



Gewirtz-O'Reilly, Flynn (2024) *Generation and assessment of useful and privacy preserving synthetic datasets*. PhD thesis.

<https://theses.gla.ac.uk/84327/>

Copyright and moral rights for this work are retained by the author

A copy can be downloaded for personal non-commercial research or study, without prior permission or charge

This work cannot be reproduced or quoted extensively from without first obtaining permission in writing from the author

The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the author

When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given

Enlighten: Theses

<https://theses.gla.ac.uk/>
research-enlighten@glasgow.ac.uk

Generation and assessment of useful and privacy preserving synthetic datasets

Flynn Gewirtz-O'Reilly

Submitted in fulfilment of the requirements for the
Degree of Doctor of Philosophy

School of Mathematics & Statistics
College of Science and Engineering
University of Glasgow



University
of Glasgow

January 2024

Abstract

Synthetic datasets are gaining traction as a potential solution for allowing access to sensitive data while protecting the privacy of individuals. However, the assessment of both the utility and disclosure risk of synthetic data is still an open question for which there is little consensus. Solutions that are theoretically good have been proposed but these are not currently feasible for most use cases. Meanwhile, most practicable disclosure risk assessments are ad hoc, unsuitable for more than a few sensitive variables, and only consider a narrow range of risk scenarios. For greater uptake of synthetic data it is important to establish a standard for its assessment.

In this thesis, we evaluate methods for the assessment of synthetic data and identify several clear issues in the literature. We develop a practical framework for the quantitative assessment of disclosure risk for synthetic data. Hierarchical regression models are used for the evaluation and comparison of disclosure risk for multiple sensitive variables, synthetic datasets and intruder assumptions simultaneously. We demonstrate our methods on two example datasets. A small dataset containing less than 1000 samples and 9 variables, and a larger dataset that contains over 50000 samples and 40 variables. We find that the method of prediction has a significantly larger effect on attribute disclosure risk than the synthetic data generation method.

Contents

Abstract	i
List of Acronyms	ix
Acknowledgements	x
Declaration	xi
1 Introduction	1
1.1 Motivation	1
1.2 Organisation of the thesis	3
2 Statistical and machine learning methods	4
2.1 Regression models	4
2.1.1 Generalised linear models	4
2.1.2 Bayesian inference	7
2.1.3 Maximum likelihood estimation	11
2.1.4 Generalised additive mixed models	14
2.2 Decision tree models	15
2.2.1 Random forests	17
3 Literature review: information disclosure	19
3.1 Types of information disclosure	20
3.1.1 Risk factors for information disclosure	21
3.2 Protecting confidentiality: non-synthetic methods	22
3.2.1 Method 1: removal of personal identifiers	22
3.2.2 Method 2: masking techniques	24
3.3 Standards for ensuring anonymity	25
3.3.1 k -anonymity	25
3.3.2 ℓ -diversity	27
3.3.3 t -closeness	28
3.3.4 Differential privacy	29

3.4	Protecting confidentiality: synthetic data	30
4	Literature review: synthesising data	31
4.1	Frameworks for synthetic data	31
4.1.1	Rubin’s framework: fully synthetic data	31
4.1.2	Little’s framework: partially synthetic data	33
4.1.3	Addressing a common misconception with synthetic data terminology	34
4.2	Variance estimation for synthetic data	34
4.2.1	Multiple synthesis	34
4.2.2	Variance estimators for synthetic data	35
4.2.3	Synthesis of missing data	37
4.3	Joint modelling	39
4.4	Sequential modelling	44
4.4.1	Order of synthesis	45
4.4.2	Synthesising with regression models	47
4.4.3	Synthesising with decision tree models	50
4.4.4	Examples of sequential modelling	53
5	Literature Review: assessing utility	56
5.1	Measures of distributional similarity	57
5.2	Model based discriminators	58
5.3	Plots of data distributions	61
5.4	Sample statistics	63
5.4.1	Sample statistics: methods of comparison	64
5.5	Human feedback	65
5.6	Task performance	65
6	Literature Review: assessing the disclosure risk	68
6.1	Identity disclosures	68
6.2	Membership disclosure	69
6.3	Attribute disclosure	70
6.3.1	A Bayesian approach	72
6.3.2	An empirical approach	73
6.4	Disclosure risk of outliers	77
6.4.1	Local outlier factor	78
6.5	Summary of the literature review	81
7	Methods for assessing synthetic data	83
7.1	Assessing the utility of synthetic data	83
7.1.1	Plots	84

7.1.2	Task performance	84
7.2	Assessing the disclosure risk of synthetic data	84
7.2.1	The prior knowledge of the intruder	85
7.2.2	Intruder prediction method(s)	86
7.2.3	Quantifying the risk of disclosure	88
7.2.4	Evaluating acceptability of disclosure risk	89
8	Generating and assessing synthetic Pima data	92
8.1	Diabetes and the Pima dataset	92
8.2	Methods	94
8.2.1	Synthesising Pima data	94
8.2.2	Assessing the utility of synthetic Pima data	97
8.2.3	Assessing disclosures risks of synthesised Pima datasets	102
8.3	Results	114
8.3.1	Results of utility assessments of synthetic Pima data	114
8.3.2	Results of disclosure risk assessments of synthetic Pima data	122
8.4	Conclusions	137
9	Large synthetic microdata	139
9.1	Exploratory analysis	139
9.2	Methods	146
9.2.1	Synthesising Diabetes-130 data	146
9.2.2	Baseline datasets	149
9.2.3	Assessing the utility of synthetic Diabetes-130 data	150
9.2.4	Assessing the disclosure risk of synthetic Diabetes-130 data	154
9.3	Results	159
9.3.1	Results of utility assessments of synthetic Diabetes-130 data	159
9.3.2	Results of disclosure risk assessment of synthetic Diabetes-130 data	163
9.4	Conclusions	174
10	Discussion	176
10.1	Discussion of Literature Review	176
10.2	Our Framework	176
10.3	Our Utility Assessment	177
10.4	Our Disclosure Risk Assessment	177
10.5	Limitations and future work	178
A	Additional tables	180

List of Tables

3.1	Sample of observations from Pima data.	25
8.1	The percentage of subsets of attribute disclosure scores equal to zero. . . .	105
8.2	Differences between area under precision-recall curve of inliers and outliers for both intruder prior knowledge scenarios.	124
8.3	Subset of the closest pairs of observations from training and synthetic datasets.	125
8.4	The closest pairs of observations from test and synthetic datasets.	125
8.5	ELPD difference for linear regression models.	127
8.6	ELPD difference for Gaussian, log normal, and gamma distributional models.	128
8.7	ELPD difference of full and simplified Gaussian distributional models. . . .	128
8.8	ELPD difference of full and simplified Gaussian hierarchical distributional models.	129
9.1	How HbA1c and change variable was defined.	152
9.2	95% confidence interval overlaps for various sample statistics in 130 hospitals data.	161
9.3	Ratio of estimates for various sample statistics in 130 hospitals data. . . .	162
9.4	Estimates of fixed and random effects from fitted model on the log-odds scale (Equation (9.4)).	164
A.1	The closest pairs of training (original) and synthetic Pima observations for each data synthesis method when matching on all variables.	180
A.2	List of variables from Diabetes-130 dataset and their descriptions (Strack et al., 2014).	182
A.3	List of groupings made to levels of categorical variables in Diabetes-130 dataset.	184

List of Figures

2.1	Plots showing the fits of polynomial regression models on 9 observations of data.	13
2.2	Example of a decision tree with four splits that partition the co-domain into five leaves.	16
4.1	Generating synthetic data within the partially synthetic framework.	33
8.1	Prevalence of obesity in the UK for different age groups and sexes.	95
8.2	Synthesis order for Pima Indians dataset.	96
8.3	Histogram of attribute disclosure scores.	104
8.4	Histograms of y_{jrst} for each combination of target variable and disclosure method.	106
8.5	Distribution of y_{ijst} for each target variable and disclosure method.	107
8.6	Histogram of simplified attribute disclosure scores.	108
8.7	$\log(\text{lof}_i)$ against y_{ijst} for a small sample of the attribute disclosure scores.	112
8.8	Univariate variable distributions for 50 replications of synthetic Pima data.	115
8.9	Synthetic data quantiles for pregnancies and diabetes pedigree function.	116
8.10	Distribution of synthesised glucose variables.	116
8.11	Pairwise correlation difference between original and synthetic Pima data.	117
8.12	DWP scores for synthetic and 5-anonymised Pima data.	118
8.13	Boxplots of $\text{pMSE}_{\text{ratio}}$ scores for synthetic Pima data.	119
8.14	The number of synthetic Pima replications that each inference model was the best fit.	120
8.15	90% credible interval overlap for Pima inference models averaged over all coefficients.	120
8.16	90% credible interval overlap for Pima inference models averaged over all replications.	121
8.17	The number of replications of synthetic data for which the hypothesis test result aligns with the original data.	122
8.18	Precision-recall curves for changing intruder prior knowledge.	123

8.19	Precision-recall curves for changing intruder prior knowledge and training subsets.	123
8.20	PPD of area under precision-recall curves for changing intruder prior knowledge and training subsets.	124
8.21	Posterior samples from linear regression grouped by attribute and attribute disclosure method.	126
8.22	Posterior distributions $\tilde{y}_{jrst} y_{jrst}$ of distributional models.	127
8.23	Posterior distribution $\tilde{y}_{jrst} y_{jrst}$ from normal distributional model grouped by target variable and disclosure method.	128
8.24	Posterior distribution of Gaussian hierarchical distributional model grouped by target variable and disclosure method.	130
8.25	Posterior distribution of Gaussian hierarchical distributional model grouped by subject.	130
8.26	Posterior distributions of attribute prediction method coefficient, for each target variable and disclosure method.	133
8.27	Posterior distributions of synthesis method coefficient for Gaussian distributional and hierarchical distributional regression models.	134
8.28	90% HDPIs of attribute disclosure score from Gaussian distributional regression and hierarchical distributional regression models.	135
8.29	Posterior density for the coefficient of log local outlier factor.	136
8.30	Posterior predictive distribution of log attribute disclosure score given log local outlier factor.	136
9.1	Total number of visits for each patient in 130 hospitals dataset.	140
9.2	Distributions for count variables in the 130 hospitals data.	141
9.3	Distributions of medicine variables in the 130 hospitals data.	142
9.4	ICD-9 code for primary diagnosis of each visit in the 130 hospitals data.	143
9.5	Speciality of admitting physician for each visit in the 130 hospitals data.	144
9.6	Distributions of categorical variables across all visits in the 130 hospitals data.	145
9.7	Results from the initial exploration of regression models for synthesising count variables.	147
9.8	Proportion of correct attribute predictions for different attributes in the 130 hospitals training partition.	155
9.9	Probability of correct attribute predictions for small sample of subjects in the 130 hospitals training partition.	156
9.10	Paired differences in the probability of correct attribute predictions for all pairs of attribute prediction methods, and all pairs of synthesis methods and baseline datasets.	156

9.11 Comparing the proportions of levels of categorical variables in synthetic and baseline datasets.	159
9.12 Comparing the proportions of levels of diagnosis code variables in synthetic and baseline datasets.	160
9.13 Comparing the proportions of levels of medicine variables in synthetic and baseline datasets.	160
9.14 Feature prediction scores for categorical variables from the 130 hospitals dataset.	165
9.15 Comparing the distributions of numeric variables in synthetic and baseline datasets.	166
9.16 Feature prediction scores for numeric variables from the 130 hospitals dataset.	166
9.17 Pairwise correlation differences for synthesis replications.	167
9.18 Counts of structural zero observations for synthetic replications.	167
9.19 pMSE score ratios for synthetic replications of 130 hospitals data.	167
9.20 Sample statistics for synthetic and baseline 130 hospitals datasets.	168
9.21 Percentage of re-admissions for different age brackets in synthetic and baseline 130 hospitals datasets.	168
9.22 90% credible intervals for coefficients of Bayesian regression model with Gaussian priors.	169
9.23 90% credible interval overlap for coefficients of Bayesian regression model. .	170
9.24 90% credible intervals for coefficients of Bayesian regression model with horseshoe priors.	171
9.25 Observed and predicted values for each attribute in validation data.	172
9.26 Observed and predicted values for each attribute prediction model and synthesis model in validation data.	172
9.27 95% prediction intervals from the fitted model conditioned on the uncertainty of fixed effects.	173

List of Acronyms

BMI	body mass index
CART	classification and regression trees
DCAP	differential correct attribution probability
DPMPM	Dirichlet process mixture of products of multinomials
DWP	dimension-wise prediction
ELPD	expected log predictive density
GAMM	generalised additive mixed model
GAN	generative adversarial network
GMM	Gaussian mixture model
HDPI	highest density posterior interval
HIPAA	Health Insurance Portability and Accountability Act
LOF	local outlier factor
MCMC	markov chain monte carlo
PCD	pairwise correlation difference
PPD	posterior predictive distribution
SDC	statistical disclosure control
SUDA	special unique detection algorithm
SVM	support vector machine
VAE	variational autoencoder

Acknowledgements

First, I would like to thank my supervisors, Dr Surajit Ray, Dr Jeremy Voisey, and Dr Aneta Lisowska. Thank you for your patience and guidance.

