ScienceDirect

Contents lists available at sciencedirect.com
Journal homepage: www.elsevier.com/locate/jval

Methodology

Who Needs Real Data Anyway? Exploring the Use of Synthetic Data in Economic Evaluations of Health Interventions

Naomi van der Linden, PhD, Xavier G.L.V. Pouwels, PhD, Beate Jahn, PhD, Uwe Siebert, MD, Hendrik Koffijberg, PhD

ABSTRACT

Objectives: Data needed for economic evaluations in healthcare are often subject to privacy regulations and confidentiality, limiting accessibility. This poses challenges for conducting, reviewing, and validating health economic evaluations. The use of "synthetic data" may solve this problem.

Methods: An economic evaluation compared "shamectomy" with "usual care" for the prevention of a fictitious disease called shame. A data set (Dorg) was created, consisting of 1000 patients in the base case. Next, synthetic data (Dsyn) were created from Dorg. Dorg and Dsyn were used, separately, to inform a model-based economic evaluation, and the similarity of the results was assessed for various scenarios: different sizes of Dorg, order of synthetization, method of synthetization, number of synthesized data sets, and missing data.

Results: With standard settings, incremental cost-effectiveness ratio (ICER)-results for shamectomy were €25 848/quality-adjusted life-year in Dorg and on average €25 857 in 500 Dsyns, 95% CI (€16 776; €60 021). In the base case, 15% of the generated Dsyns resulted in an ICER leading to a positive reimbursement decision, as opposed to a negative decision when using Dorg. With smaller Dorg data sets (n = 50 and n = 500), ICER ranges increased to 95% CI (negative; €151 542) and 95% CI (negative; €669 717), respectively.

Conclusions: Outcomes and conclusions of economic analyses based on synthetic data may deviate from those obtained by using the original data. For data sets < 1000 patients, which are common, deviations may be substantial and lead to suboptimal policy decisions. Based on our results, we propose a stepwise approach to using synthetic data for model-based health economic evaluations, using a large number of synthetic data sets (ie, >100) with the same size as the original data.

Keywords: cost-effectiveness, simulation study, synthetic data.

VALUE HEALTH. 2025; 28(11):1722–1731

Highlights

- This simulation study seeks to illustrate the benefits and limitations of using synthetic data in economic evaluations of health interventions.
- Outcomes and conclusions of economic analyses based on synthetic data may deviate from those obtained by using the original data.
- Data synthesizers should generate a large number of synthetic data sets with the same size as the original data.

Introduction

The health and economic impact of health interventions is often evaluated based on a combination of (1) clinical data, eg, data on treatment effectiveness from randomized controlled trials, (2) patient-reported outcomes, eg, quality of life, (3) resource use, and (4) cost estimates. Compared with aggregate data, individual patient data (IPD) reflects details and heterogeneity more accurately.

In the absence of IPD, researchers often rely on aggregate data, such as published survival curves or summary statistics, which are digitized and converted into quasi-IPD using algorithms such as the Guyot algorithm.[1] Although this approach is common practice, it is not ideal. Aggregate data lack the granularity to capture patient-level variation, correlations between model parameters, and the influence of patient characteristics on event timing and outcomes. For example, it cannot adequately reflect how baseline factors might affect time to progression or resource use. This limitation can lead to oversimplified models that fail to capture the complexity of real-world patient experiences. In many cases, however, IPD with detailed information do exist, such as clinical trial data or patient registries, but they are often not collected for health technology assessment purposes and may not be directly usable because of privacy or accessibility constraints. This poses significant challenges for conducting, reviewing, and validating health economic evaluations.

Although various methods have been developed to address these hurdles,[2] these are not-yet common practice. Developments in data science from 1993[3] onward have led to methods that allow the creation of "synthetic data," which is an artificial reproduction of the original IPD with nearly identical statistical properties.[4] These methods provide data owners, including researchers, governmental parties, and clinical institutions, with a means to share synthetic data with the same relevance as the original data set, without revealing the (potentially) sensitive information contained in that original data set.

Importantly, the generation of synthetic data is typically performed by the owner of the original data set, who has direct access to the data and the necessary permissions to process it. This ensures compliance with privacy regulations while enabling external

modelers to conduct analyses using synthetic data that mimic the original data set's statistical properties. Generating synthetic data can, therefore, increase the usefulness of empirical data sets, through use of synthetic derivatives for developing, extending, or reviewing economic evaluations.

Synthetic data can be created from different types of data sets, including randomized controlled clinical trial data but also observational data from patient registries. The result could be used in different types of health economic evaluations. For example, synthetic data could be the basis for a model-based economic evaluation but also as part of a purely empirical (eg, trial-based) economic evaluation. In this article, we illustrate the impact of using synthetic data in a model-based economic evaluation because models are often needed to project results beyond the clinical trial duration or to combine data from different sources.

Although the use of synthetic data in health economic evaluations seems a promising development, its impact on outcomes and conclusions from such analyses is currently unclear. This article aims to evaluate the extent to which, and the conditions under which, the use of synthetic data leads to the same health economic outcomes and conclusions as the use of original data. It will do so based on one software package to generate synthetic data: the "synthpop" package in R.[5,6] Other software packages for synthetic data generation are also available, and several packages have been compared within review articles.[7–10] We chose to use synthpop in R because the use of R seems to be increasing among health economists, and we were most familiar with synthpop ourselves.

By providing an illustration using synthpop, our findings will contribute to a better understanding of the potential benefits and limitations of using synthetic data in economic evaluations of health interventions.

Methods

Approach

For our analysis of synthetic data sets, we will focus on a hypothetical disease called "shame."[11] Once individuals develop "shame," their quality of life decreases, and they cannot be cured. Shamectomy is an expensive procedure, to be performed during the asymptomatic state, "to prevent being ashamed in the first place"[11] and hereby delaying or preventing the onset of shame. The goal is to evaluate cost-effectiveness of "shamectomy" versus "usual care" for the prevention of "shame."

Figure 1A[11] shows the steps taken in this study. The next subsections of the article discuss each of these steps separately. In summary, to test synthetic data sets within this decision-making framework, first a fictitious data set (data set original: Dorg) was created, in which relevant characteristics (such as size and correlation between variables) can be manipulated. After this, 500 synthetic data sets (Dsyn) were created from Dorg. The model-based economic evaluation was then performed using Dorg and again using all 500 Dsyns, to assess the extent to which the results were similar. This was tested for a range of scenarios reflecting different characteristics of Dorg and different approaches taken in the synthesis approach, specified in the section "Define scenarios."

