundation for this.

Practical challenges in conducting simulation studies and, consequently, in finding or generating synthetic data for a specific research question lie in various aspects and also depend on the researcher's level of domain-specific expertise: With regard to the medical background, relevant relationships must be appropriately identified and described. These relationships need to be statistically modeled, often by specifying a data-generating process, with particular attention typically given to conditional and unconditional as well as joint and marginal distributions. Even when such a data-generating process is known, its exact simulation often requires the use of approximate algorithms. Finally, the procedures need to be implemented using statistical software.

Ideally, suitable tools and software packages would be available for straightforward simulation of synthetic data. However, the diversity of application-specific requirements makes it challenging to develop universally applicable tools: each comes with a different focus and its own strengths. Existing approaches for generating synthetic clinical data can broadly be grouped into three categories, each with a distinct objective. First, correlation-based methods (e. g. copula [21], bin-data [22], *SimMultiCorrData* [23], and *simdata* [24]) aim to reproduce predefined marginal distributions and correlation structures across mixed variable types. These methods are well suited when the primary goal is to approximate a joint distribution with specified dependence properties. However, regression parameters are no typical primary modeling targets; consequently, conditional effect sizes arise indirectly from the imposed joint dependence structure rather than being explicitly parameterized in interpretable disease models.

Second, regression-based simulation tools (e. g. R packages *simstudy* [25], *SimCorMultRes* [26]) focus on generating outcomes conditionally on covariates using generalized linear models. These approaches allow direct control of regression coefficients and are well suited for evaluating estimator performance under prespecified conditional models. However, they struggle to simultaneously control the complete global correlation structure.

Third, data-driven synthetic data generators (e. g. the software *synthea* [27], the R package *synthpop* [28], and related resampling approaches) rely on real-world datasets as a basis for generating synthetic observations. While these methods can achieve high empirical realism, they inherit the structural assumptions and dependencies present in the source data and are therefore less suited for purely theoretical or counterfactual simulation scenarios in which specific marginal distributions, prevalences, or regression parameters must be imposed independently of observed data.

In our work, we develop, describe and evaluate a process for generating synthetic medical data with the aim to support statistical method validation in the context of co-occurring diseases. Specifically, we design a four-step framework, with each step building on the results from the previous one. Within each step, we propose alternative approaches that can be combined across the different steps in a modular manner. Our framework differs conceptually from other approaches by explicitly separating covariate generation from multiple alternative disease-generating mechanisms. Rather than optimizing a single global fit criterion, the framework makes structural trade-offs transparent and controllable: it allows users to prioritize, depending on the research question, the precise specification of regression coefficients, disease prevalences, marginal distributions, or inter-disease associations. While individual components of this strategy can also be implemented using existing tools, to our knowledge no single framework integrates these alternative disease-generating mechanisms within a unified and systematically evaluated structure that explicitly documents which statistical properties are preserved under each modeling choice.

The remainder of this paper is organized as follows: In [Methods](#) section, we formulate the statistical model of interest and guide the reader through the aforementioned four-step framework. Further, we present strategies for verifying required key properties in the synthetic data. In [Use case: Co-occurring diseases in chronic pain patients](#) section, we turn to the medical use case of co-occurring pain-causing diseases which provided the motivation for the methodological question. For demonstration purposes, we utilize real data as a basis for synthetic data generation, applying a selection of presented approaches. Subsequently, we examine to what extent the intended properties can be verified and present the according results in [Results](#) section. In [Discussion](#) section, we discuss strengths and limitations of the generation approaches and scenarios in which they are most applicable. We conclude our work in [Conclusion](#) section.

## Methods

We aim to provide controlled datasets that represent a medical use case and enable us, through controllable parameters, to perform method validation by means of simulation studies. The particular application of interest concerns the situation where patients suffer from multiple diseases simultaneously, and where a holistic view of patient characteristics and comorbidities can support the diagnostic process. Before addressing this specific application, we first explain the general approach to synthetic data generation.

For this purpose, one distinguishes between three types of synthetic data generation processes, each of which is

suited to different research questions and the prevailing conditions. In all cases, we assume the existence of ground truth, that is, we specify characteristics of individuals and/or populations and require the synthetic data to resemble these properties. The three types are *knowledge-driven*, *data-driven* and *hybrid* data generation [15]. The knowledge-driven approach uses available information from scientific literature, other publication sources and expert knowledge to describe the theoretical basis for data generation. As an example, one may sample random variables from distributions motivated by expert knowledge, or using literature-derived parameters. The data-driven approach, in contrast, is solely based on real data as input for the data generating process. Examples are parametric or nonparametric bootstrap. Eventually, the hybrid approach is a combination of defining rules, models and algorithms and incorporating real data. All three approaches come with strengths and challenges, often connected to the quality of both the domain knowledge and the input data. A detailed discussion is presented by [15].

Apart from this categorization of the generation process, there is also a categorization of the resulting data: According to [29], one distinguishes between *truly synthetic*, *fully synthetic*, *partially synthetic* and *anonymized-only* data. For truly synthetic data, no real data points are considered during the generation process at all. For fully synthetic data, in contrast, information from real data is used to generate new data; however, the real data points will not be part of the generated data. Partially synthetic data combines (truly or fully) synthetic data with unmodified real data. Eventually, anonymized-only data purely consists of real data, however, with sensitive information removed, replaced or encrypted. In our work, we pursue knowledge- or data-driven generation of truly or fully synthetic data.

## Model

We are interested in the medical use case where patients suffer from several diseases and a variety of medical symptoms at the time. This may include, on the one hand, cases where different diseases exhibit similar symptoms and, as a result, a particular disease may be overlooked. On the other hand, it may also concern situations where symptoms are opposing, and the symptoms of one disease lead to the exclusion of another diagnosis.

In terms of statistical modeling, we consider a set of patients  $i = 1, \dots, n$ . Each patient suffers from one, several or none of  $d$  diseases of interest. The presence or absence of diseases is expressed by  $\mathbf{y}_i = (y_{i1}, \dots, y_{id})' \in \{0, 1\}^d$ , where  $y_{ij}$  equals one if patient  $i$  suffers from disease  $j \in \{1, \dots, d\}$ , and zero otherwise. Additionally, there is patient information available, such as core data, symptoms or laboratory

values. This is summarized in a  $k$ -dimensional vector  $\mathbf{x}_i$  for patient  $i$ , comprising values of  $k$  covariates.

While  $\mathbf{y}_i$  denotes the (in practice unknown) ground truth, we use  $\mathbf{Y}_i = (Y_{i1}, \dots, Y_{id})'$  to denote a  $d$ -dimensional random variable with state space  $\{0, 1\}^d$ , which  $\mathbf{y}_i$  is a realization of. Disease probabilities are given by

$$\pi_{ij} = P(Y_{ij} = 1) \quad \text{and} \quad \pi_{ij}(\mathbf{x}_i) = P(Y_{ij} = 1 \mid \mathbf{x}_i),$$

respectively. We also consider disease scores  $s_{ij}$  (see also [Disease probabilities  \$\pi\_i\$  and disease scores  \$s\_i\$](#)  section) which have the meaning that

$$s_{ij} \geq s_{il} \implies y_{ij} \geq y_{il} \quad (1)$$

for diseases  $j, l \in \{1, \dots, d\}$ . Other than probabilities, they determine disease outcomes without further randomness (more details provided in Approaches IIIC, IVc and IVd below). We summarize the probabilities and scores to vectors  $\boldsymbol{\pi}_i = (\pi_{i1}, \dots, \pi_{id})'$  and  $\mathbf{s}_i = (s_{i1}, \dots, s_{id})'$ , respectively.

When we speak of co-occurring diseases, particular attention is given to possible medical causes of such co-occurrences. In general, diseases may or may not have causal effect on each other; but even if they do not, their occurrence may be correlated due to covariates that are associated to both diseases. That is, the covariates may indicate the presence or absence of diseases without influencing their occurrence, or they may directly promote or prevent the occurrence of diseases, or both. In our study, we do not make any explicit assumption about the unconditional (in-)dependence of any two diseases  $j$  and  $l$ , i. e. about the joint distribution of  $Y_{ij}$  and  $Y_{il}$  for diseases  $j, l \in \{1, \dots, d\}$  with  $j \neq l$ .

### Synthetic data generation

We aim to generate synthetic data for the above described model of co-occurring diseases. We set the number of patients to  $n \geq 1$ , the number of considered diseases to  $d \geq 2$  and the number of patient covariates to  $k \geq 1$ . In what follows, we assume all patients (more precisely, the occurrence of their diseases and the values of their covariates) to be mutually independent and use a generic index  $i \in \{1, \dots, n\}$  to refer to patient  $i$ .

To generate medically realistic or statistically interesting data, a modeler should be able to specify certain target parameters. These include:

- (i) the effect of patient covariates on disease occurrence;
- (ii) the dependencies among the several diseases;
- (iii) as well as the prevalence  $p_j$  of individual diseases or disease combinations  $j$ .

It will not always be possible to satisfy every target simultaneously. For example, if two diseases are influenced by covariates in opposite ways, a positive correlation between them may be precluded. Similarly, commonly observed covariate values are less likely to be associated with rare diseases. This section describes algorithms that incorporate mathematical constraints, such as the requirement of positive definiteness for correlation matrices, and demonstrates how data can be obtained that matches the specified targets as closely as possible.

For the data generation, we start by (I) first simulating covariate data  $\mathbf{x}_i$ . Then, conditional on this, we (II) construct functions which mirror the impact of covariates on disease occurrence. These are based on linear combinations of covariate values; we call them (linear) predictors and denote them by  $\boldsymbol{\eta}_i \mid \mathbf{x}_i$ . Next, we (III) derive disease probabilities  $\boldsymbol{\pi}_i \mid \mathbf{x}_i$  or disease scores  $\mathbf{s}_i \mid \mathbf{x}_i$  (usually equivalent to  $\boldsymbol{\pi}_i \mid \boldsymbol{\eta}_i$  and  $\mathbf{s}_i \mid \boldsymbol{\eta}_i$ ). Finally, we (IV) draw realizations  $\mathbf{y}_i \mid \boldsymbol{\pi}_i$  or  $\mathbf{y}_i \mid \mathbf{s}_i$  of disease occurrences. In this last step we assume that the probabilities  $\boldsymbol{\pi}_i$  and scores  $\mathbf{s}_i$  contain sufficient information for the simulation of disease occurrences in the sense that the covariates  $\mathbf{x}_i$  do not contain any additional relevant information anymore. That is, conditional on  $\boldsymbol{\pi}_i$  or  $\mathbf{s}_i$ ,  $\mathbf{y}_i$  is independent of  $\mathbf{x}_i$ . In the following, we describe these Steps I to IV, and within each step, a number of alternative approaches. Figure 1 provides an overview of possible combinations.

### Covariates $\mathbf{x}_i$

Covariates capture patient details such as core data, life-style factors, clinical measurements, laboratory results, or genetic data. They do not yet include diagnoses but may still offer indications of potential diseases. The distribution of individual covariates, as well as the dependencies between them, are considered representative of the population of interest. Distributions should therefore be chosen accordingly, and results interpreted in this context.

#### Ia. Anonymized-only or partially synthetic data

If suitable data is available, and if data quality, data quantity, and data protection do not pose any obstacles, anonymized-only or partially synthetic data can be used as covariates, even if no corresponding diagnoses are known. By relying on such data as covariates, existing physiological and clinical correlations such as the association between systolic and diastolic blood pressure are naturally retained in the synthetic setting, provided that their joint distribution is adequately represented.