I would also like to thank the EPSRC and Canon Medical for their generous funding of this project and for the opportunities provided to me over the course of my PhD. I am grateful to the faculty and staff at the University of Glasgow and the School of Mathematics and Statistics for your support through the challenges I have faced throughout my PhD.

I would especially like to thank my family. To Freda, Eva, and Róisín, thank you for your endless affection, presence, and support. From the beginning, all three of my sisters have encouraged lots of lengthy “discussions” on all topics, big and small; I am forever grateful to them for encouraging inquisitiveness. To my parents, Sharon and Desmond, thank you for your love, generosity, encouragement, and care. You have always been so very vital in the formation of my values and ethics. To my in-laws, Mark (a.k.a. Joe) and Meris, thank you for going above and beyond. Your support has been vital to the completion of this thesis. To my daughter, Isla, thank you for arriving halfway through this PhD and being my biggest inspiration to finish. You have never ceased to put a smile on my face. Thank you.

My greatest thanks go to my wife. Kellan, we started our PhDs together and you have fully supported me since day one. Thank you for your unwavering belief in me and your ability to always see the positive. I would not be where I am today without you.

Declaration

I declare that, except where explicitly stated otherwise in the text, this thesis is the result of my own original work and has not been submitted for any other degree at the University of Glasgow or any other institution.

Chapter 1

Introduction

Statistics and machine learning are two areas of research undergoing rapid growth and innovation. Two factors primarily drive this growth. The first is the increased gathering and accessibility of large sets of individual-level data. The second is the development of modelling approaches that use advances in computing power. These areas of research have the potential to improve massively the lives of many. In fact, there are already numerous examples of the benefits of these new technologies, such as providing decision support to doctors through the automatic detection of time-critical health issues from clinical notes (Cutforth et al., 2023), triage of COVID-19 patients (Ray et al., 2022), and segmentation of tumours (W. Zhang & Ray, 2023) to name a few. Despite these benefits, there is growing recognition of and concern about the gathering, storage, and (willing or unwilling) dissemination of personal data.

1.1 Motivation

Legislation has been introduced to protect the privacy rights of individuals (see, e.g., California Consumer Privacy Act (2018), California Privacy Rights Act (2020), Data Protection Act 2018 (2018), Health Insurance Portability and Accountability Act (1996), and The European Union (2016)). These protections include providing individuals with the rights of ownership over their data and requiring organisations to protect sensitive data adequately. Organisations that breach these laws can be subject to hefty penalties (Information Commissioner’s Office, 2018).

One example, the Health Insurance Portability and Accountability Act (HIPAA) focuses on protecting personal medical information. For data to be deemed HIPAA compliant, the data has to go through a de-identification procedure where names, identifying numbers, and dates (except years) are removed from the data, geographical information must be generalised to areas with greater than 20,000 people, and the age of any person older than 89 years must be truncated (OCR, 2012). There is agreement within the

statistics and machine learning communities that these de-identification methods reduce the utility of the data (Drechsler & Reiter, 2010; Purdam & Elliot, 2007; Winkler, 2007). Even more concerning is the evidence that the de-identification procedures do not adequately protect from disclosure (Sweeney et al., 2017).

Given the need to protect personal data, and particularly personal medical data, and the need to use this data in research, there is a significant privacy-utility trade-off with de-identified data. In comes synthetic data. *Synthetic data* is artificial data generated from a model that approximates the distribution of real data and was initially proposed by Rubin (1993) and Little (1993). It is often touted as a solution for releasing personal data that has a more favourable privacy-utility trade-off. Synthesising data (and, therefore, generative models) has increased in popularity in the statistics and machine learning communities over the past several years. Numerous organisations use synthetic data, including the United States Census Bureau and the Institute for Employment Research (IAB) (Abowd et al., 2006; Benedetto et al., 2017; Benedetto et al., 2013; Dennett, 2017; Dennett et al., 2016; Drechsler, Dundler, et al., 2008; Kinney et al., 2014; Kinney et al., 2011; Nowok et al., 2017; U.S. Census Bureau, 2013, 2018). Synthetic data allows researchers to access data that would have been unavailable previously.

Due to the novelty of synthesising data and the implementation of generative models, there is a lack of consensus in the literature (both statistics and machine learning) about assessing synthetic data and generative models. In terms of assessing synthetic data, there are two aspects of interest to this thesis. The first aspect is measuring the utility of synthetic data. The second is measuring the disclosure risk of synthetic data. A consensus on these two aspects must be reached to enable more effective comparison of synthetic data generation methods, which would help with the development of new data generation methods. This thesis focuses on exploring these two aspects. Specifically, this thesis does the following:

- Answers the question: “*How should synthetic data’s utility and disclosure risk be assessed?*”,
- provides a comprehensive review of the synthetic data literature. In particular, our review of the methods for assessing synthetic data does not yet exist in the literature,
- investigates the impact that the choices of synthetic data evaluation methods have on the results of the evaluations,
- develops and evaluates a framework for assessing the utility and disclosure risk of synthetic data (or data produced by other statistical disclosure control (SDC) methods), and
- introduces methods for assessing attribute disclosure risk.

1.2 Organisation of the thesis

Chapter 2 contains an overview of methods from the statistical and machine learning literature utilised throughout the remainder of this thesis. Our novel review of the synthetic data literature is separated into four chapters. Specifically, in Chapter 3, we review the literature related to information disclosure and SDC methods, including synthetic data. In Chapter 4, we summarise the theory underpinning synthetic data and review the literature on synthetic data generation. We review the literature relating to assessing utility and disclosure risk of synthetic data in Chapters 5 and 6, respectively. In Chapter 7, a framework for assessing the privacy and utility of synthetic data is introduced, and we describe methods for the assessment of synthetic data assessment developed over the course of this research. In Chapter 8, we demonstrate the framework and methods that were described using the Pima diabetes data as a simple example. Chapter 9 demonstrates the framework and methodology but with the more complex 130 Hospitals Diabetes dataset. Finally, in Chapter 10, we summarise the key findings of the thesis and discuss possible avenues for future work.

Chapter 2

Statistical and machine learning methods

This chapter summarises methods from the wider statistical and machine learning (ML) literature. These methods are required background knowledge for the synthetic data generation models and synthetic data assessment methods that we implement through this thesis. However, most content in this chapter will be familiar to those with a background in statistics or machine learning. Section 2.1 describes regression models in both a Bayesian and a frequentist setting. Then, in Section 2.2, we describe classification and regression trees (CART), and random forests.

2.1 Regression models

2.1.1 Generalised linear models

Consider the linear regression model of the form:

$$\mathbb{E}(Y_i) = \mu_i = \mathbf{x}_i^T \boldsymbol{\beta}, \quad (2.1)$$

with Gaussian probability density function (PDF)

$$p(y_i | \mu_i, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2} (y_i - \mu_i)^2\right), \quad (2.2)$$

where

- $Y_i \sim \mathcal{N}(\mu_i, \sigma^2)$ is the i^{th} observation of the response variable,
- $\mathbf{x}_i = (1, x_{i1}, \dots, x_{ip})^T$ is the i^{th} row of the design matrix of known covariates \mathbf{X} , and
- $\boldsymbol{\beta} = (\beta_0, \dots, \beta_p)^T$ is the vector of unknown coefficients.

We can generalise this for response variables with non-Gaussian distributions.

Let Y be a random variable with a probability distribution that belongs to the exponential family, a flexible class of distributions that includes many of the most commonly used probability distributions. The PDF of any distribution in the exponential family is written as

$$p(y|\theta, \phi) = \exp\left(\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right), \quad (2.3)$$

where θ is a location parameter called the canonical parameter, and ϕ is a dispersion parameter. The values of these parameters depend on the probability distribution of Y . The functions $a(\phi)$, $b(\theta)$, and $c(y, \phi)$ are known functions that also depend on the probability distribution of Y .

The mean and variance of probability distributions in the exponential family are expressed in terms of $a(\phi)$ and $b(\theta)$:

$$\begin{aligned} \mathbb{E}[Y] &= \mu = b'(\theta), \\ \text{Var}[Y] &= b''(\theta)a(\phi). \end{aligned}$$

Consider the generalised linear regression model of the form

$$g(\mathbb{E}(Y_i)) = g(\mu_i) = \mathbf{x}_i^T \boldsymbol{\beta},$$

where

- Y_i is an independently distributed response variable, that belongs to a distribution in the exponential family given in Equation (2.3).
- g is a monotonic link function that describes the relationship between μ_i and $\mathbf{x}_i^T \boldsymbol{\beta}$.
- $\mathbf{x}_i = (1, x_{i1}, \dots, x_{ip})^T$ is the i^{th} row of the design matrix of known covariates \mathbf{X} .
- $\boldsymbol{\beta} = (\beta_0, \dots, \beta_p)^T$ is the vector of unknown parameters.

The link g is any monotone, continuous and differentiable function. One family of convenient link functions are *canonical link functions*. These are link functions where $\theta = g(\mu)$.

Generalised linear models for binary response variables

The standard model to represent an unordered categorical response variable is a logistic regression. To start, consider the case where the response variable is binary. In this case, the response is assumed to follow a Bernoulli distribution:

$$Y_i \sim \text{Bernoulli}(p_i),$$

where

- $Y_i \in \{0, 1\}$,
- $P(Y_i = 1) = p_i$ is the probability of “success”,
- $P(Y_i = 0) = 1 - p_i$ is the probability of “failure”, and
- $\mathbb{E}(Y_i) = p_i$.

The Bernoulli PDF is written as

$$p(y_i|p_i) = \exp\left(y_i \log\left(\frac{p_i}{1-p_i}\right) - \log\left(\frac{1}{1-p_i}\right)\right). \quad (2.4)$$

The standard choice of link function is its canonical link function, the *logit* written as

$$g(p_i) = \log\left(\frac{p_i}{1-p_i}\right) = \mathbf{x}_i^T \boldsymbol{\beta}. \quad (2.5)$$

The inverse of the logit function is the *logistic* function,

$$p_i = g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta}) = \frac{1}{1 + \exp(-\mathbf{x}_i^T \boldsymbol{\beta})}. \quad (2.6)$$

Equation (2.4) can be rewritten in terms of Equation (2.6) such that

$$p(y_i|\mathbf{x}_i, \boldsymbol{\beta}) = g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})^{y_i} (1 - g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta}))^{1-y_i}. \quad (2.7)$$

Generalised linear models for nominal response variables

The logistic regression model is extended for an unordered categorical response variable with K categories as follows

$$\mathbf{Y}_i \sim \text{Categorical}(\mathbf{p}_i),$$

where

- $\mathbf{Y}_i = (Y_{i1}, \dots, Y_{iK})$ for $Y_{ik} \in \{0, 1\}$ and $\sum_{k=1}^K Y_{ik} = 1$,
- $P(\mathbf{Y}_{ik} = 1) = p_{ik}$, and
- $\mathbf{p}_i = (p_{i1}, \dots, p_{iK})$ for $\sum_{k=1}^K p_{ik} = 1$.

The PDF of the categorical distribution is written as

$$p(\mathbf{Y}_i = \mathbf{y}_i|\mathbf{p}_i) = \exp\left\{\sum_{k=1}^{K-1} y_{ik} \log\left(\frac{p_{ik}}{1 - \sum_{k=1}^{K-1} p_{ik}}\right) - \log\left(\frac{1}{1 - \sum_{k=1}^{K-1} p_{ik}}\right)\right\}.$$

The standard choice of link function is the multinomial *logit* function,

$$g(\mathbf{p}_i)_k = \log \left(\frac{p_{ik}}{1 - \sum_{s=1}^{K-1} p_{is}} \right) = \mathbf{x}_i^T \boldsymbol{\beta}_k,$$

where

- $\mathbf{x}_i = (1, x_{i1}, \dots, x_{ip})^T$ is the i^{th} row of the design matrix, and
- $\boldsymbol{\beta}_k = (\beta_{k1}, \dots, \beta_{kp})$ is a column in the $(p \times K)$ matrix of coefficients $\boldsymbol{\beta}$.

We enforce the constraint that the predicted probabilities for the K categories must sum to one by setting the coefficients for the arbitrary K^{th} level to be zero,

$$\boldsymbol{\beta}_K = \mathbf{0}.$$

The inverse of the multinomial logit is the *softmax* function,

$$p_{ik} = \frac{\exp(\mathbf{x}_i^T \boldsymbol{\beta}_k)}{\sum_{t=1}^K \exp(\mathbf{x}_i^T \boldsymbol{\beta}_t)}.$$

There are two schools of thought commonly used to make inferences on the unknown parameters of these regression models: Bayesian inference and maximum likelihood estimation (MLE).