Create Model for the Economic Evaluation

A Markov model for cost-effectiveness analysis was created using the R package "heemod." Similar to a prior illustration of this package by Filipović-Pierucci et al,[11] a model is created to reflect an imaginary disease called "shame." The model has 3 health states: shameless (asymptomatic), ashamed (symptomatic), and death being an absorbing state, see Figure 1B.[11] In the model, once patients become symptomatic, their quality of life decreases, and they cannot be cured. The model aims to determine the cost-effectiveness of a preventative surgical treatment called "shamectomy" compared with "usual care," which slows the progression from shameless to ashamed. The model is run for 25 cycles, with a cycle length of 1 year (90% of modeled individuals died within 25 years).

Transition probabilities were estimated for the Markov model, separately based on Dorg and based on Dsyn. All transition probabilities were dependent on the time spent in the health state and estimated from the data set (described in next section) using Weibull models, which were fitted separately for each treatment arm, using the "flexsurv" package. In the first cycle of the shameless state, diagnostic costs were incurred. Additionally, in the shamectomy strategy, costs of the procedure were incurred in this first cycle. Mean utilities for each health state (shameless, ashamed, and death) were calculated from the data. Final outcomes were expressed as costs (in Euros, discounted at 3%) per quality-adjusted life-year (QALY) gained (discounted at 1.5%), in line with the Dutch guideline for health economic evaluations.[12] In the base-case analysis, total discounted costs and discounted QALYs of shamectomy were compared with usual care. Incremental costs and QALYs of shamectomy versus usual care were summarized in an incremental cost-effectiveness ratio (ICER) and compared with a willingness-to-pay (WTP) threshold of €20 000/QALY. Incremental net monetary benefit (iNMB) was calculated as the difference in QALYs multiplied by a WTP value of €20 000/QALY, minus the difference in costs ((deltaE $\times$ WTP) $-$ deltaC).

Determine Data Set Characteristics for Dorg

All analyses were performed in R version 4.3.0.[6] We created a data set to reflect the results of the fictitious SHAME-OFF trial, which investigated the cost-effectiveness of "shamectomy" versus "usual care" for the prevention of an imaginary, terminal disease called "shame."[11] Table 1 provides the steps taken in generating Dorg. Dorg is a dataframe of 1000 rows, each representing a simulated individual, and 10 columns. Each column is explained in Table 1. Although the data set is fictitious, we made our data generation choices to reflect "commonly seen" data characteristics, similar to what one would encounter in, for example, oncology trials.

Determine Dsyn

The synthetic data set was created with the "synthpop" package in R.[6] The R code used for the analyses is provided in Appendix 1 in Supplemental Materials found at https://doi.org/10.1016/j.jval.2025.06.007.

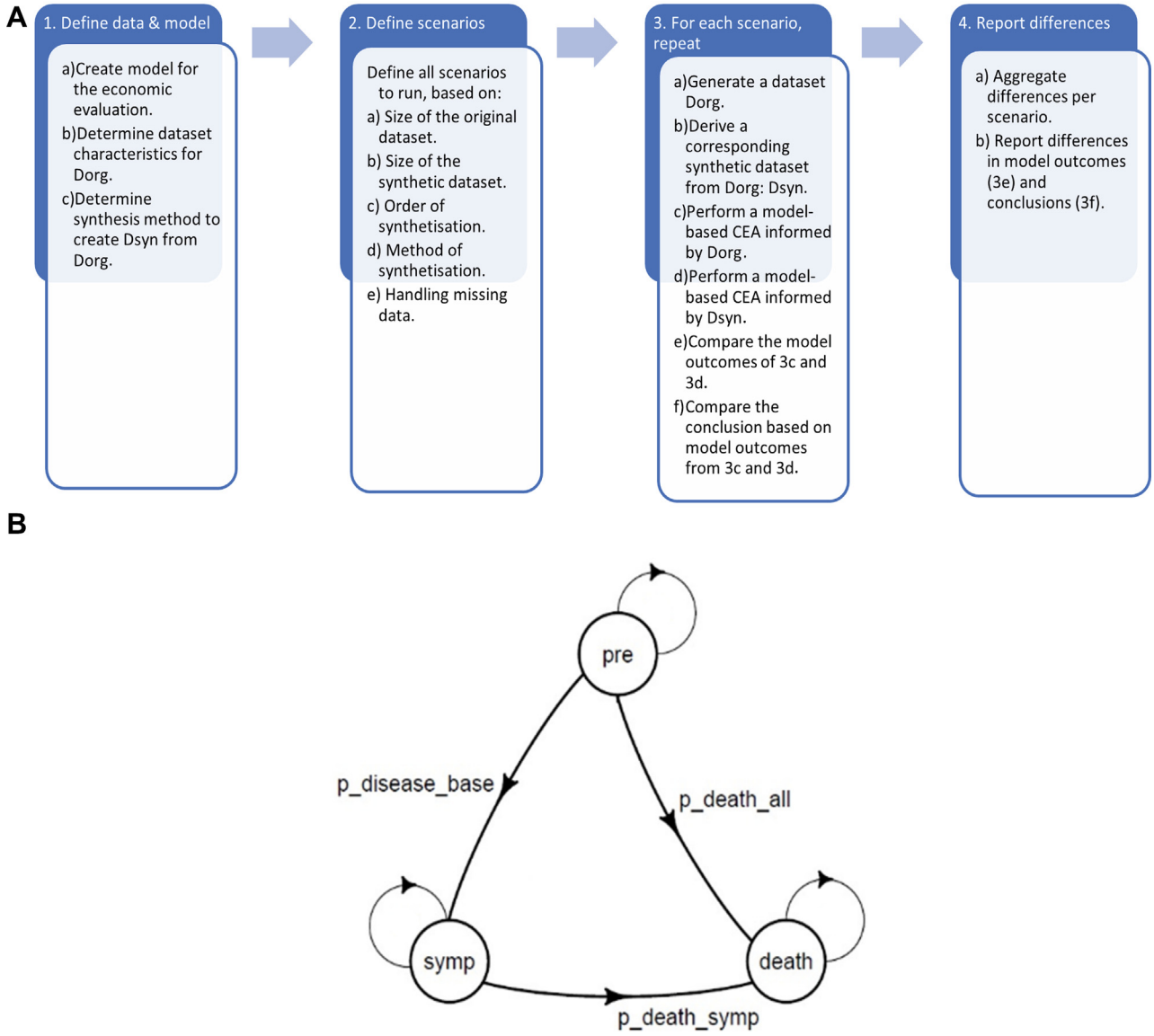
All data were synthesized using the default command "syn()" in the synthpop package.[5] By default, the algorithm synthesizes all variables in columns from left to right. When synthesizing the data, for every variable to be synthesized, a value is drawn from a conditional distribution, which is defined based on the variable of interest and values of previously synthesized columns: this is called "simple synthesis."