#### Ib. NORTA (NORmal To Anything)

In the following, we focus on the more complex case in which we aim for truly synthetic or fully synthetic covariates. In our approach, we create synthetic data by the NORmal To Anything (NORTA) approach [30]. This approach requires the modeler to specify the correlation

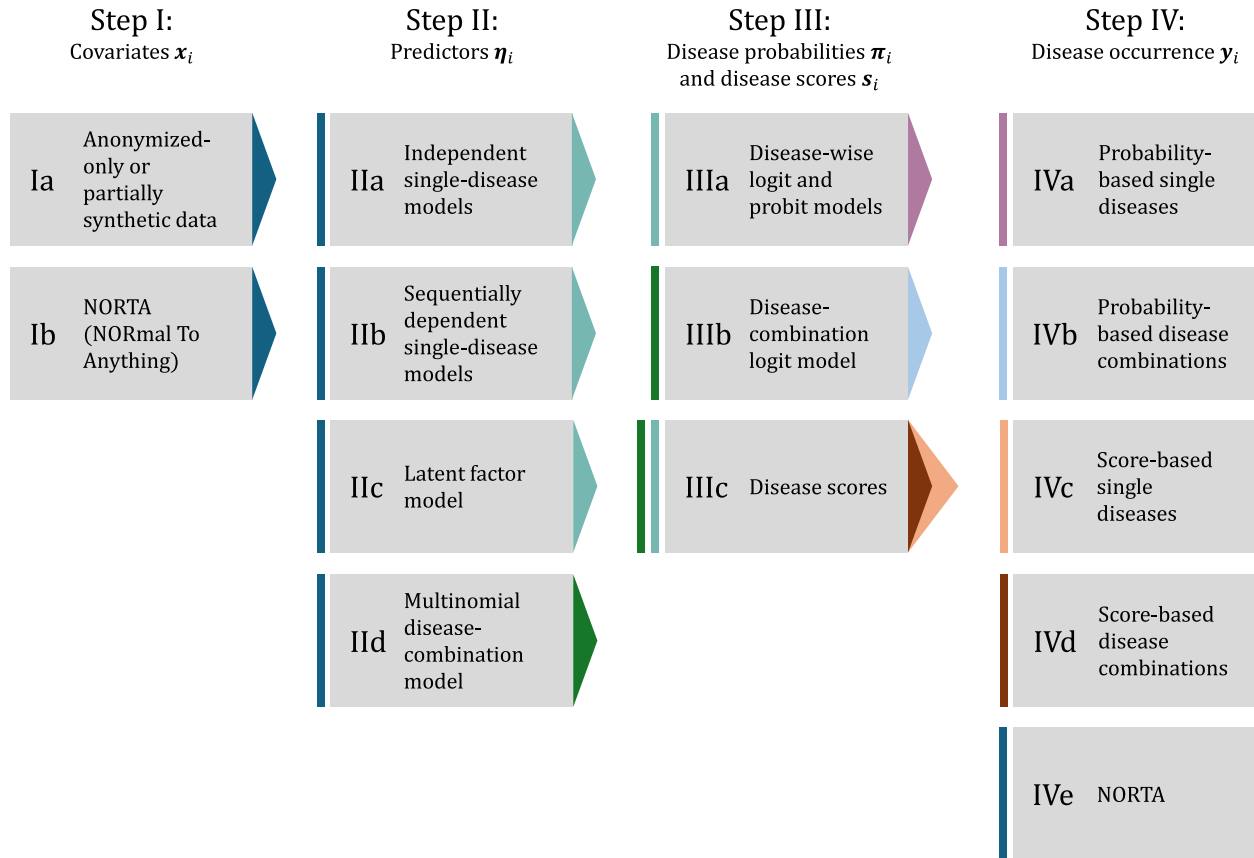


**Fig. 1** Overview of Steps I to IV for synthetic data generation. Each column shows various approaches, representing alternative options that can be combined across steps. The colors of arrows at the right side of each box indicate which approach can follow next, namely one with a bar with matching color at the left. Note that Approach IVe can follow directly after Step I, meaning that Steps II and III are not required here

structure and the marginal distributions of the covariates. The information basis for this specification determines whether the resulting data is regarded truly or fully synthetic. Based on Gaussian copulas, the algorithm leads to the multivariate generation of data with the desired properties, i. e. association structure and distribution. The steps are as follows:

1. Specify the correlation structure of the  $k$  covariates by a  $(k \times k)$ -dimensional matrix  $\mathbf{R}$  that contains all pairwise correlations. The matrix must, of course, satisfy the properties of a correlation matrix; that is, it must be symmetric, positive semi-definite, have values in the range  $[-1, 1]$  and ones on the main diagonal.
2. Describe the marginal distributions of the covariates  $h \in \{1, \dots, k\}$  by defining their cumulative distribution functions  $F_h$ .
3. Incorporate the desired dependence structure between the covariates by drawing a random variable  $z_i$  from a multivariate normal distribution with mean zero and covariance matrix  $\mathbf{R}$  for each

patient  $i$ , i. e.  $z_i = (z_{i1}, \dots, z_{ik})' \sim \mathcal{N}_k(\mathbf{0}, \mathbf{R})$  independently for all  $i = 1, \dots, n$ .

4. Transform the  $z_i$  into (correlated) uniformly distributed variables in the domain  $[0, 1]^k$  by component-wise application of the standard normal cumulative distribution function  $\Phi$ . This way, we generate random variables  $u_i = (u_{i1}, \dots, u_{ik})'$  with  $u_{ih} = \Phi(z_{ih})$  for  $i \in \{1, \dots, n\}$  and  $h \in \{1, \dots, k\}$ . Each of these variables is marginally uniform, while the sought dependence structure between the variables remains.
5. Convert the uniform variables  $u_i$  into the desired target distribution using the quantile functions (inverse cumulative distribution functions)  $F_h^{-1}$  of the respective covariates by setting  $x_i = (x_{i1}, \dots, x_{ik})'$  with  $x_{ih} = F_h^{-1}(u_{ih})$  for all  $i \in \{1, \dots, n\}$  and  $h \in \{1, \dots, k\}$ .

This method explicitly requires the specification of a target correlation structure between the covariates and thereby ensures that relevant dependencies are directly embedded in the data-generating process. In particular,

clinically meaningful associations can be predefined and systematically incorporated into the synthetic data.

### Predictors $\eta_i$

The predictors represent the central mechanism for establishing the association between covariates and disease occurrence, and the relationship between diseases. We present four different approaches that differ in the dependencies they generate and in their complexity, in terms of the parameters that need to be specified: independent single-disease models, sequentially dependent single-disease models, a latent factor model, and a multinomial disease-combination model. Table A1 in the Appendix summarizes required input specifications for these approaches.

Although mathematically not required, we recommend standardizing metric covariates  $h$  such that the empirical mean over  $x_{1h}, \dots, x_{nh}$  equals zero and the empirical variance becomes one, or that they cover identical value ranges. This leads to better comparability of the effect sizes to be specified. In the following, all computations of predictors  $\eta_i$  are conditional on  $x_i$ , even if not flagged explicitly in equations. We employ vectors  $\tilde{x}_i = (1, x'_i)' \in \mathbb{R}^{k+1}$  for the purpose of vector product notation.

#### Ila. Independent single-disease models

We use regression models [31] to describe the probabilities of patients suffering from the considered diseases, conditional on the covariate information. To that end, we start by combining the covariates to linear predictors. In our first approach, we model the diseases to occur independently of each other. In particular, we do not assume that diseases are mutually exclusive; instead, a patient can suffer from any combination of them. In detail, we follow these steps for patients  $i \in \{1, \dots, n\}$  and diseases  $j \in \{1, \dots, d\}$ :

1. Define coefficient vectors  $\beta_j = (\beta_{j0}, \dots, \beta_{jk})'$ , where  $\beta_{jh} \in \mathbb{R}$  describes the association between the  $j$ th disease and  $h$ th covariate, conditional on all others. The desired proportion  $p_j$  of patients affected by disease  $j$  can be regulated by the intercept  $\beta_{j0}$ . In case of normally distributed covariates with mean zero, it should be chosen as

$$\beta_{j0} = \log \left( \frac{p_j}{1 - p_j} \right). \quad (2)$$

2. Calculate a linear predictor  $\eta_{ij} | x_i$  as

$$\eta_{ij} = \tilde{x}_i' \beta_j = \beta_{j0} + \beta_{j1}x_{i1} + \dots + \beta_{jk}x_{ik}.$$

#### Iib. Sequentially dependent single-disease models

Next, we extend Approach Ila by using sequentially dependent rather than independent regression models to describe disease probabilities. This modeling approach is based on the assumption that information about the presence of one disease alters the probability of another disease being present, beyond what is already explained by the information contained in the covariates. To include disease dependencies in the simulation process, we model the disease occurrences sequentially and condition on the already generated disease variables when calculating the linear predictor for subsequent diseases. To this end, we follow these steps for every patient  $i \in \{1, \dots, n\}$ :

1. Define coefficient vectors  $\beta_j = (\beta_{j0}, \dots, \beta_{jk})'$  as in Step 1 of Approach Ila. Note, however, that the coefficients will have a different impact here due to the conditioning on other diseases (inclusion of coefficients  $\gamma_{jl}$ ) in the following.
2. Sort the diseases to obtain an order such that, for  $j < l$ , we can model the probability of occurrence of disease  $j$  independently of disease  $l$ . In particular, the occurrence probability of disease 1 will be modeled independently from all other diseases.
3. Determine coefficients  $\gamma_{jl}$  which describe how the occurrence probability of disease  $l$  is influenced by the occurrence of disease  $j \in \{1, \dots, l-1\}$ , conditional on  $x_i$  and the occurrence information of the diseases  $1, \dots, j-1, j+1, \dots, l-1$ .
4. For  $l = 1$ , do the following:

– Calculate

$$\eta_{i1} = \tilde{x}_i' \beta_1.$$

– Proceed with Steps III and IV to simulate  $y_{i1}$ .

5. For  $l = 2, \dots, d$ , do the following:

– Calculate

$$\eta_{il} = \tilde{x}_i' \beta_l + \sum_{j=1}^{l-1} \gamma_{jl} y_{ij}.$$

– Proceed with Steps III and IV to simulate  $y_{il}$ .

**Iic. Latent factor model** So far, we have computed the predictors conditional on observed covariates and (in Approach Iib) information about the presence of other diseases. In a further approach, we now include additional factors that are neither observed nor assigned a substantive interpretation. These are latent factors that influence

the joint occurrence of each pair of diseases. Using shared latent factors makes it possible to simulate disease occurrences which depend on both known patient covariates and unknown other influences [32]. We assume the latent factors to be independent of the observed covariates: While correlation between them is possible in real-world data, maintaining strict independence isolates the unobserved dependence between diseases from the effects of the covariates, preventing a confounded structure and making the synthetic data more useful for method validation. We proceed as follows:

1. Define coefficient vectors  $\beta_j = (\beta_{j0}, \dots, \beta_{jk})'$  as in Step 1 of Approach IIa. Note, however, that the coefficients will have a different impact here due to the conditioning on latent factors (inclusion of coefficients  $\lambda_{jl}$ ) in the following.
2. For every disease  $j$ , determine a set  $B_j$  of diseases which share a latent factor with  $j$ , i. e.  $B_j \in \{1, \dots, d\} \setminus \{j\}$ . By definition, shared latent factors influence two diseases; thus, the sets must fulfill

$$j \in B_l \iff l \in B_j.$$

It is allowed that  $B_j$  is empty. To avoid using vectors of varying dimensions in the following, we introduce the (symmetric and potentially sparse) matrix  $B^*$  with entries  $b_{jl} = 1$  if  $j \in B_l$ , i. e. if diseases  $j$  and  $l$  share a latent factor, and zero otherwise.