2.1.2 Bayesian inference

With *Bayesian inference*, prior distributions and a likelihood function are specified. The prior distributions describe prior knowledge of the probability distribution for each unknown parameter, and the likelihood describes the distribution of the data given the parameters. Bayes' rule is used to update the prior with the data according to the likelihood. This results in a posterior distribution for the unknown parameters, which describes the probability distribution of each unknown parameter after the data has been considered.

More formally, these Bayesian updates can be written as

$$\begin{aligned} P(\boldsymbol{\theta}, \phi | \mathbf{y}, \tau, \gamma) &= \frac{P(\mathbf{y} | \boldsymbol{\theta}, \phi, \tau, \gamma) P(\boldsymbol{\theta} | \tau) P(\phi | \gamma)}{P(\mathbf{y} | \tau, \gamma)}, \\ &\propto P(\mathbf{y} | \boldsymbol{\theta}, \phi, \tau, \gamma) P(\boldsymbol{\theta} | \tau) P(\phi | \gamma), \end{aligned} \quad (2.8)$$

where

- $P(\mathbf{y} | \boldsymbol{\theta}, \phi, \tau, \gamma) = \prod_{i=1}^n P(y_i | \boldsymbol{\theta}, \phi)$ is the likelihood,
- $\boldsymbol{\theta} \sim p(\tau)$ and $\phi \sim p(\gamma)$ are the prior distributions for the model parameters, and

- τ and γ are the prior distribution hyperparameters.

The marginal distribution of the data,

$$P(\mathbf{y}|\tau, \gamma) = \iint P(\mathbf{y}|\boldsymbol{\theta}, \phi, \tau, \gamma)P(\boldsymbol{\theta}|\tau)P(\phi|\gamma) d\boldsymbol{\theta} d\phi,$$

is a constant that normalises the posterior distribution. In practice, the marginal distribution is intractable. As such, Bayesian methods focus on estimating the posterior distribution without directly computing the marginal distribution of the data.

Today, the most common approach is using *Markov chain Monte Carlo* (MCMC) methods to efficiently sample from the posterior distribution. The MCMC family of methods includes Gibbs sampling, the Metropolis algorithm, the Metropolis-Hasting’s algorithm, Hamiltonian Monte Carlo (HMC) and the No-U Turn Sampler (NUTS). This thesis uses NUTS. It is a variant of Hamiltonian Monte Carlo that uses a “momentum” variable to efficiently sample from the posterior distribution by taking large jumps in low-density posterior regions and smaller jumps in high-density posterior regions (Hoffman & Gelman, 2011). NUTS is implemented in the `Stan` programming language (Carpenter et al., 2017) and is interfaced with R through the packages `RStan` (Stan Development Team, 2024), `rstanarm` (Stan Development Team, 2018), and `brms` (Bürkner, 2017). These packages provide high-level interfaces for fitting a wide array of regression models using Stan. For more information on the HMC and MCMC methods, see Gelman et al. (2014, Chapters 11 & 12).

Expected log posterior density

One method for evaluating the fit of Bayesian models is leave-one-out (LOO) expected log predictive density (ELPD)

$$\text{elpd}_{\text{loo}}(\mathbf{y}) = \sum_{i=1}^n \log p(y_i|y_{-i}), \quad (2.9)$$

where $p(y_i|y_{-i})$ is the leave-one-out predictive density for the i^{th} data point. The leave-one-out predictive density is efficiently computed using Pareto smoothed importance sampling Vehtari et al. (2017).

Care needs to be taken if comparing a relatively large number of models that are fit to a comparatively small dataset. In such cases, model selection can induce bias and overfitting, so a model averaging or projection predictive approach would be more appropriate (Piiironen & Vehtari, 2017a).

When the response variable scale differs between models, ELPD must be transformed to a common scale by making a Jacobian adjustment before model fit can be compared.

For example, let y_i be a positive response variable, then under the change of variables $z_i = \log(y_i)$,

$$\begin{aligned} p(y_i|y_{-i}) &= p(z_i|z_{-i}) \left| \frac{\partial z_i}{\partial y_i} \right|, \\ &= p(z_i|z_{-i}) \cdot \frac{1}{y_i}. \end{aligned}$$

Therefore,

$$\begin{aligned} \text{elpd}_{\text{loo}}(\mathbf{y}) &= \sum_{i=1}^n \log \left(p(z_i|z_{-i}) \cdot \frac{1}{y_i} \right), \\ &= \sum_{i=1}^n (\log p(z_i|z_{-i}) - \log y_i), \\ &= \text{elpd}_{\text{loo}}(\mathbf{z}) - \sum_{i=1}^n \log y_i. \end{aligned}$$

Regularised horseshoe prior

One approach for specifying simpler Bayesian models is to use shrinkage priors such as the regularised horseshoe. The regularised horseshoe prior, as described in Piironen and Vehtari (2016, 2017b), is written as

$$\beta_j | \lambda_j, \tau, c \sim \mathcal{N}(0, \tau^2 \bar{\lambda}_j^2), \quad (2.10)$$

where

$$\bar{\lambda}_j^2 = \frac{c^2 \lambda_j^2}{c^2 + \tau^2 \lambda_j^2} \quad \text{and} \quad \lambda_j \sim \text{C}^+(0, 1).$$

Here, λ_j is a local hyperparameter that allows some coefficients to escape shrinkage, c is the standard deviation of β_j as the amount of shrinkage tends towards zero, and τ is a global hyperparameter that shrinks coefficients towards zero.

When β_j is close to zero

$$\tau^2 \lambda_j^2 \ll c^2.$$

As such, the prior tends towards the horseshoe prior (Carvalho et al., 2009)

$$\mathcal{N}(0, \tau^2 \lambda_j^2).$$

Conversely, when β_j is far from zero,

$$\tau^2 \lambda_j^2 \gg c^2,$$

and so the prior tends towards the spike and slab prior

$$\mathcal{N}(0, c^2).$$

According to Piironen and Vehtari (2017b), a sensible choice of prior for c is

$$c^2 \sim \text{Inv-Gamma}(\nu/2, \nu s^2/2),$$

which corresponds to a student- $t_\nu(0, s^2)$ slab for the coefficients that are far from zero.

Following Gelman (2006), τ is generally treated as a half-Cauchy distributed random variable. This is weakly informative and shrinks variances but allows for occasionally large values. A more informative variation of this prior, given by Piironen and Vehtari (2016, 2017b), is

$$\tau|\sigma \sim \text{C}^+(0, \tau_0^2), \quad (2.11)$$

where

$$\tau_0 = \frac{p_0}{D - p_0} \frac{\sigma}{\sqrt{n}}.$$

p_0 is a prior guess for the number of relevant parameters, and σ^2 is the variance of the observations in a Gaussian regression model or the “pseudo variance” for generalised linear models.

Bayesian regularising linear regression model

One interesting implementation of a regularising linear regression model is described by Gabry and Goodrich (2020). The method builds on the work of Lewandowski et al. (2009), who describe an approach for generating random correlation matrices from partial correlations. The standard linear regression model can be reparameterised using QR-composition to a form that allows for a prior to be placed on the R^2 value (Gabry & Goodrich, 2020).

Model 2.1 (Linear regression model with prior on R^2).

$$\mathbf{y}|R^2, \alpha, \omega \sim \mathcal{N}(\alpha + \omega s_y \sqrt{R^2(N_I - 1)} \mathbf{X}^T \mathbf{R}^{-1} \mathbf{u}, \omega^2 s_y^2 (1 - R^2)),$$

with a weakly informative prior on R^2 of a Beta distribution with a mode of p

$$R^2 \sim \text{Beta}(p/2, p/2),$$

a Jeffrey’s prior on ω

$$\omega \sim p(\omega) \propto 1/\omega,$$

and an improper uniform prior on the intercept

$$\alpha \sim \mathcal{U}(-\infty, \infty).$$

In contrast to the standard approach of specifying prior knowledge for the model parameters, specifying a prior for R^2 allows us to indicate how well we expect the model to fit the data. For cases where we do not have much prior knowledge about the model parameters, this can be a more sensible approach for specifying weakly informative priors. The R package `rstanarm` (Stan Development Team, 2018) contains an implementation of Model 2.1.

2.1.3 Maximum likelihood estimation

Recall that the likelihood

$$P(\mathbf{y}|\boldsymbol{\theta}, \phi) = \prod_{i=1}^n p(y_i|\boldsymbol{\theta}, \phi),$$

describes the distribution of the data given the parameters. Maximum likelihood estimation (MLE) aims to find the parameters that maximise the likelihood, or equivalently, the log-likelihood. The log-likelihood is generally more convenient to work with and, for distributions in the exponential family (Equation (2.3)), takes the form

$$\ell(\boldsymbol{\theta}; \mathbf{y}) = \sum_{i=1}^n \left(\frac{y_i \theta_i - b(\theta_i)}{a(\phi)} + c(y, \phi) \right), \quad (2.12)$$

$$= \frac{\mathbf{y}^T \boldsymbol{\theta} - b(\boldsymbol{\theta})}{a(\phi)} + c(\mathbf{y}, \phi), \quad (2.13)$$

where $g^{-1}(\mathbf{X}^T \boldsymbol{\beta}) = \boldsymbol{\theta} = (\theta_1, \dots, \theta_p)^T$ for $\theta_i = g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})$.

The parameters that maximise the log-likelihood

$$\hat{\boldsymbol{\theta}} = \arg \max_{\boldsymbol{\theta}} \{ \mathbf{y}^T \boldsymbol{\theta} - b(\boldsymbol{\theta}) \}, \quad (2.14)$$

are found by computing the roots of the first derivative

$$\frac{d\ell(\boldsymbol{\theta}; \mathbf{y})}{d\boldsymbol{\theta}} = \mathbf{0}.$$

The log-likelihood of the linear regression model

$$\ell(\boldsymbol{\beta}; \mathbf{y}) = \sum_{i=1}^n \left(\frac{(y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2}{2\sigma^2} - \log(\sigma\sqrt{2\pi}) \right), \quad (2.15)$$

is minimised by the least squared estimate

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}, \\ \hat{\sigma}^2 &= \frac{1}{n - p - 1} (\mathbf{y} - \mathbf{X} \hat{\boldsymbol{\beta}})^T (\mathbf{y} - \mathbf{X} \hat{\boldsymbol{\beta}}),\end{aligned}\tag{2.16}$$

where

$$\text{Var} [\hat{\boldsymbol{\beta}}] = (\mathbf{X}^T \mathbf{X})^{-1} \hat{\sigma}^2.$$

To avoid the numerical instability of calculating $(\mathbf{X}^T \mathbf{X})^{-1}$, the least squared solution is solved using QR decomposition (Hastie et al., 2009, p. 55).

For other generalised linear models, the maximum likelihood estimate is found using numerical optimisation methods. The standard optimisation method, implemented in R by the `glm` function, is iteratively reweighted least squares (Green, 1984; R Core Team, 2023). An alternative method is coordinate descent (Friedman et al., 2010), which is implemented in the R package `glmnet` and can efficiently fit regression models with very large numbers of observations or parameters.

Penalised maximum likelihood

The maximum likelihood estimate (Equation (2.14)) is the set of parameters that maximises the probability of the observed data, given the generalised linear model. However, maximising the likelihood does not always result in a model that fits unobserved samples of data well.

Bishop (2006, p. 7) uses a toy example to illustrate how maximising the likelihood for a model with a large number of parameters relative to the number of observations in the data can result in a fit that is not generalisable. Polynomial regression models are fit to a dataset with 9 observations using maximum likelihood

$$Y_i \sim \mathcal{N}\left(\sum_{j=0}^M \beta_j x_i^j, \sigma^2\right),$$

where $M = (1, \dots, 9)$.

The training set error decreases as the order of the polynomial increases, however, the fit of the 9th order polynomial model (Figure 2.1) shows clear over-fitting. Choosing a smaller model or gathering more data would be one solution to the problem, but it is not always feasible to gather more data. Another solution is to remove predictors from the model. This requires subjective decisions about how many parameters to choose and, in real-world examples of data with more variables than observations, removing predictors can remove helpful information.