Note that "Time to death" was transformed to "Time from progression to death" before creating the synthetic data sets so that the correct order of events remained intact.

Define Scenarios

After the base-case analysis on Dorg, this analysis was repeated for 500 different sets of Dsyn, created by using 500 different seeds (1:500), to check the variation in outcomes. Subsequently, scenario analyses were performed as described in Table 2.

Figure 1. (A) Analyses steps. (B) Model structure for the "usual care" strategy, adapted from Filipović-Pierucci et al.[11]



CEA indicates cost-effectiveness analysis; Dorg, original data set; Dsyn, synthetic data set.

Analyses per Scenario

To validate all variables in the data sets, conditional means by treatment arm, and covariance matrices for Dorg versus 500 Dsyns were determined. Subsequently, estimated ICERs and NMB were compared between the original analysis and the analysis based on synthetic data in the base-case scenario and all additional scenarios. Mean ICER based on all Dsyns was reported, to evaluate whether any systematic deviation existed compared with the ICER based on Dorg. Intermediate outcomes were also provided: event-free survival (EFS), overall survival, costs, and QALYs.

A measure for the utility of synthetic data, the propensity mean squared error (pMSE), was calculated to signal the extent to which the data can be classified as "original data" versus "synthetic data." The pMSE quantifies how well the synthetic data replicate the statistical properties of the original data set by comparing the propensity scores of the original and synthetic data sets. The propensity score represents the probability that a record belongs to the synthetic data set rather than the original data set. The formula for pMSE is as follows:

$$pMSE = \frac{1}{n}\sum_{i=1}^{n}\left(p_i - 0.5\right)^2$$

in which n is the number of records in the data set, p_i is the predicted probability (propensity score) that the i-th record is synthetic. If all p_i values equal 0.5, the synthetic and original data sets are indistinguishable, and pMSE would equal 0.[13] Note that pMSE quantifies the similarity of data sets, not the similarity of any outcomes obtained from analyzing those data sets.

Table 1. Steps taken in generating Dorg.

Person ID, column 1
Numeric, person identifier
Time-to-event data (clinical), columns 2 and 3
Time_to_shame and time_to_death were generated for both treatment strategies separately, from correlated Weibull distributions. For time_to_shame, the shape parameter was set at 1.5 and the scale parameter was set at 8 for usual care. For death, the shape parameter was set at 1.5 and the scale parameter was set at 12 for usual care. Correlation was set at 0.8, using a Gaussian copula (transforming the marginal distributions of the variables into uniform distributions and then using the correlation parameter to adjust the dependency between them). Time_to_shame was capped at time_to_death, so that progression from shameless to ashamed could never take place after death. Furthermore, time_to_death was transformed to time_from_shame_to_death before creating the synthetic data sets, to ensure that progression will always take place before or at the time of death. The probability of developing symptoms was assumed lower for individuals in the "shamectomy" strategy. This was implemented by multiplying time_to_shame by 2 for all patients in the 'shamectomy' strategy.
Status death, column 4
Dichotomous variable indicating whether a simulated participant was death or alive at the end of follow up. 1 = death, 0 = alive.
Status shame, column 5
Dichotomous variable indicating whether a simulated participant experienced shameless during follow up. 1 = ashamed, 0 = shameless.
Strategy, column 6
Dichotomous variable indicating in which strategy a simulated participant belongs. 1 = shamectomy, 0 = usual care. Per strategy, data were generated for 500 individuals.
Cost data, columns 7 and 8
The diagnostic costs incurred in the shameless state were randomly drawn from a normal distribution with a mean of €5000, SD of €2000 and truncated at €0 for each patient, irrespective of the arm. Additionally, patients in the shamectomy arm were assigned an additional cost of the surgery, defined by a right-skewed beta distribution with shape 1 = 2, shape 2 = 7, multiplied by €100 000. This results in a different cost per patient, with mean surgical cost of ∼€22 000.
Utility data, columns 9 and 10
Each unique patient was assigned a random utility from a normal distribution with a mean of 0.8, SD of 0.05 and truncated between 0 and 1. This utility value represents their quality of life before developing "shame." Utility values associated with the ashamed state of each individual were approximately halved, by multiplying the "shameless" utility by 0.5 and by a random noise variable drawn from a uniform distribution with minimum 0.95 and maximum 1.05, to prevent collinearity in the data set. Dead patients were assigned a utility of 0.

For each of the scenarios, it was evaluated in which proportion of cases the policy-relevant advice would change when using synthetic instead of original data. This was based on whether the ICERs were above or below the WTP threshold of €20 000/QALY.

In addition to generating results for 500 Dsyn, based on 1 seed value for Dorg, all scenarios 1 to 4 were performed with 50 Dsyn for 20 different seed values for Dorg.

Results

Point Estimates

A comparison of conditional means by treatment arm and covariance matrices for Dorg versus 500 Dsyns is provided in Appendix 2 in Supplemental Materials found at https://doi.org/10.1016/j.jval.2025.06.007.

The base-case ICER was €25 848/QALY gained, based on Dorg (with n = 1000), which is above the €20 000/QALY-gained WTP threshold and would lead to the advice not to reimburse shamectomy. The iNMB was −€4857.

The ICER estimated based on dividing the mean incremental costs by the mean incremental QALYs across the 500 Dsyns was €25 857/QALY gained. The mean iNMB in the synthetic data sets was −€4,865 and ranged from −€20,739 to €11,166 (Fig. 2, Table 3).

Scenario Analyses and Policy Decision

See Appendix 2 in Supplemental Materials found at https://doi.org/10.1016/j.jval.2025.06.007 for the results of the missing data scenarios (5A-5F, Appendix Table 4), and the scenario results based on the average of 20 runs (repeats with different seeds) with 50 synthetic data sets (see Appendix Table 5 in Supplemental Materials found at https://doi.org/10.1016/j.jval.2025.06.007). For scatterplots of each of the scenarios, see Appendix 3 in Supplemental Materials found at https://doi.org/10.1016/j.jval.2025.06.007. In summary, results were largely determined by the size of Dorg (scenario 1). Larger sizes of Dsyn reduced spread in the outcomes (scenario 2).