3. Define coefficients  $\lambda_{jl} \in \mathbb{R}$  which indicate the influence of the latent factor shared by diseases  $j$  and  $l$  on disease  $j$ . Although the latent factor is shared, the influence on diseases  $j$  and  $l$  may differ, i. e. we do not require  $\lambda_{jl} = \lambda_{lj}$ . For  $b_{jl} = 0$ , we set  $\lambda_{jl} = 0$ .
4. For each patient  $i$ , simulate independent standard normally distributed latent factors  $w_{ijl} \sim N(0, 1)$  if  $b_{jl} = 1$  and zero otherwise. Since latent factors are shared, we require  $w_{ijl} = w_{ilj}$  for all  $j, l \in \{1, \dots, d\}$ .
5. For each patient  $i$  and disease  $j$ , calculate a linear predictor which takes both the influence of covariates  $\mathbf{x}_i$  as well as latent factors  $w_{ij1}, \dots, w_{ijd}$  into account:

$$\eta_{ij} = \tilde{\mathbf{x}}_i' \beta_j + \sum_{l=1}^d \lambda_{jl} w_{ijl}.$$

## II.d. Multinomial disease-combination model

So far, we have considered the  $d$  diseases side by side and, for each patient, examined a  $d$ -dimensional combination  $(y_{i1}, \dots, y_{id})' \in \{0, 1\}^d$  of occurrences and non-occurrences. An alternative approach is to describe each combination as one of  $2^d$  possible (unordered) classes of aggregated disease profiles. This is what we do in the following. To that end, we introduce variables  $c_i$  with  $c_i = r$  indicating that patient  $i$  suffers from disease combination  $r$ , and  $r \in \{1, \dots, 2^d\}$ . In particular, we number the classes such that combination  $(j_1, \dots, j_d)' \in \{0, 1\}^d$  corresponds to class  $1 + \sum_{m=1}^d j_m 2^{m-1}$ . For example, in the case of three possible diseases, the combination  $(0, 0, 0)'$  corresponds to class 1,  $(0, 1, 0)'$  to class 3, and  $(1, 1, 1)'$  to class 8. For any  $d$ , the first class refers to the combination where none of the single diseases occurs, and  $r = 2^d$  to the case where all of them do. We undertake the following steps:

For every disease combination  $r$ , we calculate linear predictors after defining regression coefficients as intercepts and the effects of the  $k$  covariates on the disease outcome. For patient  $i$ , these linear predictors  $\eta_{ir}$  represent the log-odds of class  $r$  relative to the reference category. Since  $r = 1$  serves as reference category and its occurrence probability will later (in Step III) be computed as the complement to all other class probabilities, we do not specify coefficients for this class.

We propose three ways to define the regression coefficients and linear predictors:

- (i) If knowledge is available about occurrences of disease combinations, and how they are influenced by the covariates, the regression coefficients  $\beta_r = (\beta_{r0}, \beta_{r1}, \dots, \beta_{rk})'$  can be chosen individually for every disease combination  $r$ . This procedure requires the user to define  $(2^d - 1)(k + 1)$  coefficients  $\beta_{rh}$  ( $r \in \{2, \dots, 2^d\}$ ,  $h \in \{0, \dots, k\}$ ). In case of normally distributed covariates with mean zero, the intercept  $\beta_{r0}$  should be chosen as

$$\beta_{r0} = \log \left( \frac{p_r}{p_1} \right), \quad (3)$$

where  $p_r$  is the desired proportion of patients affected by disease combination  $r$ . This regulates the proportion of patients suffering from  $r \geq 2$  compared to the proportion of healthy patients ( $r = 1$ ). For every patient  $i$ , calculate the linear predictor for disease combination  $r \geq 2$  as

$$\eta_{ir} = \tilde{\mathbf{x}}_i' \boldsymbol{\beta}_r = \beta_{r0} + \beta_{r1}x_{i1} + \cdots + \beta_{rk}x_{ik}.$$

- (ii) Alternatively, we can model the impact of covariates on disease occurrence for every disease individually. That means, the regression coefficients are independently defined as  $\boldsymbol{\beta}_j = (\beta_{j0}, \beta_{j1}, \dots, \beta_{jk})'$  for every disease  $j \in \{1, \dots, d\}$ . This requires the user to specify only  $d(k+1)$  coefficients  $\beta_{jh}$  ( $j \in \{1, \dots, d\}, h \in \{0, \dots, k\}$ ). For every patient  $i$ , one calculates a linear predictor for disease  $j$  as

$$\tilde{\eta}_{ij} = \tilde{\mathbf{x}}_i' \boldsymbol{\beta}_j = \beta_{j0} + \beta_{j1}x_{i1} + \cdots + \beta_{jk}x_{ik}.$$

The linear predictor for disease combination  $r$ , corresponding to  $(j_1, \dots, j_d)'$ , then results as

$$\eta_{ir} = \prod_{m=1}^d \tilde{\eta}_{im}^{j_m}. \quad (4)$$

Since all exponents equal either zero or one, that is the product of all linear predictors corresponding to the single diseases in combination  $r$ . While the choice of the coefficients  $\beta_{jh}$  can be made exactly analogously to Approach IIa, the difference arises at the point where the predictors  $\eta_{ir}$  are calculated per disease combination rather than per disease. This changes the dependency structure among the generated disease outcomes, also in interaction with the disease prevalences.

- (iii) Via the product (4), the previous approach mingles individual linear predictors and thus eventually leads to synthetic disease occurrences which are dependent both among each other and on the covariates. However, it does not retain the individual relationship between single diseases and covariates. This can be achieved if the linear predictor for a disease combination is calculated dependent on the linear predictors of a subset of the disease combinations which reflect mutually exclusive single diseases (e. g.  $(1, 0)'$  and  $(0, 1)'$  for  $d = 2$ ). To this end, we make use of the odds ratio  $\text{OR}_{ir}$ , which quantifies for patient  $i$  the association between the single diseases contained in combination  $r$ , and therefore how the presence or absence of one single disease changes the odds of observing a particular disease category. With the detailed calculations in Section A.1 in the Appendix, we show the following:

Let  $r = (j_1, \dots, j_d)'$  be some disease combination with  $j_1 = 1$  and  $j_2 = 1$  (and arbitrary values for  $j_3, \dots, j_d$ ). Let  $r'$  be identical to  $r$  with the difference that  $j_1 = 0$ , and  $r''$  like  $r$  with  $j_2 = 0$ . Then, the linear predictor for patient  $i$  and disease combination  $r$  can be calculated as:

$$\eta_{ir} = \log(\text{OR}_{ir}) + \eta_{ir'} + \eta_{ir''}.$$

We can set  $\text{OR}_{ir}$  according to the desired association between the diseases 1 and 2. It then holds that

$$\text{OR}_{ir} \begin{cases} > 1 & \text{in case of positive association,} \\ = 1 & \text{in case of no association,} \\ < 1 & \text{in case of negative association.} \end{cases}$$

With this procedure, we constrain the linear predictor for the joint disease outcome instead of building independent linear predictors. This allows us to explicitly model an association between the single diseases. In Section A.1 in the Appendix, we also present the calculations for a combination of three diseases. The derivation for more than three diseases is analogous; however, combining more diseases decreases their practical relevance.

### Disease probabilities $\pi_i$ and disease scores $s_i$

After having constructed predictors that link the values of covariates with the occurrence of single or combined diseases, we now aim to derive probabilities  $\pi_i$  or scores  $s_i$  for disease occurrences based on them. In the following, we use the common notation to condition on the covariates  $\mathbf{x}_i$  rather than on the predictors  $\eta_i$ , because the relation between  $\mathbf{x}_i$  and  $\eta_i$  is often deterministic.

Probabilities have the advantage of being clearly interpretable, with a known range and an intuitive sense of how common or rare events are, and they directly represent the uncertainty associated with random events. Depending on the context, however, one may instead be interested in scores, which preserve the structure of the (linear) predictors and allow for easier control of the simulated disease occurrences. In the following, we present methods for deriving both: Approaches IIIa and IIIb for probabilities, and IIIc for scores.

#### IIIa. Disease-wise logit and probit models

Approaches IIa, IIb and IIc delivered predictors  $\eta_{ij}$  for patients  $i \in \{1, \dots, n\}$  for every single disease  $j \in \{1, \dots, d\}$ . Dependencies between diseases are integrated in the predictors already. A straightforward procedure from regression modeling is to use the logistic function to translate the predictors into occurrence probabilities

$$\pi_{ij}(\mathbf{x}_i) = P(y_{ij} = 1 \mid \mathbf{x}_i) = \frac{1}{1 + e^{-\eta_{ij}}}. \quad (5)$$

Alternatively, corresponding to the well-known probit model, one may calculate

$$\pi_{ij}(\mathbf{x}_i) = P(y_{ij} = 1 \mid \mathbf{x}_i) = \Phi(\eta_{ij}) \quad (6)$$

with the cumulative distribution function  $\Phi$  of the standard normal distribution. While the logit model (5) is generally more straightforward to interpret, the probit model (6) comes with properties such as thinner tails.

### IIIb. Disease-combination logit model

Approach IId considered disease combinations  $(y_{i1}, \dots, y_{id})' \in \{0, 1\}^d$ , numbered by  $r \in \{1, \dots, 2^d\}$ . Accordingly, we derive multinomial probability vectors  $\pi_i = (\pi_{i1}, \dots, \pi_{i2^d})$ . These differ from the probabilities in Approach IIIa in that the components are required to sum up to one. Based on  $\eta_{ir}$ , we calculate the occurrence probability for disease combination  $r \geq 2$  as

$$\pi_{ir}(\mathbf{x}_i) = P(c_i = r \mid \mathbf{x}_i) = \frac{e^{\eta_{ir}}}{1 + \sum_{m=2}^{2^d} e^{\eta_{im}}}$$

where  $c_i$  stands for the disease combination class of patient  $i$  as introduced in Approach IId. The occurrence probability for the reference category  $r = 1$  can be calculated as  $\pi_{i1}(\mathbf{x}_i) = 1 - \sum_{r=2}^{2^d} \pi_{ir}$  [31].

### IIIc. Disease scores

Unlike probabilities, disease scores determine disease outcomes directly, without introducing further uncertainty, through Formula (1) for diseases or disease combinations. In the simplest case, one may leave the predictors from Step II unaltered and set  $s_{ij} = \eta_{ij}$  for all  $i$  and  $j$ . Alternatively, one may add independent and identically distributed noise  $\varepsilon_{ij} \sim N(0, \sigma^2)$  for some  $\sigma > 0$ , such that  $s_{ij} = \eta_{ij} + \varepsilon_{ij}$ . The individual- and disease-specific noise term represents the fact that, even if two individuals agree in their covariates and potentially latent factors, the disease outcome may differ.

### Disease occurrence $y_i$

Finally, we proceed from the continuous-valued probabilities and scores to discrete-valued realizations of whether a disease or disease combination does or does not occur. While the probability-based Approaches IVa and IVb represent the uncertainty associated with such transitions more realistically, the score-based Approaches IVc and IVd offer greater controllability, particularly with respect to desired disease prevalences. Approach IVe revisits the NORTA algorithm and generates disease occurrences based on correlations.

### IVa. Probability-based single diseases

Given we have calculated occurrence probabilities  $\pi_{ij}$  for individual diseases  $j$  (via Approaches IIIa or IIId), we now draw independent Bernoulli distributed random variables:

$$y_{ij} \mid \mathbf{x}_i \sim \text{Bernoulli}(\pi_{ij}(\mathbf{x}_i))$$

for all patients  $i \in \{1, \dots, n\}$  and diseases  $j \in \{1, \dots, d\}$ .