In Section 2.1.2, we described one Bayesian approach to address over-fitting by placing

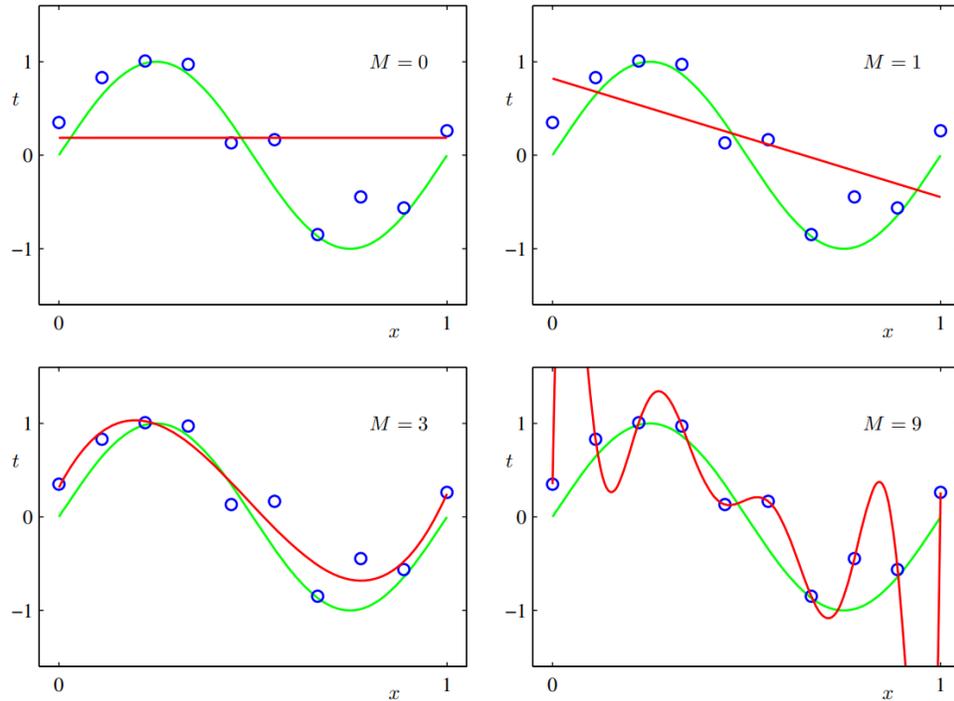


Figure 2.1: Plots showing the fits of polynomial regression models (in red) on 9 observations of data (Bishop, 2006, p. 7).

stronger priors on the model parameters. When fitting models with maximum likelihood estimation, a weight penalty can be added to regularise. This weight penalty would penalise the model as more parameters are added or as parameters grow larger.

For distributions in the exponential family (Equation (2.3)), the general form of the solution to penalised maximum likelihood is

$$\hat{\boldsymbol{\beta}} = \arg \max_{\boldsymbol{\beta}} \{ \mathbf{y}^T \boldsymbol{\theta} - b(\boldsymbol{\theta}) + \lambda P_{\alpha}(\boldsymbol{\beta}) \}, \quad (2.17)$$

where $\boldsymbol{\theta} = g^{-1}(\mathbf{X}^T \boldsymbol{\beta})$ is the canonical parameter,

$$P_{\alpha} = \sum_{j=1}^p \alpha \beta_j^2 + (1 - \alpha) |\beta_j|, \quad (2.18)$$

is the elastic net penalty (Zou & Hastie, 2005), $\lambda > 0$ is a hyper-parameter that controls the degree of regularisation, and $\alpha \in [0, 1]$ is a hyper-parameter controlling the mixing between the lasso penalty P_0 and the ridge penalty P_1 .

Note that the intercept coefficient β_0 is not penalised. Penalising the intercept coefficient would make the fitting procedure dependent on the origin of the response variable. Furthermore, penalised regression methods are not equivariant under the scaling of the inputs. As such, \mathbf{X} is generally centred and scaled before fitting (Hastie et al., 2009).

For several of the most common generalised linear models, including linear regression, Poisson regression, and binary and multi-class logistic regression, Equation (2.17) is solved using cyclical coordinate descent. In R, this is implemented with the package `glmnet` (Friedman et al., 2010).

2.1.4 Generalised additive mixed models

Generalised additive mixed models (GAMMs) are an extremely flexible and diverse class of models. GAMMs have been well described in many texts. For in-depth reading and complete proofs of the results discussed in this section, see Wood (2006, ch. 4 & 6) and Ruppert et al. (2003, ch. 8 & 11).

Consider the non-parametric additive model of the form

$$g(\mu_i) = f(\mathbf{x}_i),$$

where $\mu_i = \mathbb{E}(y_i)$, g is a monotonic link function and \mathbf{x}_i is the vector of covariates for the i^{th} observation. f is a smooth function with a single smoothing parameter

$$f(\mathbf{x}) = \sum_{j=1}^J b_j(\mathbf{x})\theta_j, \quad (2.19)$$

and associated wiggleness measure

$$J(f) = \boldsymbol{\theta}^T \mathbf{S} \boldsymbol{\theta},$$

where $\boldsymbol{\theta}$ is a vector of unknown coefficients and \mathbf{S} is a known positive semi-definite matrix¹.

We place a prior on the model coefficients

$$p(\boldsymbol{\theta}|\lambda) \propto \exp(-\lambda \boldsymbol{\theta}^T \mathbf{S} \boldsymbol{\theta}/2),$$

which reflects the belief that f is smooth. As \mathbf{S} is positive semi-definite, this prior is improper. However, the model can be reparameterised into the mixed model representation

$$g(\mu_i) = \tilde{\mathbf{x}}_i^T \boldsymbol{\beta} + \mathbf{Z}_i \mathbf{b}, \quad (2.20)$$

where $\tilde{\mathbf{x}}_i$ and \mathbf{Z}_i are the i^{th} rows of known design matrices for the fixed and random effects,

$$\mathbf{b} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}/\lambda),$$

is a vector of random effects. In this setting, $\boldsymbol{\beta}$ and λ are unknown parameters. These

¹A positive semi-definite matrix allows for some coefficients with zero wiggleness.

unknown parameters can be estimated with restricted maximum likelihood (REML). Alternatively, the posteriors of the unknown parameters can be modelled with Bayesian inference.

The smooth function described in Equation (2.19) can be represented with various spline basis functions. One such function is the penalised low-rank thin-plate splines (Wood, 2003). It is a lower dimension approximation of the multidimensional generalisation of smoothing splines. Low-rank thin-plate splines are an efficient low rank approximation of the higher dimensional thin-plate spline basis. They are constructed by the eigendecomposition of a matrix consisting of thin-plate splines with a knot at every unique observation of the data. The eigenvectors that correspond to the K largest eigenvalues then form the low-rank thin-plate spline basis.

The smoothing parameter will control the effective degrees of freedom. Therefore, the choice of K is arbitrary as long as K is not so large that computation becomes an issue and not so small that the model lacks the degrees of freedom to represent the data (Wood, 2006, p. 220).

2.2 Decision tree models

Decision trees are a flexible class of models. They have been successfully applied to a wide range of prediction problems, including the synthesis of both continuous and categorical data. For a more comprehensive introduction to these models, we direct the reader to James et al. (2013, pp.327–352).

Definition 2.1 (Decision Trees). Let $f(x)$ be a decision tree that partitions the data space R into T disjoint subsets $\{R_1, \dots, R_T\}$ through a series of recursive binary splits. Each split point is a non-terminal *node*, and each terminal node or *leaf* maps to one of the T subsets.

Subsets of decision trees are called subtrees.

Definition 2.2 (Subtree). A subtree $g(x) \in f(x)$ is any tree that can be obtained by collapsing any number of non-terminal nodes in $f(x)$.

Let $f_0(x)$ be a large tree grown until some pre-determined stopping point. For each value of α there exists a subtree $f_\alpha(x) \subset f_0(x)$ that minimises the cost complexity criterion

$$C_\alpha(f_0(x)) = \sum_{t=1}^T Q(R_t) + \alpha T, \quad (2.21)$$

where T is the number of terminal nodes in the subtree $f_\alpha(x)$ and Q is a cost function for a terminal node that depends on whether the tree is predicting a continuous or a discrete variable.

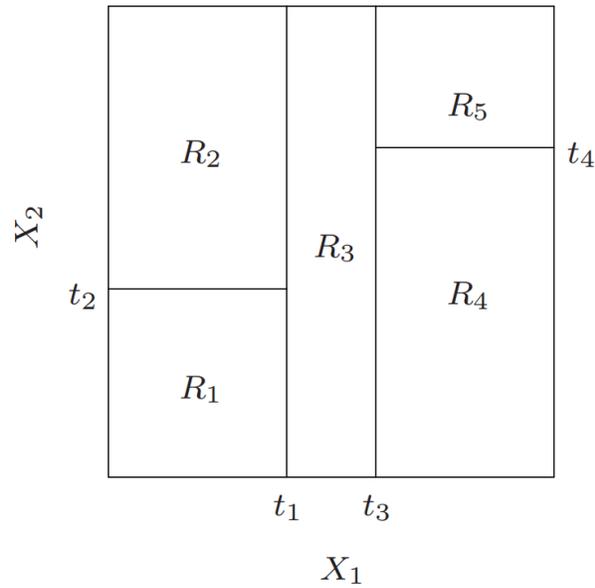


Figure 2.2: Example of a decision tree with four splits, t_1, \dots, t_4 that partition the co-domain into five leaves, R_1, \dots, R_5 (Hastie et al., 2009, p. 306).

For continuous variables, $Q(R_t)$ is the residual sum of squares

$$RSS(R_t) = \sum_{i \in R_t} (y_i - \bar{y}_{R_t})^2,$$

where \bar{y}_{R_t} is the mean of the observations in R_t . While for discrete variables, $Q(R_t)$ is the Gini index

$$G(R_t) = \sum_{k=1}^K \hat{p}_{tk}(1 - \hat{p}_{tk}),$$

where \hat{p}_{tk} is the proportion of observations in the leaf R_t that belong to the class c_k .

Breiman (1984) shows that, for any α , a sequence of trees can be constructed that must contain the optimal $f_\alpha(x)$. That sequence is constructed by sequentially collapsing the non-terminal node that gives the smallest decrease in the cost complexity criterion (Equation (2.21)), until a tree with a single node is reached (Ripley, 1996, pp.222–225).

James et al. (2013, Algorithm 8.1) describes the training procedure for CART as follows:

1. Use recursive binary splitting to grow a large tree on the training data, stopping only when each terminal node has fewer than some minimum number of observations.
2. Apply cost complexity pruning to the large tree to obtain a sequence of best sub-trees as a function of α .
3. Use cross validation to find the value of α that minimises prediction error.
4. The final trained tree is the sub-tree corresponding to the chosen value of α .

For unordered categorical predictor variables, recursive binary partitioning scales poorly with the number of classes. If a predictor has K classes, there are $2^{K-1} - 1$ possible binary partitions (Hastie et al., 2009). Consequently, searching through all possible partitions will be computationally prohibitive when K is large. In addition, the fitting algorithm is biased in favour of selecting categorical variables that have many levels. The more possible partitions that exist for a variable, the more likely that the partition that optimises Equation (2.21) will involve splitting that variable.

2.2.1 Random forests

In a random forest, each tree is trained on a bootstrapped sample of the data and restricted to only considering a random subset of the variables in the data at each split (Breiman, 2001). This restriction de-correlates the trees in the random forest by forcing them to use different splits, decreasing the model's variance (Hastie et al., 2009).

Definition 2.3 (Random Forest). Let g_b be a decision tree that is trained on the b^{th} bootstrapped sample of a training dataset that contains p variables. Each g_b is grown until some minimum node size is reached, unlike CART the trees are not pruned. At each split we randomly sample m of the p variables and the best split is chosen from those. The random forest is the ensemble of trained decision trees

$$f = \bigcup_{b=1}^B \{g_k\}.$$

A prediction for a data point $\mathbf{x} = (x_1, \dots, x_p)$ is

$$f(\mathbf{x}) = \frac{1}{B} \sum_{b=1}^B g_b(\mathbf{x}),$$

for regression. For a classification problem we predict the class that is predicted by the majority of trees

$$f(\mathbf{x}) = \text{majority}\{g_1(\mathbf{x}), \dots, g_B(\mathbf{x})\}.$$

We can specify several hyperparameters to affect the fit of random forest models, minimum node size, the number of predictors per split m and the number of trees B . As with CART, specifying a minimum node size can prevent overfitting. However, Hastie et al. (2009) find that the gains do not justify the requirement to optimise an additional parameter and instead they recommend growing trees to full. The optimal value of m depends on the number of predictor variables and the fraction that are relevant (Hastie et al., 2009). If a low fraction of the predictors are relevant, then there is a low probability of selecting a relevant predictor when m is too small relative to the number of predictors.

Out of sample prediction error is inversely related to B (Hastie et al., 2009). However, there are diminishing returns, at which point the computational cost of fitting additional trees no longer justifies the tiny reduction in error. The point that the error stabilises depends on the data.

Ideally, hyperparameters should be selected by minimising the validation loss, but this is not always computationally feasible. Unless stated, the random forest hyperparameters in this thesis are $B = 500$, with both m and minimum node size depending on whether a categorical or continuous variable is being predicted. In the continuous case, minimum node size is 5 and $m = \max(p/3, 1)$ variables are considered at each split, where p is the number of predictor variables. For categorical variables, minimum node size is set to 1 and $m = \text{floor}(\sqrt{p})$. These are the default hyperparameters for the R package `randomForests` (Liaw & Wiener, 2002), which is the implementation we use for all random forest models.