With the smallest Dorg (n = 50), the parameters of the model could not be defined for 17% of the synthesized sets because of a lack of progression and/or death events generated in Dsyn. For all scenarios with n = 1000 in Dorg, the ICERs were rather similar, with less than €800 difference between ICERs. However, in up to 24% of the generated Dsyn sets with a Dorg of n = 1000, an advice on reimbursement based solely on the ICER would have been positive, as opposed to negative in the base case. With smaller Dorg data sets (n = 50 and n = 500), estimated ICERs lay further apart, with more than €35 000 difference between ICERs. Smaller values of pMSE were observed for larger Dsyns, which captured all variation present in Dorg.

Table 2. Scenario analyses.

Scenario 1: Size of the original data set	The base case on Dorg includes (2 × 500 =) 1000 individuals. Three subscenarios are performed to determine the impact of size of the original (training) data set: 1A: Dorg contains (2 × 25 =) 50 patients to create the synthetic data. 1B: Dorg contains (2 × 250 =) 500 patients to create the synthetic data. 1C: Dorg contains (2 × 5000 =) 10 000 patients to create the synthetic data.
Scenario 2: Inflated synthetic data set	In the base case and scenario 1, the size of Dsyn is equal to the size of Dorg. In scenario 2, the synthetic data sets are inflated to contain 10 000 individuals. Three subscenarios are performed, with different sizes of Dorg: 2A: Dorg contains 50 patients and is used to create Dsyn with 10 000 individuals. 2B: Dorg contains 500 patients and is used to create Dsyn with 10 000 individuals. 2C: Dorg contains 1000 patients and is used to create Dsyn with 10 000 individuals.
Scenario 3: Order of synthetization	By default, the algorithm synthesizes all variables in columns from left to right. To determine the impact of synthetization order, the analysis is repeated with the opposite order: from right to left. This is done by using the visit.sequence argument in the synthpop package.[4]
Scenario 4: Proper synthesis	When synthesizing the data, observed values are replaced by sampling from a probability distribution conditional on the following: the variable to be synthesized, the values from previously synthesized columns of the original data set, and the fitted parameters of either (1) the conditional distribution (this is called "simple synthesis") or (2) the posterior predictive distribution of parameters (this is called "proper synthesis").[4] In the base-case analysis, "simple synthesis" is performed. This scenario uses "proper synthesis," by setting the argument "proper" to "TRUE" in the synthpop package.[4]
Scenario 5: Missing data, prior	When part of the data are missing, several approaches are possible to impute the missing values. We test 2 of these: (1) to impute missing values before synthesis (scenarios 5A-5C), and (2) not to impute missing values before synthesis, but complete them as part of the synthesis procedure (scenarios 5D-5F). 5A: In this scenario, 10% missings are introduced to the data in Dorg. This was done at random locations before imputation and the generation of each of the 500 Dsyn. The missing data were imputed (m = 5) using the mice package, before synthesis, creating 3 synthetic data sets for each of the 2500 imputed Dorg files. 5B: In this scenario, 25% missings are introduced to the data in Dorg. This was done at random locations before imputation and the generation of each of the 500 Dsyn. The missing data were imputed (m = 5) using the mice package, before synthesis, creating 3 synthetic data sets for each of the 2500 imputed Dorg files. 5C: In this scenario, 50% missings are introduced to the data in Dorg. This was done at random locations before imputation and the generation of each of the 500 Dsyn. The missing data were imputed (m = 5) using the mice package, before synthesis, creating 3 synthetic data sets for each of the 2500 imputed Dorg files. 5D: In this scenario, 10% missings are introduced to the data in Dorg, at random locations for each of 500 different sets. The data are not imputed before synthesis but completed as part of the synthesis procedure, creating 3 synthetic data sets for each of the 500 files. 5E: In this scenario, 25% missings are introduced to the data in Dorg, at random locations for each of 500 different sets. The data are not imputed before synthesis but completed as part of the synthesis procedure, creating 3 synthetic data sets for each of the 500 files. 5F: In this scenario, 50% missings are introduced to the data in Dorg, at random locations for each of 500 different sets. The data are not imputed before synthesis, but completed as part of the synthesis procedure, creating 3 synthetic data sets for each of the 500 files.

The mean of 20 runs (repeated with different seeds) of 50 Dsyns confirmed that the outcomes and conclusions based on an individual synthetic data set may deviate substantially from the means obtained using all synthetic data sets.

The missing data scenarios suggest that the imputation of missing data as part of the synthesis procedure outperforms our current approach in scenarios 5A to 5C, in which missing data were imputed before synthesis, with respect to the similarity of the scenario results to the Dorg results.

Stepwise Approach

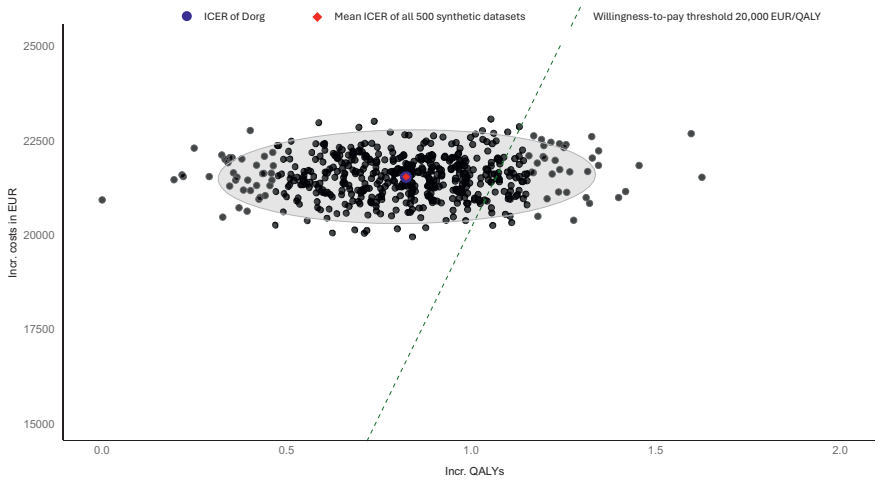
In Figure 3,[14] we suggest a stepwise approach to the use of synthetic data to inform model-based health economic evaluations, based on the findings of this simulation study.