In Approach IIa, we indicated a choice of intercept in Eq. (2) which (in theory) regulates the proportion of patients suffering from disease  $j$  such that it matches the desired fraction  $p_j$ . In case of non-normally distributed or non-centered covariates, however, the target proportion may be missed. In this case, an iterative correction can be used where one adjusts the intercept  $\beta_{j0}$  iteratively to

$$\beta_{j0}^* = \beta_{j0} + \log\left(\frac{p_j}{1 - p_j}\right) - \log\left(\frac{\bar{y}_j}{1 - \bar{y}_j}\right) \quad (7)$$

until the difference between the target prevalence  $p_j$  and sample proportion  $\bar{y}_j$  falls below a chosen tolerance value (e. g. 0.001) [33].

### IVb. Probability-based disease combinations

If we obtained occurrence probabilities  $\pi_{ir}$  for disease combinations  $r \in \{1, \dots, 2^d\}$  rather than individual diseases (via Approach IIIb), we proceed by drawing independent multinomial distributed random variables:

$$c_i \mid \mathbf{x}_i \sim \text{Multinomial}(\pi_{i1}(\mathbf{x}_i), \dots, \pi_{i2^d}(\mathbf{x}_i))$$

for all patients  $i \in \{1, \dots, n\}$ , where  $c_i$  denotes the disease combination for patient  $i$ . Before, we numbered the classes such that combination  $(j_1, \dots, j_d)'$  corresponded to class  $1 + \sum_{m=1}^d j_m 2^{m-1}$ . We can thus determine the individual disease occurrences for patient  $i$  via

$$y_{im} = \begin{cases} 1 & \text{if } j_m = 1 \text{ in combination } c_i, \\ 0 & \text{otherwise} \end{cases} \quad (8)$$

for  $m \in \{1, \dots, d\}$ .

In analogy to the intercept correction in Eq. (7) from Approach IVa, the intercept  $\beta_{r0}$  for disease combination  $r \geq 2$  can be iteratively adjusted by

$$\beta_{r0}^* = \beta_{r0} + \log\left(\frac{p_r}{p_1}\right) - \log\left(\frac{\bar{y}_r}{\bar{y}_1}\right). \quad (9)$$

### IVc. Score-based single diseases

Based on disease scores  $s_{ij}$  for single diseases  $j \in \{1, \dots, d\}$ , one can determine binary disease outcomes by comparing the scores with thresholds  $\tau_j$  which are chosen by the modeler for every disease  $j$ :

$$y_{ij} = \begin{cases} 1 & s_{ij} > \tau_j, \\ 0 & s_{ij} \leq \tau_j. \end{cases}$$

Adjusting the threshold  $\tau_j$  enables one to reach the desired prevalence  $p_j$  of disease  $j$ . In particular, one can chose  $\tau_j$  to be the empirical  $p_j$ -quantile across  $s_{1j}, \dots, s_{nj}$ .

#### IVd. Score-based disease combinations

If scores  $s_{ir}$  for disease combinations  $r \in \{1, \dots, 2^d\}$  are provided, a combination can be selected as

$$c_i = \arg \max_{r \in \{1, \dots, 2^d\}} s_{ir}$$

for every patient  $i \in \{1, \dots, n\}$ . If the score was obtained by adding normally distributed noise in Approach IIIc, this procedure corresponds to a multivariate probit model. Via Eq. (8), one can obtain single-disease realizations from the combinations.

#### IVe. NORTA

In this approach, we utilize the NORTA procedure which has already been described in Approach Ib in the context of generating synthetic covariate data. It directly builds on the covariate generation (Step I) without taking the model-based path through Steps II and III: It simulates disease occurrence solely based on specified marginal distributions and correlations. With the following algorithm, we extend the framework of [30]. In particular, we describe how occurrences  $y_{ij}$  are generated conditional on covariates  $x_i$ . We provide this description according to the chosen structuring into Steps I through IV. However, if NORTA was already applied in Step I (i. e. Approach Ib), these steps could be merged for easier implementation, without the need to consider the subsequent formulas for conditional probabilities. If covariate data was obtained via Approach Ia, or if one seeks to separate the levels, one should follow these steps:

1. Determine the correlation structure of the  $k$  covariates. This is either known, if the covariates were synthetically generated by Approach Ib, or it can be estimated via the empirical correlation matrix of the  $x_i$ . In any case, denote the resulting  $(k \times k)$ -dimensional matrix by  $R_{xx}$ .
2. Specify the correlation structure of the  $d$  diseases (i. e.  $(y_{i1}, \dots, y_{id})'$  across all patients  $i$ ) by a  $(d \times d)$ -dimensional matrix  $R_{yy}$  that contains all pairwise correlations. Further, specify the pairwise correlations between  $(y_{i1}, \dots, y_{id})'$  and  $(x_{i1}, \dots, x_{ik})'$  across all  $i$  in a  $(d \times k)$ -dimensional matrix  $R_{yx}$ .
3. Describe the marginal distributions of the occurrences  $y_{ij}$  by defining their cumulative distribution functions  $F_{y,j}$  for  $j \in \{1, \dots, d\}$ .

4. Generate auxiliary random vectors  $z_{x,i}$  which correspond to the vectors  $z_i$  in Step 3 of Approach Ib. If these vectors are still known from a previous NORTA application, they can be used. Otherwise, conduct the following calculations (reversal of Steps 4 and 5 from Approach Ib) based on the available covariate data:
  - Compute the empirical cumulative distribution functions  $F_{x,h}$  for all covariates  $h$ .
  - Calculate  $u_{x,ih} = F_{x,h}(x_{ih})$  and  $z_{x,ih} = \Phi^{-1}(u_{x,ih})$  for all  $i$  and  $h$ , where  $\Phi$  denotes the standard normal cumulative distribution function. Let  $z_{x,i} = (z_{x,i1}, \dots, z_{x,ik})'$ .

- Compute the empirical cumulative distribution functions  $F_{x,h}$  for all covariates  $h$ .
  - Calculate  $u_{x,ih} = F_{x,h}(x_{ih})$  and  $z_{x,ih} = \Phi^{-1}(u_{x,ih})$  for all  $i$  and  $h$ , where  $\Phi$  denotes the standard normal cumulative distribution function. Let  $z_{x,i} = (z_{x,i1}, \dots, z_{x,ik})'$ .
5. Incorporate the desired dependence structure among the diseases and between the diseases and covariates by drawing random variables  $z_{y,i} = (z_{y,i1}, \dots, z_{y,id})'$  from a multivariate normal distribution with mean

$$R_{yx} R_{xx}^{-1} z_{x,i}$$

and covariance matrix

$$R_{yy} - R_{yx} R_{xx}^{-1} R'_{yx}$$

independently for all patients  $i \in \{1, \dots, n\}$ .

6. Transform the  $z_{y,i}$  into (correlated) uniformly distributed variables in the domain  $[0, 1]^d$  by component-wise application of  $\Phi$ . This way, we generate variables  $u_{y,i} = (u_{y,i1}, \dots, u_{y,id})'$  with  $u_{y,ij} = \Phi(z_{y,ij})$  for  $i \in \{1, \dots, n\}$  and  $j \in \{1, \dots, d\}$ . Each of these variables is marginally uniform, while the sought dependence structure between the variables remains.
7. Convert the uniform variables  $u_{y,i}$  into the desired target distribution using the quantile functions (inverse cumulative distribution functions)  $F_{y,j}^{-1}$  of the respective occurrences by setting  $y_{ij} = F_{y,j}^{-1}(u_{y,ij})$  for all  $i \in \{1, \dots, n\}$  and  $j \in \{1, \dots, d\}$ .

The NORTA approach can also be used to generate occurrences of disease combinations rather than single diseases.

#### Verification of desired properties

In the beginning of [Synthetic data generation](#) section, we outlined three target properties which we aim to fulfill with the generated data: (i) covariate effects on disease

occurrence, (ii) associations between diseases, and (iii) disease prevalences. Once the synthetic data has been generated using the above described approaches, we seek to assess to what extent it reflects the desired ground truth. We proceed as follows:

- (i) **Covariate effects:** The connection between patient covariates and disease occurrence was established in Step II through the construction of predictors. In the approaches presented, regression coefficients were defined to combine the covariates into linear combinations. Due to the superposition of covariate effects with inter-disease dependencies, it can, however, be assumed that the original relationship has become biased. We estimate the coefficients of regression models to verify the consistency between the input parameters and the reverse-engineered estimates. To that end, we look at different paths to generate data, and apply five different ways to re-estimate the input parameters: independent binomial logit models, dependent binomial logit models, a probit model, a multinomial logit model, and the computation of empirical correlations.
- (ii) **Association between diseases:** To examine the relationship between the generated disease occurrences, we use Cramer's  $V$  [34], which is a measure of the association strength between two nominal variables. The measure ranges between zero and one, where  $V = 0$  means no association and  $V = 1$  means perfect association between the according variables. Another meaningful measure, though not applied here for verification purposes, is the tetrachoric correlation, which measures the correlation between two binary variables for which one assumes that they arose from an underlying normally distributed latent variable [32].
- (iii) **Target prevalence:** To assess if the desired disease prevalences are reached, we compare for each disease the target proportion with the arithmetic mean of the according generated disease occurrence.

**Table 1** Overview of the five methodological paths that are chosen to generate synthetic data in the presented use case. For each Path A to E, the table lists the sequence of chosen approaches in Steps I to IV

| Path | Step I | Step II | Step III | Step IV |
|------|--------|---------|----------|---------|
| A    | Ib     | Ila     | IIla     | IVa     |
| B    | Ib     | IIb     | IIla     | IVa     |
| C    | Ib     | IIc     | IIla     | IVa     |
| D    | Ib     | IIId    | IIlb     | IVb     |
| E    | Ib     | –       | –        | IVe     |

### Use case: Co-occurring diseases in chronic pain patients

To test and demonstrate the practical application of the four-step framework in a real-world context, we create and evaluate synthetic datasets for the medical use case which motivated this study.

### Medical context and data

We consider the situation of chronic pain patients who often suffer from multiple co-occurring diseases. In [20], we analyze data from a German rheumatism outpatient clinic, the Department of Internal Medicine and Rheumatology at Hospital Bielefeld Rosenhöhe. All patients in this dataset are suspected to suffer from none, one or several of the following diseases: immune-mediated rheumatic diseases (IMID), fibromyalgia syndrome (FMS), osteoarthritis (OA) and other pain-causing diseases (OPCD). The aim of [20] was to investigate whether the data-driven diagnosis of IMID can be improved by the simultaneous examination of FMS, OA and OPCD.

As a result of this project, methodological questions emerged that could be appropriately investigated on well-suited data. However, although a basis of real data is available in this use case, it cannot be employed for rigorous method validation due to the lack of controllability of ground truth, apart from limitations such as small sample size, measurement uncertainty and missing values. For that reason, we use the real dataset to extract a knowledge base which serves to specify distributions, ranges and interdependencies. To revisit the categorization introduced at the beginning of [Methods](#) section, we focus on data-driven, fully synthetic data generation.

### Procedure

For a focused demonstration of the data generation, we will subsequently concentrate on two of the aforementioned diseases, namely IMID and FMS. Even if we do not aim for medical correctness in this demonstration example, we seek to define realistic value ranges for the variables. Therefore, we use the real patient data for knowledge discovery (data-driven), while no real data is included in the final generated dataset (fully synthetic).

In the following, we describe the relevant design choices in applying the algorithms from [Methods](#) section. All details, including input and target values, can be found in the corresponding R scripts at [https://github.com/fuchslab/synthData\\_diseases](https://github.com/fuchslab/synthData_diseases).

### General settings and target properties

We take five methodological paths to generate synthetic data, illustrated in [Table 1](#). With each path, we generate 1,000 independent and identically distributed synthetic datasets. In each of these datasets, we simulate data

for  $n = 10,000$  patients. For each of the patients, in turn, we simulate the occurrence of  $d = 2$  diseases as binary outcomes, namely of IMID and FMS. For each patient, we assume information from  $k = 6$  covariates to be available.