Chapter 3

Literature review: information disclosure and protecting confidentiality

This chapter begins the novel literature review, which continues through Chapter 6. We summarise the various types of information disclosure in Section 3.1 before motivating the use of synthetic data with a review of statistical disclosure control (SDC) methods for protecting confidentiality in Sections 3.2, 3.3 and 3.4.

Information disclosure occurs when a party (a person, group, etc.) gains access to sensitive information for which they have not been permitted access. The *subject* or *target* is the party for whom the disclosed information concerns. The *attacker* or *intruder* is the party attempting to discover information about a subject or subjects.

Before we discuss the types of disclosure, consider the question, *what is the identity of the person or organisation attempting to learn this sensitive information?* This thesis focuses on synthetic, subject-level data for public release. As such, we presume that the data is not so sensitive that it is of interest to extremely well resourced intruders, such as state-sponsored hackers. We also do not consider disclosures that result from allowing direct access to unmodified sensitive data. For example, if data was leaked by hackers or if an employee viewed data without permission. These would be a failure of internal security protocols and can not be solved with synthetic data.

A helpful definition to consider when thinking about the attacker in this thesis is the *motivated intruder* (Information Commissioner’s Office, 2012).

Definition 3.1 (Motivated intruder). The motivated intruder is a party that is attempting to discover sensitive information. It is assumed that they are competent but do not have access to highly restricted information or specialist techniques. Instead, their attempts at discovering sensitive information utilise resources available to the general public, such as public documents and social media.

The motivated intruder is the hypothetical antagonist in the motivated intruder test. The *motivated intruder test* determines if a motivated intruder could have made such a

disclosure. This helps establish if reasonable steps were taken to protect the data from disclosure. During FOIA appeals and Data Protection Act appeals, the ICO and Information Tribunal, respectively, use the motivated intruder test for assessing the disclosure risk of the datasets (Data Protection Act 2018, 2018; Information Commissioner’s Office, 2012).

3.1 Types of information disclosure

We will focus on three types of information disclosure due to their relevance to synthetic data. The three types are identity disclosure, attribute disclosure, and membership disclosure. While other types of disclosure exist, they either do not apply to synthetic data or fall into subcategories of the three types of information disclosure mentioned above (see, e.g., Willenborg and de Waal (2001, pp. 19–20, 42–52), Duncan and Lambert (1989)).

Identity disclosure

Identity or *re-identification disclosure* occurs when an intruder identifies the target that corresponds to a given record in the data (Elliot, 2005, pp. 663–664). Identity disclosures require observations in the data to be mapped to an individual in the population (Willenborg & de Waal, 2001, p. 41). If that mapping does not exist, there is no ground truth to verify whether an intruder has correctly identified the target. It is important to note that, for entirely synthetic data, this mapping does not exist. As such, the definition of identity disclosure is unworkable for completely synthetic data. However, if a dataset contains real and synthetic records or variables, the mapping exists and identity disclosure is possible.

Attribute disclosure

Following the definition in (Elliot, 2005, pp. 663–664), *attribute* or *predictive disclosure* occurs when an intruder learns information about the variables of a record in the data. The sensitive information can be learned directly from an identity disclosure or inferred without identification. However, in practice, the manner in which the information is acquired is irrelevant to the harm inflicted on the target of the disclosure. In fact, the disclosure risk can be grossly underestimated if the possibility of an attribute disclosure that occurs without identification is ignored. One example of attribute disclosure without identification is a *homogeneity attack*, which is further discussed in Section 3.3.1 (Machanavajjhala et al., 2007).

Membership disclosure

Membership disclosure occurs when an intruder identifies that a record belongs to a dataset (Li et al., 2013). Membership disclosure attacks (or membership inference attacks) have been studied in a variety of fields including genomics and machine learning (Backes et al., 2016; Choquette-Choo et al., 2020; Dwork et al., 2015; Hayes et al., 2018; Homer et al., 2008; Shokri et al., 2017). Recall that, for completely synthetic data, the definition of identity disclosure is unworkable due to a lack of ground truth. In contrast, membership disclosure has a known ground truth and is a workable alternative to identity disclosure.

There is discourse in the literature about how often a disclosure of membership can be considered a disclosure of sensitive information (Reiter, 2023). To illustrate a clear example of when membership disclosure is a direct disclosure of sensitive information, consider a database of cancer patients. The disclosure that an individual is a member of this database would disclose the subject’s cancer diagnosis.

Now, we consider a more complex scenario, a sample of census respondents. The data within the census dataset is definitively sensitive, but the membership of a household is not necessarily sensitive information. The 2021 England and Wales Census had a 97% response rate and the 2022 Scottish Census had a 79% response rate (Office for National Statistics, 2022; Office for Statistics Regulation, 2023). It is likely that any given household in the United Kingdom has participated in the census. As such, it is questionable whether, in this scenario, membership disclosure would be a disclosure of sensitive information.

It is our opinion that, for much of the literature, the choice to assess the membership disclosure risk of synthetic data is a choice of convenience rather than because membership is sensitive. When membership of a dataset is *not* genuinely sensitive, assessing the attribute disclosure risk is a more direct assessment of the risk of disclosure.

3.1.1 Risk factors for information disclosure

Now that we have explored the types of disclosure, a sensible follow up question is, *what are the aspects of a record that affect its risk of disclosure?* Most of the time, some records will be more at risk than others, so in this section we discuss the risk factors for identity, attribute, and membership disclosure.

Throughout the literature, there is agreement that outlying or unique observations are more vulnerable to membership and identity disclosures (Elliot et al., 2002; Longhurst & Vickers, 2007; Taylor et al., 2018). There are two reasons why outliers are especially vulnerable. The first is that outliers have rarer combinations of attributes, so fewer observations will match. Consequently, there is a higher likelihood of correctly selecting the outlier from the set of matching observations. The second is that the outlying attributes of the outlier will influence some summary statistics more than a subject that closely

matches the average.

The risk factors for attribute disclosure risk are not as widely discussed in the literature. Palley and Simonoff (1987) find that a lack of a relationship between the attributes in a dataset will make learning a model which approximates the dataset more difficult. This suggests that if an intruder utilises predictive modelling, the attribute disclosure risk of outliers would be lower.

3.2 Protecting confidentiality: non-synthetic methods

In this section, we briefly discuss non-synthetic methods for protecting against disclosures of confidential data. For a more in-depth discussion of these methods and their drawbacks, we direct the reader to Matthews and Harel (2011) and Winkler (2007).

3.2.1 Method 1: removal of personal identifiers

One obvious method for protecting against disclosures is the removal of personal identifiers. This includes but is not limited to, the removal of unique personal identifiers such as name, date of birth and identification number. However, the removal of unique personal identifiers alone does not guarantee protection against disclosure.

To illustrate, let's assume that an intruder wishes to identify a target that they know is contained in a confidential dataset, and that the intruder knows the values for some of the variables in the dataset for their target. One strategy that the intruder could implement is to identify a set of observations that contains their target by comparing variables in the confidential dataset against publicly available information for their target. As the number of identifying variables increases, the set containing the intruder's target becomes smaller. Therefore, the probability of disclosure increases. If the target has a rare value for an identifying variable, then the size of the set will be smaller than for a more common value. As such, given enough identifying information, it is possible for the intruder to identify a small enough set that the risk of identity disclosure is unacceptably high. Dalenius (1986) refers to these identifying combinations of variables as "quasi-identifiers". In this thesis and throughout the literature, such an attack is referred to as a *linkage attack*, one famous example of a linkage attack is described in Sweeney (2002).

Example 3.1 (Disclosure of Massachusetts governor's medical records). *The then-governor of Massachusetts was identified in a supposedly anonymous database containing medical histories of Massachusetts state employees. The authors were able to cross-reference the date of birth, sex, and ZIP code information in the database with the voter registration list for Cambridge, Massachusetts.*

This example was referenced during the development of the HIPAA Privacy Rule (ASPE, 1999). The HIPAA Privacy Rule introduced numerous safeguards to prevent the disclosure of personally identifiable healthcare information, including national standards for de-identifying protected health information and severe penalties for violations of the act. Other examples of linkage attacks include the successful identification of subjects in the Personal Genome Project (Sweeney et al., 2013), and the identification of subjects in the “anonymised” Netflix Prize Dataset (Narayanan & Shmatikov, 2007), which is described below.

Example 3.2 (Disclosure of Netflix prize dataset). *In 2006, Netflix organised the first of several planned competitions to help them improve the quality of their film recommendations. As part of the competition, they released a dataset that contained over 100 million film ratings that had been made for nearly 18,000 films by over 480,000 subscribers (Bennett & Lanning, 2007). The dataset contained no information about the subscribers, beyond their film ratings and the dates that those ratings were made. Users of the website IMDb can create public profiles that display their ratings for films (“IMDb. The Internet Movie Database”, 2023). Optionally, these profiles can also include personal information.*

Researchers were able to link users in the Netflix database with their IMDb profiles, by cross referencing the score and date of film reviews in each database (Narayanan & Shmatikov, 2007). Consequently, all of the information in the supposedly anonymous Netflix dataset was then linked to the IMDb profiles of those users.

Due to the issues demonstrated by this linkage attack, staff at Netflix raised concerns about customer privacy, the Federal Trade Commission opened an investigation into Netflix’s release of the data, and a lawsuit claiming that the release violated federal privacy laws was filed against Netflix (Hunt, 2010; Video Privacy Protection Act, 1988). In response, Netflix suspended plans for the second competition and reached a settlement with the plaintiffs of the lawsuit. Netflix also agreed that, prior to any future releases of customer data, it would implement safeguards to prevent re-identification and discuss the release with the FTC. These assurances satisfied the FTC, who opted to close the investigation without taking any action against Netflix (Mithal, 2010).

These examples highlight two problems with removing personal identifiers as a method of disclosure protection. First, Example 3.1 demonstrates that quasi-identifying attributes, such as age, location, and sex, constitute enough information to make identity disclosures for some individuals with high probability. These are useful attributes for many data analyses, so their complete removal could severely impact the utility of a dataset. Second, Example 3.2 demonstrates how seemingly innocuous variables can be identifying variables when cross-referenced with publicly available datasets. On the surface, the linkage of Netflix film ratings with IMDb ratings is not an especially harmful disclosure. However,

the authors were still able to infer information about the user’s political opinions and sexual preferences based on their Netflix film ratings (Narayanan & Shmatikov, 2007).

Even non-sensitive information that is gained from such attacks can potentially allow an intruder to build up a detailed profile of a person over several linkage attacks. As more information is gained, there is an increased risk of successful linkage attacks in the future. In fact, social media profiles are a major source of vulnerability to linkage attacks, such as those that we have described. Public profiles contain a large amount of information that is directly associated with a person’s identity. *Any* information in an “anonymous” database that has been shared on a person’s public social media profile can potentially be linked back to that person.

To summarise, removing personally identifying information is not, in itself, a viable method to prevent linkage attacks. Therefore, if we are to prevent disclosures, we must look at methods beyond the removal of identifying information.

3.2.2 Method 2: masking techniques

Masking techniques are methods that can be applied to reduce the risk of disclosures. Examples include random blanking, naïve truncation, sampling, rounding, adding random noise, swapping, micro-aggregation and removal of outliers (suppression). In the literature, it is agreed that these masking techniques reduce the utility of the data (Drechsler & Reiter, 2010; Purdam & Elliot, 2007). According to Winkler (2007), of the masking methods listed, only the addition of random noise will preserve the analytic properties of the data. Furthermore, the application of masking techniques may not adequately protect subjects in a dataset from re-identification. An example of this inadequate protection is seen in the following example:

Example 3.3. *The HIPAA Privacy Rule specifies requirements to de-identify data for HIPAA compliance. These requirements include removing all names, identifying numbers, rounding geographical information to areas with greater than 20,000 people, removing all dates (except year), and truncating the age of any person older than 89 years (OCR, 2012). Sweeney et al. (2017) compare a HIPAA-compliant database with other publicly available sources of data. The researchers were able to identify 25% of participants in the study by name.*

The identification of subjects in a supposedly anonymised database is extremely concerning and suggests that more needs to be done to prevent harmful disclosures. A more extensive application of the masking methods that we have described in this section would likely help to reduce the risk. However, one must question how much more severe the application of masking procedures would have to be, in order to reach a sufficiently low risk of disclosure? Furthermore, how much utility would such a dataset have? In the following

section, we discuss various standards that have been proposed in the literature for ensuring that the disclosure risk of a dataset is sufficiently low.

3.3 Standards for ensuring anonymity

Various standards have been proposed for ensuring that sensitive data is sufficiently protected from disclosure. In this section, we discuss k -anonymity, ℓ -diversity, t -closeness, and differential privacy. Later in this thesis we compare synthetic data generation, as a method of statistical disclosure control, with k -anonymisation, see Chapter 8 and Chapter 9.

3.3.1 k -anonymity

Table 3.1: Sample of observations from Pima data.