Discussion

Findings

This simulation study using the synthpop package in R indicates that both outcomes and conclusions of cost-effectiveness analyses based on synthetic data may deviate from those obtained by cost-effectiveness analyses based on the underlying original data. Unless the original data set is very large, deviations may be substantial and lead to different, that is, suboptimal, policy decisions regarding reimbursement of interventions. Based on our findings, we recommend considering the use of synthetic data only when the original data set contains more than 1000 records. For substantially smaller data set sizes, a comprehensive analysis demonstrating sufficient similarity between synthetic and original

Figure 2. Base-case results for Dorg and results for 500 sets of Dsyn (not a probabilistic analysis).



Dorg indicates original data set; Dsyn, synthetic data set; ICER, incremental cost-effectiveness ratio; QALY, quality-adjusted life-year.

data would be warranted but may still not lead to acceptance of results based on synthetic data.

Currently available synthetic healthcare data sets reported on in a literature review by Gonzales et al,[15] contain data on 30 000 to 6.8 million people, with up to 300 million records. An example is Medicare and Medicaid claims data, with 6.8 Million beneficiary records and 112 million claims records.[15] However, although large underlying ("true") databases are common for administrative health record data, this is not the case for primary health data collections, such as clinical trial data. In clinical trials, sample sizes < 100, and between 100 and 1000 are very common, and only 6% of trials have a sample size > 1000.[16] Consequently, one should be careful when using synthetic data based on such trials to inform health economic analyses and guide policy making in healthcare. At minimum, extensive and transparent validation of the similarity of the synthetic data to the original data is needed to start building confidence in health economic analyses using those synthetic data.

For data owners, when generating synthetic data, we strongly recommend publishing of a large number of synthetic data sets (ie, at least 100) all generated based on the same original data set and with that sample size and using the same synthetization settings. This allows analysts to gain insight into the uncertainty introduced by the synthetization process, by exploring the spread in outcomes. In the absence of bias in this process, this also allows for obtaining a more realistic approximation of the original data outcomes by averaging the synthetic data outcomes (such as in Fig. 2).

Some data owners may be tempted to inflate synthetic data sets to a large and potentially "round" number (eg, generating 10 000 records from an original data set of 1234). This may incorrectly appear to be beneficial because larger data sets are often associated with greater statistical power and precision and narrower confidence intervals, potentially leading to broader use. However, this practice is misleading because inflated synthetic data no longer properly reflect the true variability and uncertainty present in the original data set. Therefore, synthetic data sets should maintain the same sample size as the original data set to ensure that the variability and uncertainty inherent in the original data are accurately represented.

When receiving synthetic data from a data owner, the modeler should first verify whether their size matches the original data set. If the synthetic data set is inflated, uncertainty in model parameters may be underestimated. To address this, probabilistic estimates can be derived from repeated random subsamples of the synthetic data, each matching the original data set's size. Although a single subsample might not perfectly reflect the original data's properties, aggregating results across many subsamples helps approximate the true variability.

For rigorous uncertainty quantification, particularly in probabilistic or value of information analyses,[17,18] a more comprehensive approach is needed. Similar to combining bootstrapping with multiple imputation, the modeler can bootstrap the original data set and generate multiple synthetic data sets per bootstrap sample. Robust methods[19,20] can then be applied to derive parameter estimates and distributions that appropriately reflect the combined uncertainty.

Synthetic data can offer advantages over aggregate data by capturing patient-level variation, correlations between variables, and the influence of patient characteristics on event timing and outcomes. These features are lost when relying on aggregated summaries. The added value of synthetic data depends largely on the comprehensiveness of the underlying clinical data. When the original data are limited in scope or detail, synthetic data may offer little advantage over aggregate data. However, when the original data set is rich and detailed, synthetic data can provide a more nuanced representation of patient-level variability, correlations, and interactions. Further research would be needed to illustrate this because this simulation study used highly simplified data and modeling.

Limitations of This Study

Our results depend directly on the use of the synthpop package in R. The synthpop package uses sequential modeling, generating replacements by drawing from conditional distributions fitted to the original data using parametric or classification and regression trees models.[5] This is just one of many existing methodologies to generate synthetic data. Other methods could potentially result in different outcomes. However, experimental

analysis by Goncalves et al[21] suggests that there is no single method that outperforms the others in all considered metrics, such as metrics for data utility (eg, Kullback-Leibler divergence) and privacy disclosure risk. Future studies could compare the performance of alternative packages for creating synthetic data for use in cost-effectiveness analyses and their influence on decision making.

Although the use of the synthpop package in this study provides a practical and accessible approach to synthetic data generation, it is important to acknowledge that alternative synthesis methods, such as Bayesian models, generative adversarial networks, or tree-based models, were not explored.[8] These methods may offer distinct advantages in terms of data fidelity, flexibility, or scalability, depending on the context and complexity of the data set. For instance, Bayesian approaches can incorporate prior knowledge and uncertainty quantification, generative adversarial networks excel in capturing complex data distributions, and tree-based models may provide interpretability and robustness.[8,22] Future research could focus on a comparative analysis of these methods to identify their relative strengths and limitations in generating synthetic health data, particularly for use in health economic evaluations. Such exploration could help determine whether alternative methods might reduce the variability in outcomes observed in this study, especially for smaller data sets, and provide more robust synthetic data for policy decision making.

In this study, we purposefully chose to use default settings in most analyses. For example, the order of synthesis was left to right in the base case, and right to left in scenario 3. In reality, another, customized order of synthesis may better preserve the statistical properties of the original data set (in our simulations, alternative positioning of the "strategy" column in Dorg did not bring Dsyn ICER or iNMB closer to Dorg outcomes, see Appendix Table 3 in Supplemental Materials found at https://doi.org/10.1016/j.jval.2025.06.007). Therefore, it may be possible to further reduce the differences in estimates between Dorg and Dsyn by modifying some of the synthesis choices. This study purposefully used readily available functions to generate synthetic data, with their default settings. We expect that these are likely to be used in practice because optimizing the settings of such functions based on specific data set(s) would require more advanced knowledge of the synthetization algorithm, which might be lacking.

A range of measures have been proposed to quantify the utility of synthetic data sets to compare different synthesis approaches or to diagnose cases in which the original and synthetic data distributions differ and thus tune the synthesis methods.[23] Although this could lead to more representative and useful synthetic data sets, it also requires substantial statistical experience

Table 3. Scenario results based on 1 Dorg and 500 synthetic data sets.