Regarding the target properties (covariate effects, disease dependencies, disease prevalences), we base the following values on characteristics from the real dataset:

- (i) Desired covariate effects are described further below.
- (ii) As an association between the two diseases IMID and FMS, we target at  $V_{\text{IMID}, \text{FMS}} = 0.42$ .

**Table 2** Description and specification of patient variables that are generated in Step I according to distributions and parameters from real data. The simulated variables serve as covariates in Paths A to E.  $\mathcal{TN}(\mu, \sigma, a, b)$  denotes a truncated normal distribution with mean  $\mu \in [a, b]$ , standard deviation  $\sigma > 0$  and values being constrained to the range  $[a, b]$ .  $\Gamma(\text{shape}, \text{rate})$  denotes the gamma distribution with positive shape and rate parameters

| Variable | Description                                                                   | Type                   | Distribution/Parameters                                                                                                                                                                                             |
|----------|-------------------------------------------------------------------------------|------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Age      | Age of patient at admission<br>(in years)                                     | Metric<br>(continuous) | $\mathcal{TN}(\mu = 55, \sigma = 15.5, a = 18, b = 89)$                                                                                                                                                             |
| Sex      | Sex at birth                                                                  | Binary                 | Bernoulli( $p = 0.31$ );<br>1 = male, 0 = female                                                                                                                                                                    |
| CRP      | C-reactive protein (in mg/l)                                                  | Metric<br>(continuous) | $\Gamma(\text{shape} = 0.5, \text{rate} = 0.05)$                                                                                                                                                                    |
| FiRST    | Sum of positively answered questions of the Fibromyalgia Rapid Screening Tool | Metric<br>(discrete)   | Sample from $\{0, 1, \dots, 6\}$ with probabilities<br>0.03, 0.10, 0.13, 0.17, 0.19, 0.18, 0.20                                                                                                                     |
| NRS      | Numeric rating scale to quantify the pain level                               | Metric<br>(discrete)   | Sample from $\{0, 0.5, 1, \dots, 9.5, 10\}$ with probabilities<br>0.031, 0.002, 0.005, 0.002, 0.020, 0.002, 0.046, 0.009, 0.046, 0.005, 0.147, 0.016, 0.153, 0.024, 0.121, 0.021, 0.178, 0.019, 0.093, 0.004, 0.056 |
| Tender   | Number of pain-sensitive pressure points on body                              | Metric<br>(discrete)   | Sample from $\{0, 1, \dots, 18\}$ with probabilities<br>0.268, 0.016, 0.156, 0.021, 0.093, 0.015, 0.077, 0.005, 0.051, 0.006, 0.053, 0.011, 0.036, 0.010, 0.029, 0.015, 0.058, 0.006, 0.074                         |

- (iii) Regarding marginal prevalences for disease occurrences, we aim to achieve a proportion of  $p_{\text{IMID}} = 0.52$  for IMID and  $p_{\text{FMS}} = 0.21$  for FMS. Additionally, we define the following target prevalences for the four possible disease combinations:  $p_{\text{onlyIMID}} = 0.49$ ,  $p_{\text{onlyFMS}} = 0.19$ ,  $p_{\text{IMID\&FMS}} = 0.03$  and  $p_{\text{healthy}} = 0.29$ . We deliberately choose  $p_{\text{IMID\&FMS}} \neq p_{\text{onlyIMID}} \cdot p_{\text{onlyFMS}}$ , where equality would indicate marginal independence of the occurrence of the two diseases.

### Step I: Covariates

We generate data for six patient-specific covariates: age, sex, C-reactive protein (CRP), sum of positively answered questions of the Fibromyalgia Rapid Screening Tool (FiRST), numeric rating scale (NRS) to quantify the pain level, and the number of pain-sensitive pressure points on the body (tender points). In the following, we refer to these variables as Age, Sex, CRP, FiRST, NRS and Tender. We empirically derive the distributions and their corresponding parameters from the real hospital data as listed in Table 2.

We generate the covariates using Approach Ib, that is the NORTA procedure. To that end, we calculate the empirical  $(6 \times 6)$ -dimensional correlation matrix of the covariates (results shown in Table A2 in the Appendix) and pass it to the algorithm as input matrix  $\mathbf{R}$ . In particular, we proceed as follows: For pairs of variables that are both continuous in the real data, we use Pearson's correlation coefficient [35]; for pairs containing one continuous and one binary variable, we calculate the polyserial correlation [36, 37]. As marginal distributions, we use the ones listed in Table 2, which are motivated through empirical analysis of the real data. For all five paths listed in Table 1, we use the exact same covariate data. We standardize the metric covariates before proceeding with Step II.

### Step II: Predictors

We apply the four model-based Approaches IIa to II d to generate predictors, entering Paths A to D. Path E does not require the generation of a predictor since the NORTA technique in Approach IVe directly builds on Step I without visiting Steps II and III. More precisely, the covariates and occurrences are generated in a joint NORTA application, and then separated for use in Paths A to D. Approaches IIa to II d require specifying coefficients that describe the effect of covariates on disease occurrences. To obtain realistic values for these coefficients, we apply regression models to the real data

and use the estimated coefficients as input values for data generation. Specific choices are as follows:

- As a basis for Approaches IIa to IIc, we estimate two logistic regression models on the real data after standardizing metric covariates — one with IMID as target variable, one with FMS. This results in coefficients  $\beta_{\text{IMID}}$  and  $\beta_{\text{FMS}}$  that are shown in Table A3 in the Appendix.
- In Approach IIb, we choose the order of diseases such that FMS occurrence is simulated first, and IMID occurrence is generated conditional on the occurrence of FMS. For the latter, we employ the coefficient value  $\gamma_{\text{FMS, IMID}} = 4.6$  to model the impact of the occurrence of FMS on IMID.
- For Approach IIc, we create latent factors  $w_{i,\text{IMID, FMS}}$  for every patient  $i$ , which represent a shared influence on the occurrence of both IMID and FMS. This latent factor enters the linear predictor with weights  $\lambda_{\text{IMID, FMS}} = \lambda_{\text{FMS, IMID}} = 5.7$ .
- As a basis for Approach IID, we estimate a multinomial regression model on the real data after standardizing the metric covariates, where the target variable can fall into one of the four categories ‘onlyIMID’ (patients who are exclusively diagnosed with IMID), ‘onlyFMS’ (patients who are exclusively diagnosed with FMS), ‘IMID&FMS’ (patients who are diagnosed with both IMID and FMS) and ‘healthy’ (the reference category). This yields regression coefficients  $\beta_{\text{onlyIMID}}$ ,  $\beta_{\text{onlyFMS}}$  and  $\beta_{\text{IMID\&FMS}}$  that are shown in Table A3 in the Appendix. Within the first step of Approach IID, we choose Option (i) to build linear predictors; that is, we define regression coefficients for each of the disease combination classes (except for the reference category).

### Step III: Probabilities and scores

To reduce the variability between datasets across paths to a few sources, we generate only probabilities in Step III and refrain from producing scores. That is, we apply Approach IIIa for the single-disease Paths A to C and Approach IIIb for the disease-combination Path D.

### Step IV: Disease occurrence

As a logical consequence of generating probabilities in Step III along Paths A through D, these paths now lead to probability-based realizations of disease occurrences: the application of Approach IVa in the single-disease cases (Paths A to C) and of Approach IVb in the disease-combination case (Path D).

In Approaches IIa to IID, we had chosen the intercept as estimated from the real data. Here, in Approaches IVa and IVb, we apply the iterative intercept optimization,

described through Eqs. (7) and (9), respectively, in order to achieve the target proportions of diseased patients.

On Path E, we move directly from the covariate data generation in Step I to the simulation of disease occurrences via NORTA in Approach IVe. To that end, we specify all pairwise correlations between the diseases IMID and FMS, and between the two diseases and the covariates. These are empirically derived from the real data and listed in Table A2 in the Appendix. As before, for pairs containing one continuous and one binary variable, we calculate the polyserial correlation [36, 37]. For pairs of binary variables, we calculate the tetrachoric correlation (see [Verification of desired properties](#) section). In terms of the description of Approach IVe in [Disease occurrence  \$y\_i\$](#)  section, they form the matrices  $R_{yy}$  and  $R_{yx}$ . In practice, however, we generate the covariates and occurrences in one joint NORTA application and separate them afterwards. The marginal distributions of disease occurrence are binomial with success probabilities  $p_{\text{IMID}}$  and  $p_{\text{FMS}}$  as specified before.

### Results

The data generation process involved several steps in which different model layers overlapped and interacted. We now examine to what extent the input parameters can be reproduced from the obtained datasets, and, more importantly, to what extent the synthetic data matches the main target properties (covariate effects, associations between diseases, target prevalences). To that end, we employ the measures described in [Verification of desired properties](#) section (estimated regression coefficients, associations in terms of correlation coefficients and Cramer’s V, and empirical disease prevalence). By comparing the results from Paths A to E, we can capture the influence of, in particular, Approaches IIa through IID, as well as the difference between the model-based Paths A to D and the correlation-based Path E.

For all data generation paths and each characteristic of interest (regression coefficients, correlations), we report the deviation of the estimated value (which we compute from the synthetic data) from the input value (that was provided as input to the data generation process and is known to the modeler). In particular, we look at the average of relative differences across all 1,000 generated datasets per path; the relative difference is the difference between the estimated value and the input value, divided by the input value. Further, we display 95% empirical ranges of the relative differences. Depending on the context, however, the value of the regression coefficient or the correlation resulting from the synthetic data may not be particularly important. More crucial might be that the direction of the effect is preserved, that is, a positive association should remain positive. Therefore, we also examine whether the sign between the input and

estimate is retained. Overall, the metrics allow us to assess central tendencies while accounting for variation across all datasets.

Paths A to E differ in how they establish the relationship between covariates and diseases, as well as in how co-occurrences of diseases are represented. Accordingly, the calculated difference between estimate and input will depend on which model is used for reverse-engineering. To ensure that each data generation path is treated equally, we perform the parameter estimation in five different ways: independent binomial logit models, dependent binomial logit models, a probit model, a multinomial logit model, and the computation of empirical correlations (where the type of correlation depends on the data type of each pair of variables). Figure 2 displays relative differences for the regression coefficient  $\beta_{j, \text{Age}}$  for all combinations of data generation paths and reverse-engineering methods. It also indicates how often the signs of the estimate and the input, i. e. the direction of the effect coincide. Regarding the employed methods, the combinations on the main diagonal match best, and thus are expected to deliver the smallest differences. These

method-wise optimal combinations are thus compiled for all six covariates in Fig. 3. A more detailed representation, analogous to Fig. 2, for the covariates beyond Age are provided in Figs. A1 to A5 in the Appendix.

Regarding the targeted prevalences and associations of single diseases or disease combinations, we compute the difference between the target value and the estimated value for each data generation path and target measure. Figure 4 displays average values across the 1,000 datasets and empirical 95 % ranges.

In addition to the graphical representations, we provide tabular overviews of estimated values of regression coefficients and correlation coefficients in Table A4 in the Appendix, and empirical disease prevalence and disease association in simulated data in Table A5 in the Appendix.