Age	Pregnancies	Diabetes
21	0	yes
21	0	yes
21	1	no
21	1	no
21	1	no
22	0	no
22	0	yes
22	0	no
22	1	yes
22	1	no
23	1	no
23	1	no
24	1	no
24	1	no

Throughout this section we refer to Table 3.1. This table contains some sample observations from the Pima dataset in Chapter 8. The Pima dataset is the subject of the first case study in this thesis.

Now, recall the term quasi-identifier from Section 3.2.1. The k -anonymity privacy model, defined in Definition 3.2, is first described in (Sweeney, 2002).

Definition 3.2 (k -anonymity). Let D be a dataset containing the columns $\{x_1, \dots, x_p\}$, and $q \subseteq \{x_1, \dots, x_p\}$ be the set of all quasi-identifying columns. Then, D satisfies k -anonymity if, for all observed combinations of q , there are at least k observations in D that match q .

k -anonymity is achieved through some combination of removal of observations, aggregation of quasi-identifying variables, and application of other masking procedures. As

discussed in Section 3.2.1, these methods often fail to preserve the analytic properties of the data.

An intruder who knows the values of all quasi-identifying variables for a target in a k -anonymous dataset, will only be able to learn that the target corresponds to one of k other observations. As such, k -anonymity ensures a minimum level of protection against identity disclosure. However, k -anonymity hinges on the assumption that all quasi-identifiers have been identified. In the case that a quasi-identifier is not included in q , then k -anonymisation does not guarantee any level of protection against identity disclosure.

For example, let's say that an intruder wishes to identify a particular target in a k -anonymous dataset D . We assume that the intruder knows their target's values for the variables (q, x_*) , where $x_* \not\subseteq q$. k -anonymity guarantees that the set of observations in D that match the target on q is at least k . However, the intruder can utilise the variable x_* to identify a smaller subset, unless every observation in that set also matches the target on x_* . So, there is no guarantee that the set matching on (q, x_*) will contain at least k observations. In our discussion of Example 3.2, we noted that variables that seem to be non-identifying can in fact be identifiers. As such, the assumption that all quasi-identifiers are known is difficult to verify and, consequently, the privacy guarantees of k -anonymity or any similar methods are difficult to verify.

Examples of algorithms that produce k -anonymous data include Datafly (Sweeney, 1998a), μ -argus (Hundepool et al., 2014; Hundepool et al., 2012), and k -similar (Sweeney, 1998b). In the following example, we explore how Table 3.1 satisfies 2-anonymity given some assumptions.

Example 3.4. *Assume that age and number of pregnancies are quasi-identifying variables. Then, there are at least two observations per subject for all combinations of the quasi-identifiers. Therefore, Table 3.1 satisfies 2-anonymity.*

Even if data guardians make sensible guesses about which variables are quasi-identifying, protection is not guaranteed. Machanavajjhala et al. (2007) describes two types of attacks on k -anonymous data. The first type of attack is a *homogeneity attack*. A homogeneity attack can occur when all observations that match on quasi-identifying variables also share the same value of a sensitive attribute. Again, consider Table 3.1.

Example 3.5. *Assume that age and number of pregnancies are quasi-identifying variables. In addition, the intruder knows that their target exists in the sample in Table 3.1 and that their target is 21 years old and has never been pregnant. Since the subjects in the sample dataset who are 21 years old and have had zero pregnancies have diabetes, the intruder easily discovers that their target has diabetes.*

The second type of attack described by Machanavajjhala et al. (2007) is a *background knowledge attack*. A background knowledge attack occurs when the attacker uses ad-

ditional information not present in the table to learn sensitive information from a k -anonymous dataset. Again, consider Table 3.1.

Example 3.6. *Assume that the intruder knows that their target is 22 years old, obese, and has had one pregnancy. Given the subjects in Table 3.1, the intruder can conclude that there is a 50% probability that their target has diabetes. In addition, obesity is a known risk factor for diabetes Diabetes UK (2019a). As such, the probability of the target having diabetes is greater than 50%.*

3.3.2 ℓ -diversity

To protect against homogeneity and background knowledge attacks, Machanavajjhala et al. (2007) propose ℓ -diversity.

Definition 3.3 (ℓ -diversity). Let D be a dataset containing the columns $\{x_1, \dots, x_p\}$; $q \subseteq \{x_1, \dots, x_p\}$ be the set of all quasi-identifying columns; and $\{s \subseteq \{x_1, \dots, x_p\} : |s| = 1 \wedge s \setminus q\}$ be a column containing a sensitive attribute. Then, D satisfies ℓ -diversity if, for all observed combinations of q , there are at least ℓ “well-represented” values for s among the observations in D that match q .

Machanavajjhala et al. (2007) suggest three definitions for “well-represented”. *Distinct ℓ -diversity* requires that, for all observed combinations of q , there are at least ℓ distinctive values for the sensitive attribute among the observations that match on q . *Entropy ℓ -diversity* and *recursive ℓ -diversity* both place further restrictions on the distribution of the sensitive attribute within each set of observations that match on q .

Example 3.7. *Consider Table 3.1. Notice that all subjects aged 21 years with zero pregnancies are diabetic, and all subjects that are aged 21 years with one pregnancy are not diabetic. By identifying that a subject belongs to one of those sets, we have identified whether or not they have diabetes. Hence, neither of the sets satisfies 2-diversity. The sets of observed subjects aged 22 years, with zero or one pregnancy, both satisfy 2-diversity, as each set contains two distinct values for diabetes¹. However, as 2-diversity is not satisfied for all observed combinations of age and pregnancies, Table 3.1 does not satisfy 2-diversity.*

Machanavajjhala et al. (2007) extend the definition of ℓ -diversity for the case of more than one sensitive attribute. However, doing so involves treating other sensitive variables as quasi-identifiers, and the larger number of quasi-identifiers can require more extensive application of masking techniques to achieve ℓ -diversity.

As described in Li et al. (2007), ℓ -diversity has several limitations. First, achieving ℓ -diversity may be unnecessary for a dataset to protect against harmful disclosures sufficiently. In Example 3.7, we noted that all subjects aged 21 years with one pregnancy in

¹For these sets, 2-diversity is satisfied for all three definitions of “well-represented”.

Table 3.1 did not have diabetes. We were able to learn this because the set does not satisfy 2-diversity. However, the majority of the population does not have diabetes. Presumably, if the intruder did not have access to Table 3.1, their default assumption would be that their target was not diabetic. So, the removal of the subjects aged 21 years with one pregnancy will have a negligible effect on the risk of harmful disclosures. Furthermore, satisfying ℓ -diversity is challenging for imbalanced variables because some combinations of the quasi-identifiers may not exist.

A *skewness attack* occurs when the distribution of the sensitive attributes differs significantly from that of the overall population (Li et al., 2007). For example, recall the earlier example of an intruder who knows that their target is obese, 22 years old, and has been pregnant once. Previously, we completely glossed over the fact that the incidence rate of diabetes for a 21-year-old is far lower than 50%. Simply identifying that the target has a much higher than average probability of diabetes is itself a disclosure of sensitive information. A *similarity attack* occurs when sensitive attributes are technically different but are similar in nature (Li et al., 2007). For example, imagine that Table 3.1 contains an additional column that shows that all subjects without diabetes have cancer. Meanwhile, the disclosure risk of diabetes would be unchanged by the additional column. The overall risk of harmful disclosures would be drastically increased. Now, the knowledge that a subject is contained in the table is equivalent to knowing that the subject has diabetes or cancer.

3.3.3 t -closeness

To protect against skewness and similarity attacks, Li et al. (2007) propose t -closeness, as follows.

Definition 3.4 (t -closeness). Let D be a dataset containing the columns $\{x_1, \dots, x_p\}$; $q \subseteq \{x_1, \dots, x_p\}$ be a set of quasi-identifying columns and $\{s \subseteq \{x_1, \dots, x_p\} : s \setminus q\}$ a column containing a sensitive attribute.

D satisfies t -closeness if the earth mover’s distance between the distribution of s for the observations that match q and the distribution of s for all observations in D , is no more than a threshold t for all observed combinations of q .

While t -closeness can be applied to datasets with multiple sensitive attributes, the earth mover’s distance can be challenging to calculate for multivariate distributions. An alternative, in that case, is to independently calculate the earth mover’s distance for each sensitive attribute in D . Then D satisfies t -closeness if each sensitive attribute satisfies t -closeness (Li et al., 2007).

Remark. An alternative approach to calculating t -closeness for datasets with multiple sensitive attributes is to borrow the method for calculating ℓ -diversity for datasets with

multiple sensitive attributes (Machanavajjhala et al., 2007). For each sensitive attribute, all other sensitive attributes would be included in the set of quasi-identifiers when calculating t -closeness. A dataset would satisfy t -closeness if t -closeness was satisfied for each sensitive attribute.

3.3.4 Differential privacy

Differential privacy is a radically different approach to the other ideas discussed in this section. To achieve a certain level of protection against disclosure, k -anonymity, ℓ -diversity and t -closeness each of which define requirements for a tabular dataset. In contrast, differential privacy defines requirements for a mechanism (or function) to achieve a certain level of protection against disclosure (Dwork, 2006).

To define differential privacy consider the amount that a probability distribution of the output of a randomised function, applied to some dataset, is allowed to change when the dataset changes by a single observation. Let ϵ be the upper bound of this amount. Then, an ϵ -differential privacy mechanism restrict the amount of influence that a single data point can have on the output of a randomised function (Dwork, 2006). The larger ϵ is, the more the probability is allowed to change. This definition of privacy tends to be extremely restrictive. Let δ denote the probability that the differentially private mechanism, ϵ , does not hold. Then, (ϵ, δ) - *differential privacy* is a relaxation of the ϵ -differential privacy (Dwork et al., 2006).

The post-processing property of differential privacy ensures that the privacy guarantees also extend to the output of a differentially private algorithm (Dwork & Roth, 2014, p. 19). Therefore, any data that was generated by an (ϵ, δ) -differentially private model would also be (ϵ, δ) -differentially private. The ramifications of this result cannot be overstated. Consider an (ϵ, δ) -differentially private function that outputs data. If the disclosure risk of ϵ and δ was acceptable, then the disclosure risk of any data that was output by the function would also be acceptable. Furthermore, since differential privacy does not make any assumptions about the intruder, the guarantees are not subject to changes in the knowledge or capabilities of an intruder.

The theoretical guarantees of differential privacy make it a very attractive idea for disclosure protection. However, the differential privacy mechanisms add noise to the data, which can reduce utility. The challenge is to develop efficient differential privacy mechanisms that ensure differential privacy without harming utility. The development of these algorithms is an active topic of research. In Section 4.3, we discuss examples of differentially private models in the literature.

3.4 Protecting confidentiality: synthetic data, a solution

Given the issues with the SDC methods that we have discussed in Section 3.2 and Section 3.3, one solution is the creation of synthetic dataset. These datasets would match the properties of variables and the relationships between them and, in theory, the risk of disclosure should be lower, since the data is not real. Synthetic versions of confidential datasets have been released by organisations from several countries.

Examples 3.8. *Examples of synthetic versions of confidential datasets include the following:*

1. *the United States Census Bureau’s synthetic Survey of Income and Program Participation (SIPP) (Abowd et al., 2006; Benedetto et al., 2017; Benedetto et al., 2013; U.S. Census Bureau, 2018),*
2. *the United States Census Bureau’s synthetic Longitudinal Business Database (SynLBD) datasets (Kinney et al., 2014; Kinney et al., 2011; U.S. Census Bureau, 2013),*
3. *the Scottish, England & Wales and Northern Irish Longitudinal Survey datasets (Dennett, 2017; Dennett et al., 2016; Elliot, 2014; Nowok et al., 2017), and*
4. *the German Institute for Employment Research’s IAB Establishment panel (Drechsler, 2009).*

Researchers working with any of the aforementioned synthetic datasets are able to validate their results by submitting code to the organisations that publish the datasets. Then, the organisation will run the code on the non-synthetic version of the data (U.S. Census Bureau, 2023). This allows researchers who do not have permission to view confidential datasets to experiment and develop with the low disclosure risk synthetic data while still being able to obtain results for the real data and simultaneously protect those in the real data from disclosure by limiting access to the real data. In addition, it has the advantage of lowering the quality threshold for the data to be useful, therefore, easing the privacy-utility trade-off.

In the following chapter, we discuss the theory of data generation and describe synthetic data generation methods.

Chapter 4

Literature review: synthesising data

In this chapter, we cover a variety of methods for generating synthetic data that are commonly used in the synthetic data literature. Specifically, in Section 4.1, we discuss the two theoretical frameworks for describing synthetic data distributions. In Section 4.2, we describe the combining rules necessary for the correct inference of synthetic datasets. Finally, Sections 4.3 and 4.4 include the two approaches for generating synthetic data: joint synthesis and sequential synthesis, respectively.