Scenario	Mean EFS treatment	Mean EFS comp	Incr. EFS	Mean OS treatment	Mean OS comp	Incr. OS
Scenarios in which Dorg *n* = 1000						
Original data (*n* = 1000)	10.69	6.71	3.98	11.42	10.41	1.01
Base case using synthetic data (*n* = 1000 in Dorg and Dsyn)	10.73	6.71	4.02	11.49	10.38	1.11
2c (*n* = 1000 in Dorg and *n* = 10 000 in Dsyn)	10.73	6.72	4.01	11.49	10.38	1.11
3 (synthesis from right to left)	10.73	6.71	4.02	11.49	10.38	1.11
4 (proper synthesis)	10.69	6.70	3.99	11.48	10.27	1.22
Scenarios in which Dorg *n* = 50						
Original data (*n* = 50)	9.07	8.95	0.12	9.67	14.64	−4.97
1a (*n* = 50 in Dorg and Dsyn)*	8.41	9.27	−0.86	9.42	14.59	−5.17
2a (*n* = 50 in Dorg and *n* = 10 000 in Dsyn)	8.56	9.41	−0.85	9.58	14.84	−5.26
Scenarios in which Dorg *n* = 500						
Original data (*n* = 500)	9.52	6.90	2.63	10.33	10.50	−0.17
1b (*n* = 500 in Dorg and Dsyn)	9.47	6.96	2.51	10.38	10.36	0.02
2b (*n* = 500 in Dorg and *n* = 10 000 in Dsyn)	9.47	6.94	2.53	10.37	10.33	0.04
Scenarios in which Dorg *n* = 10 000						
Original data (*n* = 10 000)	9.98	6.90	3.08	10.77	10.82	−0.05
1c (*n* = 10 000 in Dorg and Dsyn)	9.96	6.90	3.06	10.75	10.82	−0.07

Comp indicates comparator; Dsyn, synthetic data; EFS, event-free survival; ICER, incremental cost-effectiveness ratio; Incr., incremental; iNMB, incremental net monetary benefit; OS, overall survival; pMSE, propensity score mean square error from the utility model; QALY, quality-adjusted life-year; WTP, willingness-to-pay threshold (in this case €20 000/QALY, see below).
*In 86 of 500 (17%) of the synthesized data sets, the distributions for the model could not be fitted because of a lack of events generated in Dsyn.
†In 52 of 500 (10%) of the synthesized data sets, ICERs were negative because of negative incremental costs.

and time, which may be limited in a practical clinical setting. Users of synthetic data sets would benefit from extensive reporting by the data owner on the synthetization steps and settings and similarity to the original data. Furthermore, tools such as "SynthRO"—a dashboard to evaluate and benchmark synthetic tabular data—may help select the most suitable synthetic data model for individual use cases.[24]

In addition to the synthesis choices, other choices were made in creating Dorg and building the Markov model. With each of these choices, we aimed to keep our approach as simple and transparent as possible, while still mimicking a realistic model-based economic evaluation. For example, our analysis was based on a relatively simple 3-state Markov model, which may not fully capture the complexity of health economic evaluations. In practice, models often include additional health states and more intricate cost and outcome structures. Generalizability of our findings to other types of data and models is, of course, not guaranteed. Future studies should explore how the use of synthetic data in such complex models influences outcomes and decisions because the increased dimensionality and interactions between variables may amplify deviations between outcomes based on synthetic versus original data. Investigating these scenarios would provide a more comprehensive understanding of the strengths and limitations of synthetic data in health economic evaluations and help identify contexts in which its use is most acceptable. To allow further exploration of this research question, we provide our R code (see Appendix 3 in Supplemental Materials found at https://doi.org/10.1016/j.jval.2025.06.007) so that readers can modify the choices we made and explore the resulting impact on the findings.

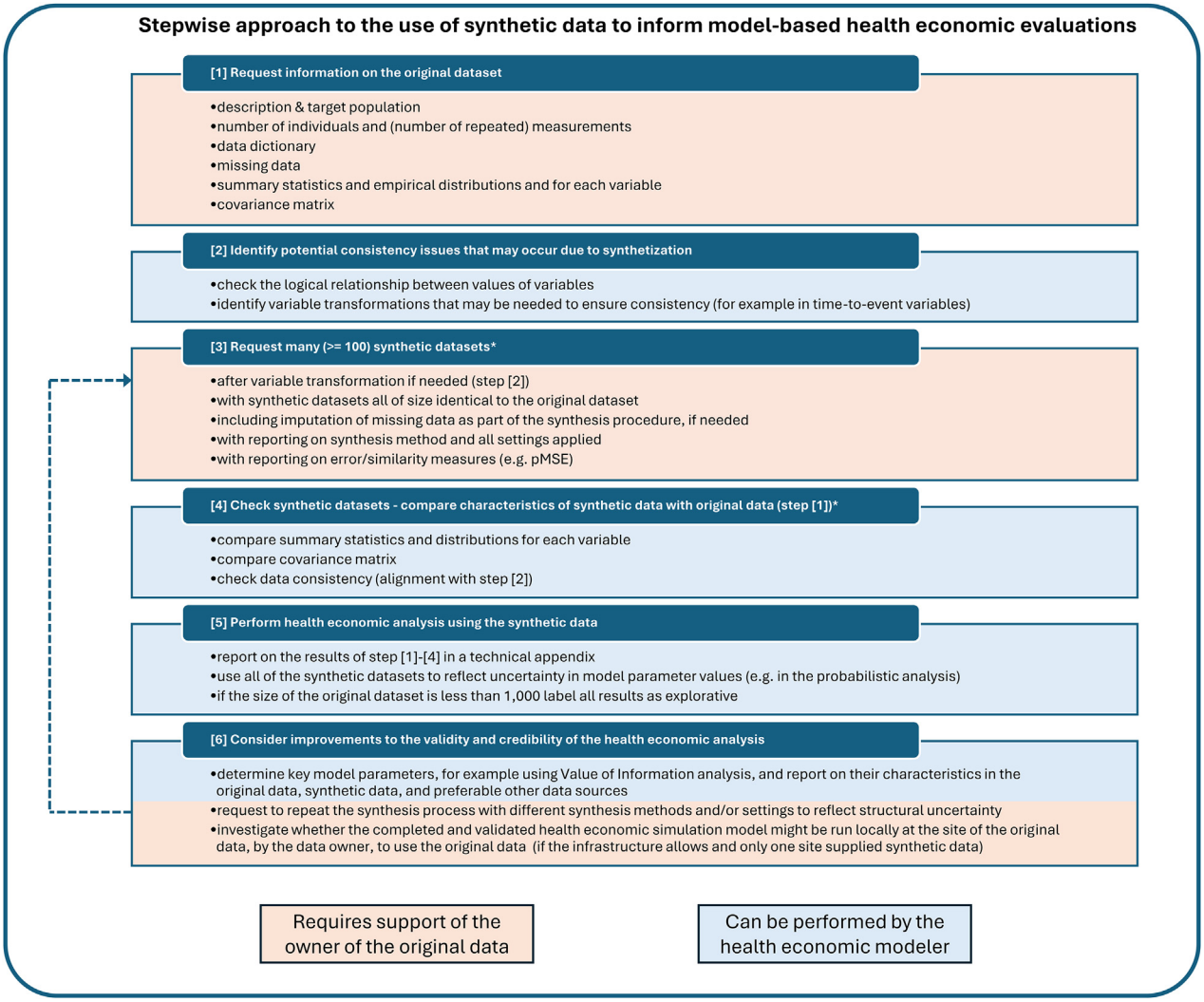
Although our study provides valuable insights into the use of synthetic data in health economic evaluations, its reliance on a self-created "original" data set limits external validity. In fact, generating synthetic data from actual real-world data may well lead to larger deviations in cost-effectiveness outcomes as shown here if the occurrence of measurement errors, missing data patterns, heterogeneity, and outliers negatively affects the accuracy of data synthetization methods. To address this, future research should replicate our methodology using real-world data sets to support a more robust assessment of the utility of synthetic data in practical settings. By validating our findings on real-world data sets, researchers can better understand the generalizability of synthetic data in health economic evaluations and refine synthesis methods to improve their reliability.