Summarizing the results, we observe that the input information (coefficients and correlations) can be largely recovered from the synthetically generated data if the data-generating process and the estimation method are similar to each other (Fig. 3 and main diagonal of Fig. 2), while the target properties (prevalences and association

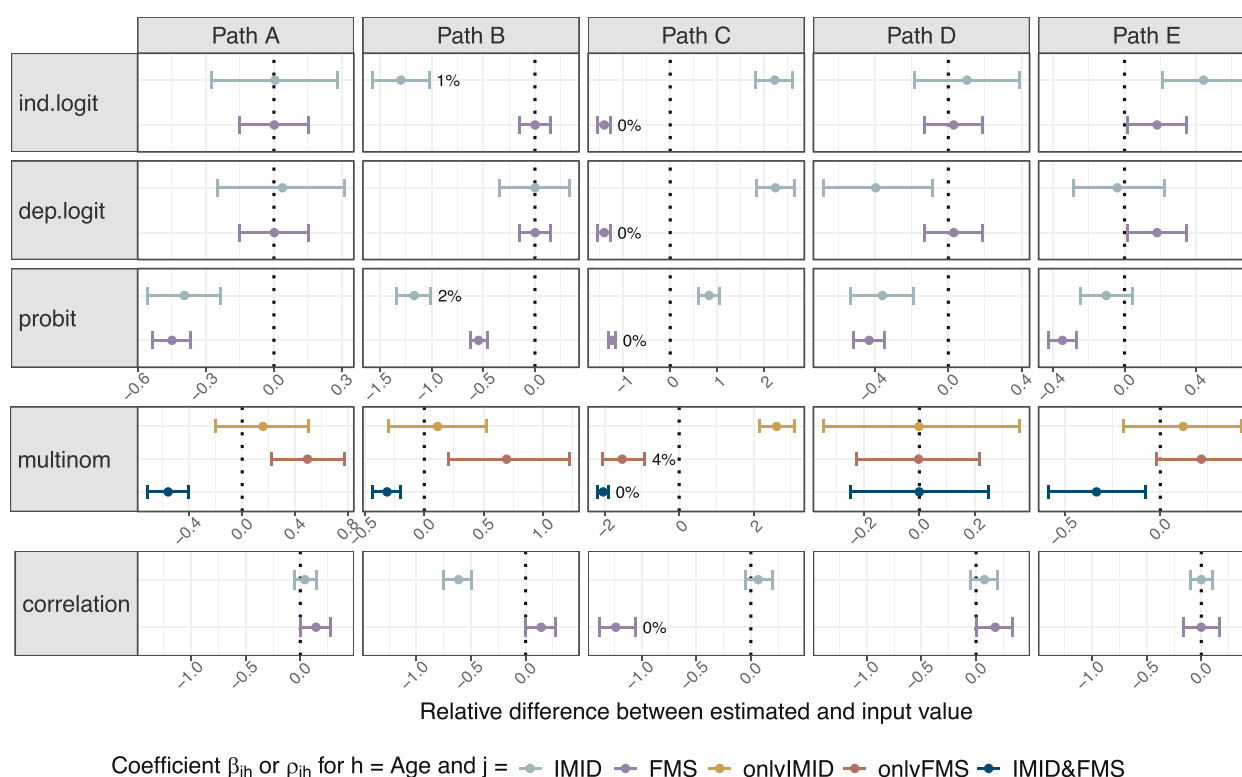


**Fig. 2** Relative difference between the estimated value of the regression coefficient  $\beta_{jh}$  or correlation coefficient  $\rho_{jh}$  and the input value. The plots display averages (dots) and 95 % empirical ranges (bars) across 1,000 generated datasets. Percentages indicate how often the signs of estimate and input coincide. Where no percentage is given, it is 100 %. Columns represent the data generating approach (Paths A to E), while rows represent ways to estimate  $\beta_{jh}$  or  $\rho_{jh}$ , respectively. These are independent binomial logit models ('ind.logit'), dependent binomial logit models ('dep.logit'), a probit model ('probit'), a multinomial logit model ('multinom'), and the computation of empirical correlations ('correlation'). The considered covariate is  $h = \text{Age}$ ; other covariates are considered in Figs. A1 to A5 in the Appendix. The target disease  $j$  covers IMID, FMS, onlyIMID, onlyFMS and IMID&FMS and is distinguished by color. Dotted vertical lines represent a difference of zero

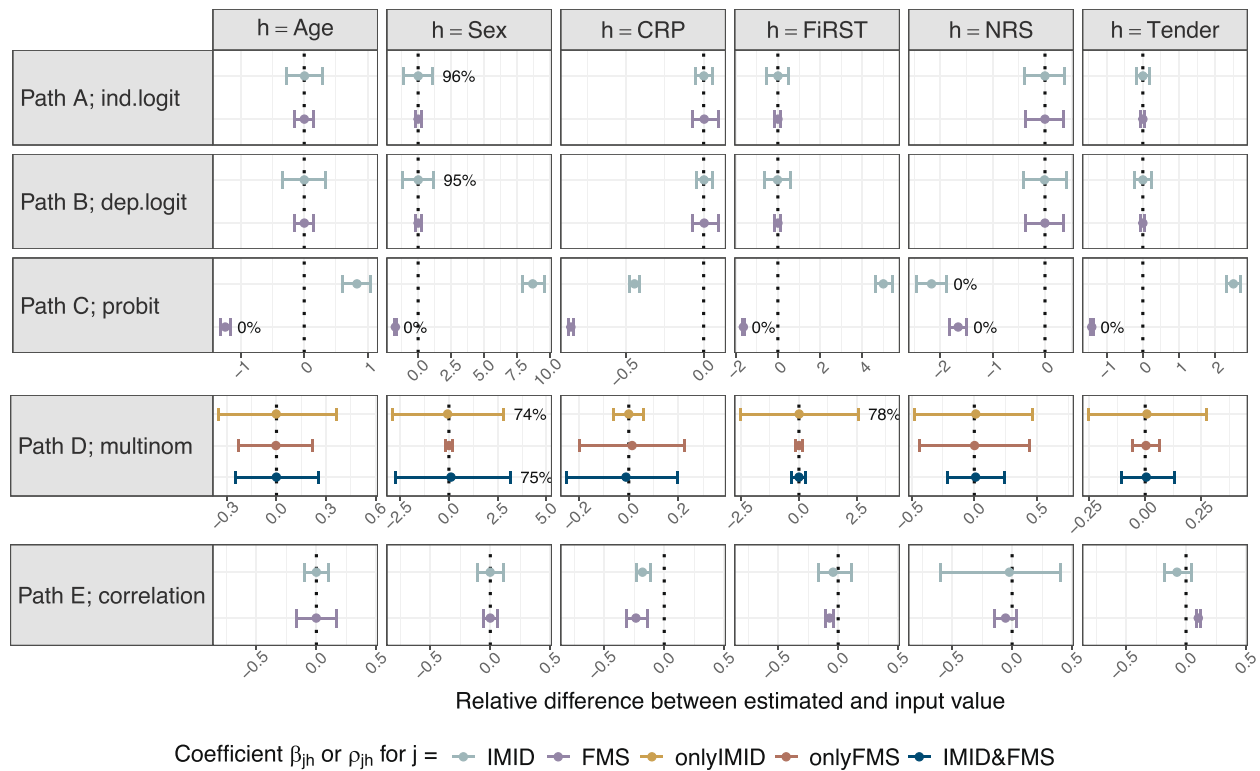


**Fig. 3** Relative difference between the estimated value of the regression coefficient  $\beta_{jh}$  or correlation coefficient  $\rho_{jh}$  and the input value. The plots display averages (dots) and 95 % empirical ranges (bars) across 1,000 generated datasets. Percentages indicate how often the signs of estimate and input coincide. Where no percentage is given, it is 100 %. Each row stands for a combination of data generation path and a way to estimate  $\beta_{jh}$  or  $\rho_{jh}$ , respectively (see also Fig. 2). Columns represent the six covariates  $h$ . The target disease  $j$  covers IMID, FMS, onlyIMID, onlyFMS and IMID&FMS and is distinguished by color. Dotted vertical lines represent a difference of zero

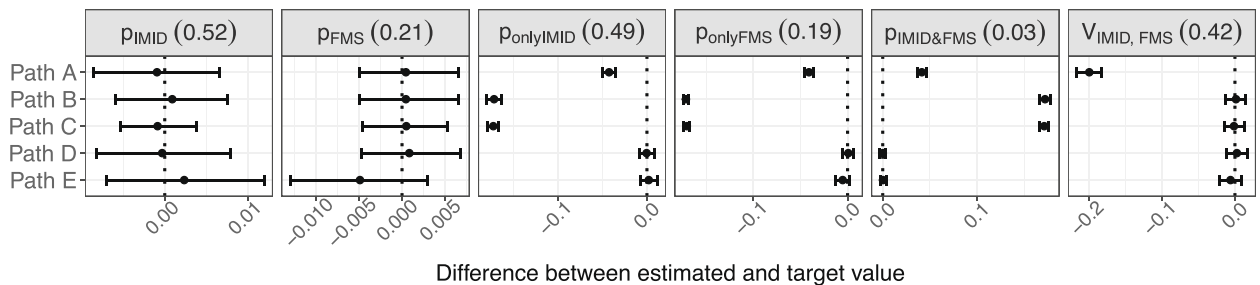


**Fig. 4** Difference between the estimated values and the target values for all disease (combination) prevalences and Cramer's V between IMID and FMS. The plots display averages (dots) and 95 % empirical ranges (bars) across 1,000 generated datasets, obtained via Paths A to E. The dotted line represents a difference of zero. In the last plot, the signs of estimate and target coincide agree for every single dataset

between diseases) are achieved to varying degrees (Fig. 4).

In particular, it hardly surprises that the regression coefficients generated from Paths A, B and D can be estimated precisely through independent logit models, dependent logit models and the multinomial logit model, where the utilized algorithms for generation and estimation directly correspond to each other. Path C, in contrast, introduces an additional dependence level. It makes use of latent factors, unknown for the estimation procedure, and consequently yields the highest deviations

between estimated regression coefficients and input values. It often even fails to maintain the direction of effect. With Path E, the correlation-based generation approach, we can reproduce the input correlations for all covariates with either no or only minor systematic deviation. The sign of the input correlation is maintained for all covariates for all datasets.

The targeted marginal disease prevalence for the single diseases IMID and FMS is obtained with all data generation paths. With respect to the prevalence of disease combinations (that is, onlyIMID, onlyFMS and

IMID&FMS), Paths D and E achieve the target and Path A comes comparably close, while Paths B and C show higher deviations.

Regarding the association between IMID and FMS, Path A leads to the largest deviation (in absolute values) for  $V_{\text{IMID, FMS}}$ , a consequence of modeling disease occurrence independently, with the only connection between diseases being established via the association with covariates. In contrast, Paths B to E reach the target value to the desired extent. With all paths, the obtained and targeted value of Cramer's V agree in their direction of association across all synthetic datasets.

## Discussion

We have developed and presented a framework that enables the generation of synthetic data in a medical context involving co-occurring and potentially dependent diseases. Our process consists of four steps: (I) the generation of patient covariates, (II) the calculation of disease predictors based on patient covariates, (III) the calculation of disease probabilities and scores based on predictors, and (IV) the simulation of disease occurrences based on probabilities and scores.

Within each step, various approaches are introduced, which can be combined in a modular fashion. Each of these approaches offers distinct advantages and limitations for specific scenarios. Some of these can be inferred theoretically from the underlying models and were already mentioned in [Methods](#) section. The use case from [Use case: Co-occurring diseases in chronic pain patients](#) section further provides insight into a subset of possible applications and practical implications using a concrete data example. At the same time, it serves exclusively to demonstrate how the framework can be used to generate synthetic data based on different modeling assumptions. The aim is not to infer or compare the real-world relationship between the diseases mentioned (IMID and FMS). The example illustrates how different data-generating approaches impose specific structural properties, and is intended solely to compare generation strategies with respect to their ability to reproduce specified assumptions.

In the following discussion, we focus on the point in the data generation process where the effect of covariates on diseases as well as the dependencies among diseases themselves are modeled and integrated into the data simulation. This corresponds to the transition from Step I to all alternatives within Step II (that is, the model-based Approaches IIa to IIId) and the direct transition from Step I to the correlation-based Approach IVe. The Paths A to E in the data simulation in [Use case: Co-occurring diseases in chronic pain patients](#) section were deliberately chosen to represent these five variants. For the remaining parts of the paths, we intentionally kept

methodological variation to a minimum in order to better isolate the sources of differences in results.