4.1 Frameworks for synthetic data

In 1993, two articles proposing the generation of synthetic data were published in a single issue of *The Journal of Official Statistics*, see Rubin (1993) and Little (1993). Within these articles, Rubin and Little each propose a framework for the generation of synthetic data. Throughout this thesis, and the wider literature, Rubin’s approach is referred to as *fully synthetic data* and Little’s approach as *partially synthetic data* (see, e.g., Drechsler (2011a, pp. 7–10)). In this section, we describe the theory and assumptions that underpin each of these approaches and advantages and disadvantages of each. In Section 4.1.3, we discuss a common misconception about these approaches and establish terminology to address this confusion.

4.1.1 Rubin’s framework: fully synthetic data

The fully synthetic data generation framework is based on Rubin’s work on multiple imputation for missing data, see Rubin (1987). In standard statistical terminology, a *population* refers to an entire group of interest and a *sample* refers to an observed subset of the population. For example, if we were conducting polling for an election, the population would be every possible voter and the sample would be the group of voters that were surveyed.

Within the fully synthetic framework, all units in the population not in the sample are treated as *missing* (Rubin, 1993). Fully synthetic data is generated in two steps (Drechsler, 2011a; T. E. Raghunathan et al., 2003). First, synthetic data is generated by imputing the unobserved, or missing, units in the population. Then, a new sample of synthetic data is randomly drawn from this population. Now, we will mathematically define the fully synthetic framework.

The following is given in (T. E. Raghunathan et al., 2003). Let N denote the size of the population. Then, the *observed data* is a sample of n observed units from the population. Let Y^o and Y^s be the observed and unobserved units from the population, respectively. Furthermore, write X for the set of variables available for the entire population. Note that X is allowed to be empty.

Recall that there are two steps to generating fully synthetic data. First, draw m sets of simple random samples X_{new} from the $N - n$ unobserved units. Second, draw synthetic replacements from the posterior predictive distribution (PPD) for each unit in the sample. That is,

$$Y^s \sim P(Y^s | X_{\text{new}}, Y^o). \quad (4.1)$$

Note, imputing the entire population is unnecessary (T. E. Raghunathan et al., 2003). Rather than imputing the entire population and sampling from that population, we have instead drawn a sample X_{new} and then imputed that sample.

As with imputation for missing data, a single synthetic sample will underestimate the uncertainty of the unobserved data (Rubin, 1987). Consequently, multiple samples must be drawn to correctly account for this uncertainty. In Section 4.2.2, we describe the combining rules that are required for valid population inferences from multiple samples of synthetic data.

Recall that the synthetic data was sampled from a population containing n observed values and $(N - n)$ synthetic values. If we were to faithfully follow the fully synthetic framework, then $n/(N - n)$ samples in the synthetic data would be real. In fact, for inferences on fully synthetic data to be valid, the synthetic data *must* be a random subsample of the $N - n$ samples of Y^s and n samples of Y^o (Drechsler, 2018). Presumably, the motivation for the release of synthetic data is an inability to release the observed (or real) data. Therefore, the inclusion of any observations seems infeasible. For this reason, most of the literature suggests that only synthetic observations should be released which implies that the real data is a sample from an infinite population (T. E. Raghunathan et al., 2003). This assumption and its validity is discussed in Section 4.2.2. For now, it suffices to say that the assumption is reasonable if N is large (Drechsler, 2018).

In this section, we summarised the theory of the fully synthetic framework for synthetic data generation. Next, we define Little's partially synthetic framework.

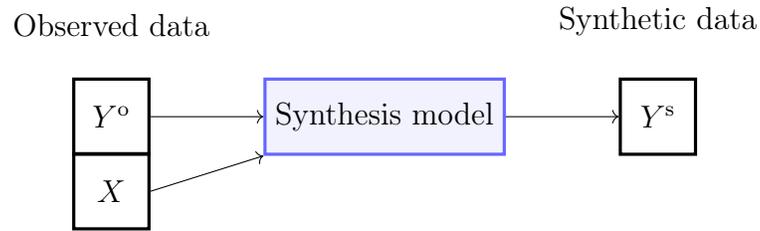


Figure 4.1: Generating synthetic data within the partially synthetic framework.

4.1.2 Little’s framework: partially synthetic data

The partially synthetic framework views the problem as a data-masking one (Little, 1993). We identify a subset of units and variables in the real sample which are too sensitive to be released. Then, synthetic values are generated for that sensitive subset. The partially synthetic dataset will consist of the non-sensitive subset of the real sample *and* the replacement synthetic values. We formally define the partially synthetic framework following Reiter (2003).

Let Y^o be N observations of real data, some of which are too sensitive to be released. Furthermore, write X for the set of variables available for the entire population. Let $Z = (Z_1, \dots, Z_N)$ denote the indicators of whether a unit will be replaced with synthetic values. Where, $Z_i = 1$ if any values of Y_i^o will be replaced with synthetic values and $Z_i = 0$ if no values will be replaced. Note that as in the fully synthetic framework, X is allowed to be empty.

Partially synthetic data is generated by drawing m sets of synthetic replacements for the sensitive values

$$Y^s \sim P(Y^s | X, Y^o, Z). \quad (4.2)$$

Then, these values are released along with the non-synthetic variables and observations

$$D = (X, Y^s, Y^o).$$

Note, partially synthetic data is conditioned only on the observed data. As such, sampling from the PPD is not necessary (Reiter & Kinney, 2012).

As we will discuss in Section 4.1.3, the phrase “partially synthetic” is a misnomer. Under the partially synthetic framework, there are no restrictions on the number of variables or observations that can be replaced. In practice, replacing synthetic values for all units in the data is common (see, e.g., Reiter and Kinney (2012)). If the entire observed sample is replaced, then the synthetic sample is not required to be the same size as the original (Drechsler, 2011b).

4.1.3 Addressing a common misconception with synthetic data terminology

Throughout this thesis, we are consistent with the literature and refer to Rubin’s approach as fully synthetic data and Little’s approach as partially synthetic data. However, in Section 4.1.1 and Section 4.1.2 we highlighted how certain aspects of these frameworks are not consistent with their naming. It is easy to think that fully synthetic data is a dataset consisting of entirely synthetic data and that partially synthetic data must contain both real and synthetic data. However, as mentioned previously, this is not the case. That is, a partially synthetic dataset can consist of entirely synthetic data and a fully synthetic dataset should, in fact, contain real data.

In fact, the distinction between the two frameworks is based on how the synthetic values are conditioned (Drechsler, 2018). Fully synthetic data is conditioned on the population, hence, new values are drawn from the PPD, Equation (4.1). In contrast, partially synthetic data is conditioned on the observed sample, Equation (4.2). Therefore, partially synthetic values can be drawn conditional on the fitted values of synthesis models (Reiter & Kinney, 2012).

To help ease the confusion surrounding the terminology, in this thesis, we follow the terminology proposed by G. Raab et al. (2016). That is, entirely synthetic data is called *completely synthetic data* and data containing a mixture of real and synthetic observations or variables is called *incompletely synthetic data*.

4.2 Variance estimation for synthetic data

In this section, we discuss methods for correctly estimating uncertainty when using synthetic data for inference. We begin by discussing the reasons to synthesise multiple datasets and then we discuss the procedures for obtaining valid uncertainty estimates from synthetic data.

4.2.1 Multiple synthesis

The parallels between synthetic data and multiple imputation are very clear, (Rubin, 1993). In both, we model the probability distribution of synthetic (or missing) observations and then draw values from that probability distribution, Equation 4.1 and Equation 4.2. A synthetic (or imputed) dataset is a single set of realisations from the synthetic (or missing) data distribution. Rubin (2004) explains the need for multiple imputation by pointing out that a single set of realisations cannot represent the uncertainty of the entire missing data distribution.

Let us consider the estimation of population quantities from synthetic data. Estimation must account for two sources of uncertainty, *sampling uncertainty* and *estimation uncertainty*. Sampling uncertainty is due to the sampling of observed values from the population. Estimation uncertainty is due to the estimation of the synthetic data distribution (Rubin, 2004). Multiple synthesis correctly accounts for both sources of uncertainty when we estimate population quantities (T. E. Raghunathan et al., 2003; Reiter, 2003).

Now let us consider the estimation of observed sample quantities from completely synthetic data. In that case, we must still account for the uncertainty due to synthesis. However, we no longer need to care about sampling uncertainty. G. Raab et al. (2016) show that, to obtain correct inferences for the observed data, a single completely synthetic dataset is sufficient. Note, this is not true for incompletely synthetic data.

In summary, the estimation of population quantities from synthetic datasets are subject to two sources of uncertainty. The solution is combining the inferences for multiple datasets. However, it is possible to correctly estimate the uncertainty of a sample quantity estimate from a single completely synthetic dataset. In the following section, we discuss the various combining rules that are available for drawing inferences with synthetic data.

4.2.2 Variance estimators for synthetic data

In the previous section, we justified the use of multiple synthetic datasets, and combining rules for correct variance estimation. The choice of appropriate combining rules for synthetic data depends on several factors. These include the synthetic data framework that was used to generate the data, whether the data is completely synthetic, and whether we are estimating a population or sample quantity (Drechsler, 2018). In this section, we shall briefly describe the estimators and the situations for which they are valid. For a more in depth discussion, Drechsler (2018) contains an excellent summary of all variance estimators, when they should be used, and provides simulation studies to demonstrate their properties.

For all variance estimators, we define n_o , n_s and M to be the number of observations in the real data, the number of observations in the synthetic datasets and the number of synthetic replications, respectively. Let q_m be the point estimate of the unknown scalar parameter Q , for the m^{th} synthetic dataset, and v_m its associated measure of uncertainty. Then, calculate the following quantities, which are made use of throughout this section.

Mean point estimate

$$\bar{q}_M = \frac{1}{M} \sum_{m=1}^M q_m,$$

within sample variance

$$\bar{v}_M = \frac{1}{M} \sum_{m=1}^M v_m,$$

and between sample variance

$$b_M = \frac{1}{M-1} \sum_{m=1}^M (\bar{q}_M - q_m)^2.$$

The variance estimator

$$T_f = (1 + M^{-1}) b_M - \bar{v}_M, \quad (4.3a)$$

is appropriate for data that is generated following Rubin's fully synthetic framework (T. E. Raghunathan et al., 2003). T_f can be negative, so Reiter (2002) proposes the alternative estimator

$$T_f^* = \max(0, T_f) + \delta \left(\frac{n_s}{n_o} \bar{v}_M \right), \quad (4.3b)$$

where $\delta = 1$ if $T_f < 0$, and otherwise $\delta = 0$. T_f is also valid for completely synthesised data that is generated with Little's framework (Drechsler, 2018).

Recall from Section 4.1.1, that Rubin's synthetic framework assumes that the sample is drawn from a population that includes non-synthetic observations. If that assumption is ignored, it implies that the synthetic data is a sample from an infinite population (T. E. Raghunathan et al., 2003). The variance estimate will be positively biased (Drechsler, 2011b), where that bias is a function of the sampling rate. Simulation studies by Drechsler (2018) show substantial bias for very large sample rates (20%), but that bias was negligible when sampling rates were smaller (1%). Small sampling rates are the norm, so the release of completely synthetic data that is generated following Rubin's framework will often be reasonable. In the rare cases that the real sample is a large proportion of the entire population, synthesisers should prefer Little's framework.

The variance estimator for synthetic data that is generated with Little's approach

$$T_p = \bar{v}_M + M^{-1} b_M, \quad (4.4a)$$

was derived by Reiter (2003). An extension for the case that $n_s \neq n_o$ was proposed by Drechsler (2011b)

$$T_{\text{alt}} = \frac{n_s}{n_o} \bar{v}_M + M^{-1} b_M. \quad (4.4b)$$

Simulation studies by Drechsler (2018) show that variance estimates for partially synthetic data are correct for a variety of conditions where the assumptions are not technically

valid. However, in cases where the inference is conditional on population level variables, the variance estimates from partially synthetic data were huge overestimates. Drechsler (2018) attributes these overestimates to a failure to account for all available population information.

Recall from Section 4.2.1, that multiple synthetic datasets are not required to estimate the variance of sample quantities. G. Raab et al. (2016) derive simpler variance estimators, for completely synthesised data, that do not require the between sample variance.

For synthetic data generated with Little’s framework

$$T_s = (n_s/n_o + M^{-1}) \bar{v}_m, \quad (4.5a)$$

and for synthetic data drawn from the PPD

$$T_{s(\text{PPD})} = (n_s/n_o + M^{-1}(1 + n_s/n_o)) \bar{v}_m. \quad (4.5b)$$

These are valid estimators for sample variance as long as the synthetic sample is completely synthesised and includes all variables on which the synthetic data generation model was conditioned.

Research has demonstrated that as M increases, attribute disclosure risk also increases (Taub et al., 2018). As such, the simple variance estimators can be utilised in situations that the disclosure risk of releasing multiple synthetic datasets is too high. This is only the case if population inference is not required. Consider, a completely synthesised dataset generated under Rubin’s Framework. Drechsler (2018) shows that these simpler variance estimators would overestimate the variance and instead recommends T_f , Equation (4.3a).