Alternatively, future studies could enhance the realism of synthetic data by introducing greater separation between the stages of data generation, synthesis, and analysis. For example, the team generating Dorg could be independent of the team synthesizing Dsyn, and analysts building the economic model could be

Table 3. Continued

Δcosts (SD)	ΔQALYs (SD)	Mean ICER in €/QALY gained	iNMB	iNMB 95% CI iNMB range (min; max)	Probability of Dsyn with ICER < WTP	pMSE (min; mean; max)
Scenarios in which Dorg n = 1000						
21 468	0.83	25 848	−4857	NA	NA	NA
21 476 (576)	0.83 (0.24)	25 857	−4865	−5286; −4443 −20 739; 11 166	0.15 (75/500)	0.000; 0.183; 0.192
21 455 (186)	0.82 (0.07)	26 148	−5045	−5172; −4918 −9274; −404	0.00 (0/500)	0.029; 0.031; 0.032
21 476 (576)	0.83 (0.24)	25 857	−4865	−5286; −4443 −20 739; 11 166	0.15 (75/500)	0.171; 0.181; 0.193
20 396 (6565)	0.77 (0.37)	26 646	−5087	−5833; −4342 −22 572; 27 817	0.24 (122/500)	0.175; 0.186; 0.194
Scenarios in which Dorg n = 50						
21 724	0.50	43 567	−11 751	NA	NA	NA
17 199 (15 302)	1.22 (0.80)	14 130[†]	7144	5456; 8832 −24 662; 55 016	0.61 (252/414)	0.218; 0.243; 0.250
22 890 (191)	1.14 (0.06)	20 084	−95	−212; 21 −4286; 3862	0.50 (248/500)	0.001; 0.002; 0.002
Scenarios in which Dorg n = 500						
22 226	0.37	59 513	−14 757	NA	NA	NA
14 989 (17 143)	0.37 (0.26)	40 875	−7655	−9101; −6209 −24 369; 43 193	0.18 (89/499)	0.190; 0.203; 0.214
22 757 (199)	0.30 (0.07)	75 973	−16 766	−16 893; −16 639 −21 098; −12 322	0.00	0.015; 0.017; 0.018
Scenarios in which Dorg n = 10 000						
22 079	0.62	35 858	−9764	NA	NA	NA
22 076 (184)	0.60 (0.07)	36 783	−10 073	−10 204; −9942 −15 267; −6062	0.00	0.069; 0.071; 0.074

Figure 3. Stepwise approach to the use of synthetic data in health economic models, distinguishing activities performed and controlled by the data owner (or individuals with full data access) from activities performed and controlled by the health economic modeler. *If characteristics or outcomes are known, or thought, to vary substantially between subgroups of individuals (for example, in different treatment arms of nonrandomized studies) then generation and checking of synthetic data sets separately per subgroup should be considered in step 3 and 4. Similar to the process of performing multiple imputation in subgroups, generating synthetic data sets per subgroup is likely slightly less efficient than ignoring subgroups but also more robust in capturing potential interaction effects.[14]



pMSE indicates propensity mean squared error.

blind to the rules used to generate Dorg. This would better reflect real-world scenarios in which modelers lack access to the underlying data generation process.

Although we investigated the use of synthetic data as an alternative to original data in performing a model-based economic evaluation, there are alternative potential applications of synthetic data in cost-effectiveness analyses and alternative methods to preserve privacy, see Appendix 5 in Supplemental Materials found at https://doi.org/10.1016/j.jval.2025.06.007.

Conclusions

Synthetic data should not be used as complete replacement of original data to inform health economic models aiming to support reimbursement decisions. Quoting Nowok et al: "The original aim of producing synthetic data has been to provide publicly available data sets that can be used for inference in place of the actual data. However, such inferences will only be valid if the model used to construct the synthetic data is the true mechanism that has generated the observed data, which is very difficult, if at all possible, to achieve."[5] In line with this, Giuffrè and Shung[25] conclude "[…] while synthetic data possess the potential to revolutionize healthcare by offering improved research capabilities and cost-effective solutions, overcoming the challenges related to biased information, data quality concerns, and potential privacy risks are of paramount importance." This is true for clinical studies, as well as economic evaluations. For now, it would therefore be safest and valuable to use synthetic data only for exploration and potentially model building, while validating results with the original data (or asking the data owner to do so) before using them to inform policy decisions.

Author Disclosures

Author disclosure forms can be accessed below in the Supplemental Material section.

Supplemental Material

Supplementary data associated with this article can be found in the online version at https://doi.org/10.1016/j.jval.2025.06.007.