We describe characteristics of the five Approaches IIa to IIId and IVe and compare them in terms of the following aspects:

- **Covariate effects:** How are the intended effects of covariates on (predictors for) disease occurrence included as input into the simulation?
- **Comorbidities:** How does the approach control or induce an association between co-occurring diseases?
- **Causal relations:** What causal structure between diseases is assumed to underlie the data-generating process?
- **Target prevalence:** How is the desired disease prevalence targeted at, and to what extent is it achieved?
- **Use case:** In which context is the approach particularly meaningful?

The following discussion of properties and suitable use cases is also summarized in [Table 3](#).

In **Approach IIa**, we model disease occurrences as conditionally independent given the covariates. The effect of covariates on each disease predictor can be specified directly by regression coefficients. While we do not explicitly define the relationship between single diseases, a positive or negative correlation can be induced by according choices of regression coefficients; that is, through the values of regression coefficients  $\beta_{jh}$  and  $\beta_{lh}$  for different diseases  $j$  and  $l$  but identical covariate  $h$ . The prevalence of each disease can be approximately targeted at by choosing the intercept according to [Eq. \(2\)](#), which can be further optimized in Approach IVa. Approach IIa is primarily useful for creating conditionally independent disease occurrences, in particular if we have no suspicion of patient characteristics that promote or exclude a simultaneous occurrence, or of causal relationships between diseases. A limitation of the approach is that target associations between diseases are not directly specified but occur as a result of the impact of covariates. The correlation between diseases is further constrained by the target prevalences of the single diseases.

**Approach IIb** overcomes one limitation of Approach IIa: By generating diseases as a successive sequence, Approach IIb allows to explicitly model disease dependence. In particular, the occurrence of one disease can be modeled to depend not only on patient covariates but also on the presence of one or more other diseases. This comes at the cost of giving up another property, namely the control over effects of covariates on disease occurrence. The final association is blurred by the sequence of model layers, as could be observed in the use

**Table 3** Comparison of characteristics of approaches introducing the central dependence structure. **Covariate effects:** How are effects of covariates on disease occurrence included in simulation? **Comorbidities:** How does the approach control or induce an association between diseases? **Causal relations:** What causal structure between diseases is assumed to underlie the data-generating process? Arrows indicate the direction of causation; here exemplary shown for two binary disease variables A and B, covariates  $\mathbf{X}$  and latent factor variables  $\mathbf{W}$ . **Target prevalence:** How is the desired disease prevalence targeted at, and to what extent is it achieved? **Use case:** In which context is the approach particularly meaningful?

| Appr. | Covariate effects                                                                | Comorbidities                                                                           | Causal relations                                                                                                                                | Target prevalence                                                                          | Use case                                                                                            |
|-------|----------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------|
| Ila   | Directly specified via regression coefficients                                   | Diseases modeled independently; correlation inducible via regression coefficients       | No causality between diseases modeled: $\mathbf{X} \rightarrow A$ and $\mathbf{X} \rightarrow B$                                                | Approximately achieved through intercept (2); precisely achieved through adjustment in IVa | Independent diseases without causal assumption but with co-occurrence due to certain covariates     |
| Ilb   | For dependent diseases, specification is blurred due to overlapping model layers | Explicitly modeled via dependent regression models with respective coefficients         | Causality through sequentially dependent models: $\mathbf{X} \rightarrow A$ and $\mathbf{X} + A \rightarrow B$                                  | Approximately achieved through intercept (2); precisely achieved through adjustment in IVa | Sequential dependence of diseases                                                                   |
| Ilc   | Specification is blurred due to inclusion of latent factors                      | Specified via inclusion of shared latent factors and respective regression coefficients | Causal effect of shared latent factor on pairs of diseases: $\mathbf{X} + \mathbf{W} \rightarrow A$ and $\mathbf{X} + \mathbf{W} \rightarrow B$ | Approximately achieved through intercept (2); precisely achieved through adjustment in IVa | Assumption about shared but unknown common cause of disease occurrence                              |
| Ild   | Directly specified for disease combinations via regression coefficients          | Modeled via disease combination categories and odds ratios                              | Models the distribution of disease combinations; no explicit causality assumption                                                               | Approximately achieved through intercept (3); precisely achieved through adjustment in IVb | Knowledge about effect of covariates on occurrence of disease combinations                          |
| Ive   | Directly specified via pairwise correlations                                     | Directly specified via pairwise correlations                                            | No causal relation but correlation modeled                                                                                                      | Precisely achieved through success probability in marginal binomial distribution           | Correlated disease occurrences without causality assumption or knowledge on data-generating process |

case (e. g. Fig. 2, Path B). For those diseases that are modeled unconditional on another disease, however, the characteristics resulting from Approach Ila and Approach Ilb are equivalent since they use the same underlying model structure. Target prevalences are achieved as in Approach Ila, followed by IVa. Approach Ilb is particularly useful if a direct influence from one or several diseases on another one is assumed. The specification of target covariate effects on disease occurrence, however, becomes more complicated.

With **Approach Ilc**, we are able to create associations between disease occurrences by including shared latent factors: The occurrence of single diseases is modeled to depend on both individual patient covariates and additional joint latent factors, without the need to define a sequence of dependencies or model a direct causal relation. A key implication of introducing latent factors is that the input regression coefficients are not precisely reproducible from the generated datasets and may even change their sign as observed in the use case example (e. g. Fig. 2, Path C). Target prevalences are achieved as in Approaches Ila and Ilb, followed by IVa. Approach Ilc is particularly useful in cases where diseases are assumed to be dependent, but due to some unknown common cause rather than due to covariate effects as in Approach Ila (or at least not to the same extent). In this context it is important to note that the definition of relationships between diseases in this manner is restricted to the data-generating process; i. e. it does not imply causality

between diseases when interpreting the synthetic data, but the unobserved latent factors act as confounders in the underlying structure.

**Approach Ild** generates the simultaneous occurrence of diseases by modeling categorical outcomes representing disease combinations, rather than by modeling separate binary outcomes. The effect of covariates on disease combinations is explicitly specified by the regression coefficients and can be precisely reproduced from the generated dataset when applying a multinomial logit model (e. g. Fig. 2, Path D). The association between single diseases is induced implicitly rather than specified explicitly. However, the conditional association between two diseases can be specified through odds ratios as explained in Approach Ild, Step 1, Option (iii). Approach Ild is most appropriate if there is knowledge about the effect of covariates on the occurrence of disease combinations. The prevalence of each disease combination can be approximately targeted at by choosing the intercept according to Eq. (3), which can be further optimized in Approach IVb.

**Approach Ive** uses the NORTA algorithm, that means, unlike the previous approaches, it is fully based on specified correlations and marginal distributions instead of using a regression model for simulation. The primary advantage of this approach is its simplicity and the possibility to directly specify the linear relationship between all variables (covariate – disease and disease – disease) via a correlation matrix. Target prevalences are precisely

obtained through the indicated success probabilities of the respective marginal binomial distributions of single diseases. If one aims to synthetically generate covariate data through Step Ib (which is NORTA-based as well), Approaches Ib and IVe can be combined to one simulation step to make implementation even easier. Also, Approach IVe can easily be adapted to the scenario of simulating diseases combinations as in Approach IId. As a drawback of Approach IIe, there is no data-generating process modeling the distribution of disease occurrence conditional on covariates or other diseases. While the direction (positive or negative) of the covariate effects on disease occurrence can be influenced through the corresponding correlation, the effect strength is rather induced than specified. Moreover, changing the distribution of covariates while keeping correlations fixed might cause unintended effects on disease occurrence. Overall, Approach IVe is most useful if correlations between diseases and covariates are intended to be strictly controlled but no model is required. In practice, this approach is most applicable if knowledge about correlations is available from real data. In that case, it might be more straightforward to specify a correlation matrix rather than regression coefficients. Still, the approach exhibits limited scalability; as dimensionality increases, selecting a meaningful and valid correlation matrix becomes increasingly difficult.

To summarize, the theoretical considerations and empirical insights show that our modular toolkit provides a wide range of possibilities for generating the desired synthetic data. Thereby, the data generation process is designed to produce synthetic data that serves to validate statistical methods through simulation studies; for example, for the methodological comparison of single-label and multi-label classification which arose from our previous investigation [20]. Importantly, there are natural limits to fulfilling all wishes for synthetic data (precise specification of covariate effects, disease associations and prevalences) simultaneously. Thus, the user will make a selection of appropriate modules based on the available information and priorities.

## Conclusion

Reliable methods of statistical inference are a fundamental prerequisite for drawing reliable conclusions from medical data. Robust synthetic data, in turn, form the basis for the validation of statistical methods by means of simulation studies. This holds in general, and in particular for the case we consider: the analysis of co-occurring diseases in connection with patient information. We addressed the need for respective data by introducing a four-step framework consisting of several viable approaches for generating synthetic data in case of co-occurring diseases.

Apart from a theoretical classification of the model-based and correlation-based approaches, we tested the framework in practice and applied it to a deliberately chosen selection of five combinations of generation approaches to a specific use case, the co-occurrence of pain-causing diseases. We analyzed the impact of the choice of methodological approaches on covariate effects, association structures and disease prevalence. The approaches differ in their ability to induce specific characteristics regarding covariate effects and disease dependency. We have delineated the strengths and limitations, and recommended use cases for each strategy.

The presented framework for generating synthetic data is specifically intended for the simulation-based validation of statistical methods, rather than for developing medical therapies or investigating clinical relationships. Consequently, while domain-informed plausibility is beneficial, absolute medical realism is not required for its intended methodological scope. The appropriateness of a data generation approach critically depends on the considered use case, relevant target characteristics of the data, and underlying assumptions. In order to ensure meaningful synthetic data, it remains essential to select modules of the data generation process as well as input and target parameters with careful, informed consideration.

It is important to distinguish our framework from data-driven synthetic data generation approaches such as SMOTE-based augmentation techniques [38] and modern deep-learning-based tabular generators (e. g. generative adversarial networks, variational autoencoders, or diffusion-based models [39–41]). These methods rely on existing datasets and aim to reproduce empirical patterns or enhance predictive performance. In contrast, our framework generates fully synthetic data from explicit structural and distributional specifications, without the need for source data. The primary objective is the transparent and controllable construction of a predefined mathematical ground truth. This enables the precise specification of marginal distributions, regression parameters, prevalences and inter-disease dependencies for theory-driven methodological validation.

Although our work focuses on co-occurring diseases and their association with patient information, the modeling framework is, of course, transferable to any case of simultaneously occurring events and covariate information, also beyond the medical context. The detailed explanations and interpretations offered in this article can serve as general guidelines for researchers regarding synthetic data generation. As the primary objective of the framework is to provide precise ground truth control for method validation, it is explicitly intended as a theoretical tool. It does not capture the full clinical realism required for medical decision-making and thus cannot

be used to draw actionable clinical inferences, investigate real-world medical relationships, or support direct clinical decision-making.

Further, our approach opens several avenues for future research. Specifically, the individual approaches could be further extended or combined to account for more complex dependencies. One natural extension would be to further develop the latent factor framework by allowing latent factors to be explicitly correlated with observed covariates. In addition, future work could systematically investigate the generation and evaluation of partially synthetic or anonymized-only datasets. Overall, our paper lays a foundation for the principled simulation of synthetic data in the context of co-occurring diseases. By enabling rigorous method validation in controlled yet flexible settings, these approaches ultimately contribute to the development and assessment of statistical tools that can later be applied in real-world medical research and thereby support evidence-based decision-making.

#### Abbreviations

|       |                                       |
|-------|---------------------------------------|
| CRP   | C-reactive protein                    |
| FIRST | Fibromyalgia rapid screening tool     |
| FMS   | Fibromyalgia syndrome                 |
| IMID  | Immune-mediated inflammatory diseases |
| NORTA | Normal to anything                    |
| NRS   | Numeric rating scale                  |
| OA    | Osteoarthritis (OA)                   |
| OPCD  | Other pain-causing diseases           |
| OR    | Odds ratio                            |

#### Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12874-026-02966-2>.

Supplementary material, including additional calculations and results, is provided in the document 'Appendix: Validating methods for inferring co-occurring diseases: a flexible framework for simulating synthetic data'.

#### Acknowledgements

We would like to thank Martin Rudwaleit and Marvin-Hendrik Röchter for the insightful collaboration in a different research project that introduced us to the medical use case treated here. That project was supported by the Medical Research Start-up Fund of the Medical School OWL, Bielefeld University, and provided patient data. Further, we thank Elmar Spiegel for methodological discussions and valuable feedback.

#### Authors' contributions

H. M. and C. F. initiated the project and conceptualized the research goals. H. Marchi conducted literature research, method development and implementation, data analysis and visualization, and drafted the original manuscript. S. S. and T. S. gave input on structure and methods. All authors edited and critically revised the manuscript.

#### Funding

Open Access funding enabled and organized by Projekt DEAL. Not applicable.

#### Data availability

Due to data privacy protection, we are not allowed to share the data of the German hospital used in this work. Data generation was done using the statistical software R version 4.4.3 [42]. The code used for data generation and analysis is available at [https://github.com/fuchslab/synthData\\_diseases](https://github.com/fuchslab/synthData_diseases).

#### Declarations

##### Ethics approval and consent to participate

The hospital data used in this work originates from a study for which ethical approval was given by the Ethics Committee of the Westphalia-Lippe Medical Association and the University of Münster (Ethik-Kommission der Ärztekammer Westfalen-Lippe und der Westfälischen Wilhelms-Universität Münster); case number: 2021-426-f-S. Informed consent was obtained from all participants or their legal guardian.

##### Consent for publication

Not applicable.

##### Competing interests

The authors declare that they have no competing interests.

Received: 6 March 2026 / Accepted: 3 August 2026

Published online: 19 August 2026

#### References

1. Chatfield C. Model Uncertainty, Data Mining and Statistical Inference. *J R Stat Soc A Stat Soc*. 1995;158–3:419–66.
2. Mayer DG, Butler DG. Statistical validation. *Ecol Model*. 1993;68(1–2):21–32. [https://doi.org/10.1016/0304-3800\(93\)90105-2](https://doi.org/10.1016/0304-3800(93)90105-2).
3. Hastie T, Tibshirani R, Friedman JH. The elements of statistical learning: Data mining, inference, and prediction. Second edition ed. Springer Series in Statistics. New York: Springer; 2017.
4. Morris TP, White IR, Crowther MJ. Using simulation studies to evaluate statistical methods. *Stat Med*. 2019;38(11):2074–102. <https://doi.org/10.1002/sim.8086>.
5. Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. *J Am Med Inform Assoc*. 2013;20(1):144–51. <https://doi.org/10.1136/amiajnl-2011-000681>.
6. Sterne JAC, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ (Clin Res ed)*. 2009;338:b2393. <https://doi.org/10.1136/bmj.b2393>.
7. Little R, Rubin D. Statistical Analysis with Missing Data. 3rd ed. Hoboken: Wiley; 2019.
8. Liu Y, Stouffs R, Theng YL. Development of Synthetic Patient Data to Support Urban Planning for Public Health. *Innov Aging*. 2020;4(Suppl. 1). <https://doi.org/10.1093/geroni/igaa057.039>.
9. Wang Z, Myles P, Tucker A. Generating and evaluating cross-sectional synthetic electronic healthcare data: Preserving data utility and patient privacy. *Comput Intell*. 2021;37(2):819–51. <https://doi.org/10.1111/coim.12427>.
10. Du Y, Lin S, Huang Z. Generation of Semantic Patient Data for Depression. In: Siuly S, editor. Health information science. Cham: Springer; 2017. p. 102–12.
11. Gonçalves A, Ray P, Soper B, Stevens J, Coyle L, Sales AP. Generation and evaluation of synthetic patient data. *BMC Med Res Methodol*. 2020;20(1):108. <https://doi.org/10.1186/s12874-020-00977-1>.
12. Sella N, Guinot F, Lagrange N, Albou LP, Desponds J, Isambert H. Preserving information while respecting privacy through an information theoretic framework for synthetic health data generation. *NPJ Digit Med*. 2025;8(1):49. <https://doi.org/10.1038/s41746-025-01431-6>.
13. Zhou G, Catic A, Shabestari M, Young M, Li C, Poppe K. From Statistical Fidelity to Clinical Consistency: Scalable Generation and Auditing of Synthetic Patient Trajectories. *arXiv:2603.06720*.
14. Gonzales A, Guruswamy G, Smith SR. Synthetic data in health care: A narrative review. *PLoS Digit Health*. 2023;2(1):e0000082. <https://doi.org/10.1371/journal.pdig.0000082>.
15. Murtaza H, Ahmed M, Khan NF, Murtaza G, Zafar S, Bano A. Synthetic data generation: State of the art in health care domain. *Comput Sci Rev*. 2023;48:100546. <https://doi.org/10.1016/j.cosrev.2023.100546>.
16. Giuffrè M, Shung DL. Harnessing the power of synthetic data in healthcare: innovation, application, and privacy. *NPJ Digit Med*. 2023;6(1):186. <https://doi.org/10.1038/s41746-023-00927-3>.
17. Siam NH, Snigdha NN, Tabasumma N, Parvin I. Diabetes mellitus and cardiovascular disease: exploring epidemiology, pathophysiology, and treatment

- strategies. *Rev Cardiovasc Med*. 2024;25(12):436. <https://doi.org/10.31083/j.rcm2512436>.
18. Hirschfeld RMA. The Comorbidity of Major Depression and Anxiety Disorders: Recognition and Management in Primary Care. *Prim Care Companion J Clin Psychiatry*. 2001;3(6):244–54. <https://doi.org/10.4088/pcc.v03n0609>.
  19. Garofalo C, Cristiani CM, Ilari S, Passacatini LC, Malafoglia V, Viglietto G. Fibromyalgia and irritable bowel syndrome interaction: a possible role for gut microbiota and gut-brain axis. *Biomedicines*. 2023. <https://doi.org/10.3390/biomedicines11061701>.
  20. Schmiegel S, Marchi H, Röchter MH, Rudwaleit M, Fuchs C. Single-label and Multi-label Classification for Disease Recognition with Special Consideration of Comorbidities. *medRxiv*. 2025. <https://doi.org/10.64898/2025.12.23.25342901>.
  21. Hofert M, Kojadinovic I, Maechler M, Yan J.: copula: Multivariate Dependence with Copulas. R package version 1.1-6. <https://CRAN.R-project.org/package=copula>. Accessed 3 Feb 2026.
  22. Leisch F, Weingessel A, Hornik K.: bindata: Generation of Artificial Binary Data. R package version 0.9-22. <https://CRAN.R-project.org/package=bindata>. Accessed 3 Feb 2026.
  23. Fialkowski AC.: SimMultiCorrData: Simulation of Correlated Data with Multiple Variable Types. R package version 0.2.2. <https://github.com/afialkowski/simmulticorrdata>. Accessed 3 Feb 2026.
  24. Kammer M.: simdata: Generate Simulated Datasets. R package version 0.4.1. <https://CRAN.R-project.org/package=simdata>. Accessed 3 Feb 2026.
  25. Goldfeld K, Wujciak-Jens J. Simstudy: illuminating research methods through data generation. *J Open Source Softw*. 2020;5(54):2763. <https://doi.org/10.21105/joss.02763>.
  26. Touloumis A. Simulating correlated binary and multinomial responses under marginal model specification: the SimCorMultRes package. *R J*. 2016;8(2):79–91. <https://doi.org/10.32614/RJ-2016-034>.
  27. Walonoski J, Kramer M, Nichols J, Quina A, Moesel C, Hall D, et al. Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record. *J Am Med Inform Assoc*. 2018;25(3):230–8. <https://doi.org/10.1093/jamia/ocx079>.
  28. Nowok B, Raab GM, Dibben C. synthpop: Bespoke Creation of Synthetic Data in R. *J Stat Softw*. 2016;74(11):1–26. <https://doi.org/10.18637/jss.v074.i11>.
  29. McLachlan S, Dube K, Gallagher T, Simmonds JA, Fenton N, et al. Realistic Synthetic Data Generation: The ATEN Framework. In: Wiebe S, Anderson P, Saggio G, Zwiggelaar R, Gamboa H, et al., editors. *Cliquet Jr A. Biomedical Engineering Systems and Technologies*. Springer eBooks Computer Science. Cham: Springer; 2019. p. 497–523.
  30. Cario MC, Nelson BL. Modeling and Generating Random Vectors with Arbitrary Marginal Distributions and Correlation Matrix. Technical Report, Department of Industrial Engineering and Management Sciences, Northwestern University, Evanston, Illinois. 1997.
  31. Fahrmeir L, Kneib T, Lang S, Marx B. Regression: Models, methods and applications. Berlin and Heidelberg: Springer; 2013. <https://doi.org/10.1007/978-3-662-63882-8>. Accessed 3 Feb 2026.
  32. Greene W. Econometric Analysis. Pearson International; 2019. <https://elibrary.pearson.de/book/99.150005/9781292231150>. Accessed 3 Feb 2026.
  33. King G, Zeng L. Logistic Regression in Rare Events Data. *Polit Anal*. 2001;9(2):137–63. <https://doi.org/10.1093/oxfordjournals.pan.a004868>.
  34. Cramér H. Mathematical Methods of Statistics (PMS-9). vol. 9 of Princeton Mathematical Series. Princeton: Princeton University Press; 1946.
  35. Pearson K. VII. Mathematical contributions to the theory of evolution.—III. Regression, heredity, and panmixia. *Philos Trans R Soc Lond A Math Phys Sci*. 1896;187:253–318. <https://doi.org/10.1098/rsta.1896.0007>.
  36. Drasgow F. Polychoric and Polyserial Correlations. In: Kotz S, Read CB, Balakrishnan N, Vidakovic B, editors. *Encyclopedia of Statistical Sciences*. Wiley; 2004.
  37. Fox J.: polycor: Polychoric and Polyserial Correlations. R package version 0.8-1. <https://CRAN.R-project.org/package=polycor>. Accessed 3 Feb 2026.
  38. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic Minority Over-sampling Technique. *J Artif Intell Res*. 2002;16:321–57. <https://doi.org/10.1613/jair.953>.
  39. Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative Adversarial Nets. In: Ghahramani Z, Welling M, Cortes C, Lawrence N, Weinberger K, editors. *Advances in Neural Information Processing Systems*, vol. 27. Curran Associates, Inc.; 2014. [https://proceedings.neurips.cc/paper\\_files/paper/2014/file/f033ed80deb0234979a61f95710dbe25-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2014/file/f033ed80deb0234979a61f95710dbe25-Paper.pdf). Accessed 3 Feb 2026.
  40. Kingma DP, Welling M. Auto-Encoding Variational Bayes. 2022. [arXiv:1312.6114](https://arxiv.org/abs/1312.6114).
  41. Ho J, Jain A, Abbeel P. Denoising Diffusion Probabilistic Models. *CoRR*. 2020;abs/2006.11239. [arXiv:2006.11239](https://arxiv.org/abs/2006.11239).
  42. R Core Team.: R: A Language and Environment for Statistical Computing. Vienna. <https://www.R-project.org/>. Accessed 3 Feb 2026.

## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.