Article and Author Information

Accepted for Publication: June 3, 2025

Published Online: September 19, 2025

doi: https://doi.org/10.1016/j.jval.2025.06.007

Author Affiliations: Institute for Health Systems Science, Multi-Actor Systems Department, Delft University of Technology, Delft, The Netherlands (van der Linden); Health Technology and Services Research, Faculty of Behavioural, Management and Social Sciences, TechMed Centre, University of Twente, Enschede, The Netherlands (Pouwels, Koffijberg); Department of Public Health, Health Services Research, and Health Technology Assessment, UMIT TIROL-University for Health Sciences and Technology, Hall in Tirol, Austria (Jahn, Siebert); Division of Health Technology Assessment, ONCOTYROL-Center for Personalized Cancer Medicine, Innsbruck, Austria (Jahn, Siebert); Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA, USA (Siebert); Department of Health Policy and Management, Center for Health Decision Science, Harvard T.H. Chan School of Public Health, Boston, MA, USA (Siebert); Department of Radiology, Institute for Technology Assessment, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA (Siebert).

Correspondence: Naomi van der Linden, PhD, Delft University of Technology, Faculty of Technology, Policy and Management, PO Box 5015 2600 GA Delft, The Netherlands. Email: n.vanderlinden@tudelft.nl

Authorship Confirmation: All authors certify that they meet the ICMJE criteria for authorship.

Funding/Support: The authors received no financial support for this research.

Acknowledgment: The authors acknowledge Professor Dr Martijn Warnier and Dr Wilbert van den Hout for their valuable feedback on this research.

REFERENCES

1. Guyot P, Ades A, Ouwens MJ, Welton NJ. Enhanced secondary analysis of survival data: reconstructing the data from published Kaplan-Meier survival curves. *BMC Med Res Methodol*. 2012;12:1–13.
2. Smith RA, Schneider PP, Mohammed W. Living HTA: automating health economic evaluation with R. *Wellcome Open Res*. 2022;7:194.
3. Rubin DB. Statistical disclosure limitation. *J Off Stat*. 1993;9:461–468.
4. Arnold C, Neunhoeffer M. Really Useful Synthetic Data–A Framework to Evaluate the Quality of Differentially Private Synthetic Data. ArXiv. https://arxiv.org/abs/2004.07740. Accessed July 11, 2025.
5. Nowok B, Raab GM, Dibben C. synthpop: bespoke creation of synthetic data in R. *J Stat Softw*. 2016;74(11):1–26.
6. R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. https://www.R-project.org/; Published 2023. Accessed July 11, 2025.
7. Endres M, Venugopal AM. *Synthetic Data Generation: A Comparative Study-Proceedings of the 26th International Database Engineered Applications Symposium*. 2025.
8. Pezoulas VC, Zaridis DI, Mylona E, et al. Synthetic data generation methods in healthcare: a review on open-source tools and methods. *Comp Struct Biotechnol J*. 2024;23:2892–2910.
9. Kaabachi B, Despraz J, Meurers T, et al. Can we trust synthetic data in medicine? A scoping review of privacy and utility metrics. MedRxiv. https://www.medrxiv.org/content/10.1101/2023.11.28.23299124v1. Accessed July 11, 2025.
10. National Cancer Institute. Synthetic data generator (SYNDATA). https://computational.cancer.gov/software/synthetic-data-generator; Published 2024. Accessed July 11, 2024.
11. Filipović-Pierucci A, Zarca K. Durand-Zaleski I Markov Models for Health Economic Evaluations: the R Package Heemod. arXiv. https://arxiv.org/abs/1702.03252. Accessed July 7, 2025.
12. Zorginstituut Nederland. Guideline for economic evaluations in healthcare. https://english.zorginstituutnederland.nl/about-us/working-methods-and-procedures/guideline-for-economic-evaluations-in-healthcare; Published 2024. Accessed July 7, 2025.
13. Snoke J, Slavković A. *pMSE Mechanism: Differentially Private Synthetic Data with Maximal Distributional Similarity. SpringerLink*. 2025.
14. Sullivan TR, White IR, Salter AB, Ryan P, Lee KJ. Should multiple imputation be the method of choice for handling missing data in randomized trials? *Stat Methods Med Res*. 2018;27(9):2610–2626.
15. Gonzales A, Guruswamy G, Smith SR. Synthetic data in health care: a narrative review. *PLOS Digit Health*. 2023;2(1):e0000082.
16. Gresham G, Meinert JL, Gresham AG, Meinert CL. Assessment of trends in the design, accrual, and completion of trials registered in Clinicaltrials.gov by sponsor type, 2000-2019. *JAMA Netw Open*. 2020;3(8):e2014682.
17. Fenwick E, Steuten L, Knies S, et al. Value of information analysis for research decisions—an introduction: report 1 of the ISPOR Value of Information Analysis Emerging Good Practices Task Force. *Value Health*. 2020;23(2):139–150.
18. Rothery C, Strong M, Koffijberg HE, et al. Value of information analytical methods: report 2 of the ISPOR value of information analysis emerging good practices task force. *Value Health*. 2020;23(3):277–286.
19. Brand J, van Buuren S, le Cessie S, van den Hout W. Combining multiple imputation and bootstrap in the analysis of cost-effectiveness trial data. *Stat Med*. 2019;38(2):210–220.
20. Li CX, Zivich PN. Invited commentary: mixing multiple imputation and bootstrapping for variance estimation. *Am J Epidemiol*. 2024;193(10):1477–1481.
21. Goncalves 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):1–40.
22. Tzikas R. How realistic is my synthetic data? A qualitative approach. Utrecht University. https://studenttheses.uu.nl/bitstream/handle/20.500.12932/43132/Thesis%20-%20Rigas%20Tzikas.pdf?sequence=1; Published 2022. Accessed July 7, 2025.
23. Raab GM, Nowok B, Assessing Dibben C. Visualizing and Improving the Utility of Synthetic Data. arXiv. https://arxiv.org/abs/2109.12717; Published 2021. Accessed July 7, 2025.
24. Santangelo G, Nicora G, Bellazzi R, Dagliati A. How good is your synthetic data? SynthRO, a dashboard to evaluate and benchmark synthetic tabular data. *BMC Med Inform Decis Mak*. 2025;25(1):89.
25. Giuffrè M, Shung DL. Harnessing the power of synthetic data in healthcare: innovation, application, and privacy. *NPJ Digit Med*. 2023;6(1):186.