If MCMC methods are used, the use of combining rules is unnecessary. Instead, chains from the multiple synthetic datasets can be combined into a single chain, which can be used to make posterior inferences (Gelman et al., 2014, p. 452; Zhou & Reiter, 2010). A key advantage of this Bayesian approach is the avoidance of issues with biased or negative variance estimates, which can occur with Equation (4.3). Simulation studies by Si and Reiter (2011) found that, if data was synthesised following Rubin’s approach, posterior simulation was preferred to Equation (4.3a), particularly for low numbers of synthetic replications.

4.2.3 Synthesis of missing data

Up to this point, we have assumed that data does not contain missing values. In practice this is rarely true. Failing to correctly account for missingness can lead to biased estimates, and either overestimate or underestimate uncertainty (Schafer & Graham, 2002; Rubin, 1987, pp. 4–15; Rubin, 2004). Many strategies that are commonly utilised to address

missing values do not correctly account for the missingness. These include but are not limited to, the removal of observations, mean value imputation and single imputation.

We briefly discussed why multiple imputation is the correct approach for the inference of data that contains missing values in Section 4.2.1. There is a vast literature on missing data imputation (see e.g., Rubin (1987) and van Buuren and Groothuis-Oudshoorn (2011)), and it is beyond the scope of this thesis to cover the topic in depth. Instead we focus on the aspects of missing data that are relevant in the synthetic data context. If we generate synthetic data from a non-synthetic dataset that contains missing values, then we must account for two sources of uncertainty (Reiter, 2004). The uncertainty that is due to our estimation of the synthetic data distribution, and the uncertainty that is due to our estimation of the missing values.

Reiter (2004) describes a two-step procedure for generating synthetic data, that accounts for both sources of variability. For the first step, we generate R sets of complete data by imputing replacements for the missing values. At the second step, we generate M sets of synthetic data for each of the R complete datasets. Following similar notation to Section 4.2.2, the combining rules for inference of data that is generated from the two-step procedure are as follows. We denote the point estimate of the unknown scalar parameter Q for the $(m, r)^{th}$ synthetic dataset and its associated measure of uncertainty as q_{mr} and v_{mr} respectively.

Then we calculate the quantities:

$$\begin{aligned}\bar{q}_M &= \frac{1}{MR} \sum_{m=1}^M \sum_{r=1}^R q_{mr} = \frac{1}{M} \sum_{m=1}^M \bar{q}_m, \\ \bar{v}_M &= \frac{1}{MR} \sum_{m=1}^M \sum_{r=1}^R v_{mr}, \\ \bar{b}_M &= \frac{1}{M(R-1)} \sum_{m=1}^M \sum_{r=1}^R (q_{mr} - \bar{q}_m)^2 = \frac{1}{M} \sum_{m=1}^M b_m, \\ B_M &= \frac{1}{M-1} \sum_{m=1}^M (\bar{q}_M - \bar{q}_m)^2.\end{aligned}$$

Finally we compute the variance estimator of \bar{q}_M ,

$$T_M = (1 + 1/M)B_M - \bar{b}_M/R + \bar{v}_M.$$

These combining rules only apply to data that is generated following Little's framework. Combining rules for fully synthetic datasets are yet to be developed and derivation of these rules is a topic for future research (Drechsler, 2011a, p. 65).

4.3 Joint modelling

In this section we discuss joint modelling, which is one of the two approaches for generating draws from the synthetic data distribution (Drechsler, 2011a, p. 14). Under the joint modelling approach, we model the data with a multivariate distribution and then draw synthetic samples from that model. We begin by outlining the general approach to joint modelling before we discuss specific joint modelling methods from the literature.

Recall that the synthetic data distribution

$$Y^s \sim P(Y^s|X, Y^o),$$

is modelled conditionally on the observed data Y^o and population variables X . Note that, in this thesis, we omit Z for convenience since we always replace the entire dataset with synthetic values. If we were only replacing some units with synthetic values we would condition all synthesis models on Z .

Joint modelling assumes that a multivariate distribution can be specified for the synthetic data,

$$\begin{aligned} P(Y^s|X, Y^o) &= \int P(Y^s, \theta|X, Y^o) d\theta \\ &= \int P(Y^s|X, Y^o, \theta)P(\theta|X, Y^o) d\theta. \end{aligned} \quad (4.6)$$

where θ denote the model parameters.

Following Drechsler (2011a, p. 14), synthetic data can be drawn from Equation (4.6) in two steps.

1. Draw $\tilde{\theta} \sim P(\theta|X, Y^o)$.
2. Draw $Y^s \sim P(Y^s|X, Y^o, \tilde{\theta})$.

Recall that the first step is not necessary for partially synthetic data (Reiter & Kinney, 2012). Under the partially synthetic approach, both X and Y^o are samples. Hence, $P(\theta|X, Y^o)$ is fixed. Consequently, Y^s can be drawn conditional on the posterior modes or maximum likelihood estimates of θ . In the completely synthetic data setting, synthetic data is drawn from the posterior predictive distribution, therefore the first step is still necessary.

Multivariate normal models

Early synthesis approaches modelled the joint data distribution (Equation (4.6)) as multivariate normal. This mirrored the state-of-the-art approach for imputation of multivariate

data at the time (Schafer & Graham, 2002). T. E. Raghunathan et al. (2003) generate synthetic data from a multivariate normal distribution with an inverse Wishart prior placed on the covariance matrix. Imputation of missing data with this model was described by Schafer (1997). Mateo-Sanz et al. (2004) introduce a method that uses Latin hypercube sampling to generate synthetic data from a multivariate normal distribution.

The assumption that continuous data follows a normal distribution is rarely true in practice. However, the standard statistical approaches, such as log transformations of skewed variables, can help achieve a data distribution that is close to normal. Synthesis models can also be robust to departures from assumptions about normality. For example, Matthews et al. (2009) generate synthetic binary variables with a discretised multivariate normal model. This model had previously been applied to the problem of missing data imputation (Bernaards et al., 2006).

Other models for multivariate categorical or mixed data, such as Dirichlet multinomial, log-linear and general location models are described in the missing data context by Schafer (1997). Little et al. (2004) replace a subset of observations in a multivariate mixed dataset with values that are generated from a general location model. However, there are not many other examples of synthesising multivariate categorical or mixed data in the early synthetic data literature.

Despite the initial popularity of the joint modelling approach for synthesis of multivariate data. The synthetic data literature followed the trend of the missing data literature, where sequential modelling became the standard approach for imputing missing values in multivariate data (T. E. Raghunathan et al., 2001; van Buuren & Groothuis-Oudshoorn, 2011). The flexibility of the sequential modelling approach, which we discuss in Section 4.4, allows us to synthesise datasets for which the multivariate normal model is completely inappropriate. More recently, new models that allow for the specification of more flexible joint distributions have been applied to the problem of generating synthetic data. For the remainder of this section, we discuss various methods for modelling the joint data distribution.

Dirichlet Process Mixture of Products of Multinomial (DPMPM) models

Dirichlet process mixture of products of multinomials (DPMPM) models are a class of non-parametric Bayesian latent variable models that can model high dimensional multinomial variables (Dunson & Xing, 2009). Note that any dataset of unordered categorical variables can be represented as a contingency table. The cells of that contingency table form a high dimensional multinomial variable, which can be modelled with a DPMPM model.

Examples of applications of DPMPM models include imputing missing values in a contingency table with 10^{30} cells (Si & Reiter, 2013), and synthetic generation of a contingency table with 8.7×10^7 cells (Hu et al., 2014). Another extension to DPMPM models

introduces a parameter that can be tuned to balance the utility-privacy tradeoff of the synthetic data (Hu & Hoshino, 2018).

DPMPM models are capable of handling very large numbers of categories, as evidenced by the extremely large cell counts of the examples we have mentioned. The number of cells in the contingency table scales exponentially with the number of variables, and the number of categories of those variables. As such, the contingency table size will quickly scale beyond what is possible.

Structural zeros are impossible outcomes for which the corresponding contingency table cell must be zero (Upton & Cook, 2014). The standard DPMPM model assumes a non-zero probability of observing any combination of values in the data. Consequently, these models are not suitable for data with structural zeros. For example, in a dataset of employment information, one would not expect to find an individual who is both unemployed and earning a salary. DPMPM models were extended to synthesise data containing structural zeros by Manrique-Vallier and Hu (2018). The contingency table that they synthesised contained 5.5×10^9 cells, 5.2×10^9 of which were structural zeros.

DPMPM models can be fit using the R package `NPBayesImputeCat` (Hu et al., 2021). This package contains implementations for both multiple imputation and synthetic data generation. In addition, it handles structural zeros.

Gaussian mixture models

Gaussian mixture models (GMMs) can be viewed as an extension of the multivariate normal models that we discussed earlier. Rather than fitting the data to a single multivariate normal distribution, we assume that the data belongs to one of multiple multivariate normal distributions. In contrast to the restriction of a single multivariate normal, GMMs are a flexible model that can fit complex distributions (Bishop, 2006).

There are a few examples in the literature of synthesising data from GMMs. In one, the minimum prediction volume of a cluster was constrained to reduce the risk of privacy disclosures occurring (Oganian, 2014). In another, probabilistic k -anonymity was enforced by constraining clusters to all have a minimum membership probability greater than k/n , where k is the k -anonymity value and n is the sample size (Oganian & Domingo-Ferrer, 2017). These examples both leverage the clustering of the GMMs for disclosure prevention. If the GMM is fit without restriction, then outlying observations may be assigned to a small outlying cluster. However, by enforcing constraints that ensure clusters are of a reasonable size, the influence of outliers on the model fit is limited.

Generative neural networks: Variational Auto-Encoders (VAEs)

Variational autoencoders (VAEs) are generative neural networks introduced by (Kingma & Welling, 2014). VAEs use an encoder-decoder pair of neural networks to map between

the data and a latent space. The latent variables are assumed to follow some known prior distribution. New data is generated by randomly sampling latent variables from their prior distribution. These latent variables are then transformed to the data space by the decoder network.

Nazábal et al. (2020) describe a general method for the generation of heterogeneous data with VAEs. This is achieved by specifying a generator that has separate outputs for each variable. An appropriate probability distribution is chosen for each variable in the data and, then, the model is trained to optimise an overall likelihood. This overall likelihood is calculated by summing the log-likelihoods of each variable.

Nazábal et al. (2020) present two formulations of their model. In one formulation, the prior distribution of the latent variable is formed of independent Gaussian distributions. This formulation can be too restrictive. As such, in the second formulation of their model, they place a Gaussian mixture prior on the latent variable.

Generative neural networks: Generative Adversarial Networks (GANs)

Another family of generative neural networks are generative adversarial networks (GANs), introduced by Goodfellow et al. (2014). GANs consist of a pair of networks, one generates the data from a random noise input, while the other network is a classifier that learns to discriminate between real samples of data and samples from the generator network. This family of models has been successfully applied to many data generation problems. Examples include image generation, text generation, music generation, and video generation (Dong et al., 2017; Guo et al., 2017; Saito et al., 2020; Sauer et al., 2021).

While GANs are powerful generators that can learn complex data distributions, they are also known to be unstable and difficult to train (Arjovsky & Bottou, 2017). One common issue is *mode collapse*, a phenomenon where the generator only learns to produce a small set of outputs. More recently, the Wasserstein GAN (WGAN) variant was introduced, which rectifies some of these issues (Arjovsky et al., 2017; Gulrajani et al., 2017).

Choi et al. (2017) use a combination of GANs and autoencoders to generate discrete medical data. They use an autoencoder to map between data and a latent space. The generator outputs data in the latent dimension, while the discriminator differentiates between the real data and the decoded generator output. Torfi and Fox (2020) use a similar setup with a GAN and autoencoder but with a different architecture for the generator. Specifically, they include 1-dimensional convolutional layers that allow the model to capture temporal relationships in the data. This model can generate both discrete and continuous data.

Generative neural networks: Differentially private GANs

The development of differentially private mechanisms (Section 3.3.4) for training neural networks is an exciting direction of research. Differential privacy ensures strong privacy guarantees. Furthermore, the post-processing property ensures that those guarantees will extend to the output of the model (Dwork & Roth, 2014, p. 19).

Abadi et al. (2016) introduce an (ϵ, δ) -differentially private algorithm for stochastic gradient descent (DP-SGD) and a method for tracking the privacy loss incurred during training. Their algorithm achieves differential privacy by clipping and adding noise to the gradients. Several improvements for differentially private gradient descent were introduced by Mironov (2017). They propose a different relaxation of ϵ -differential privacy. Their relaxation allows tighter estimates for privacy loss than (ϵ, δ) -differential privacy. In addition, they simplify the analysis of the Gaussian noise mechanism.

The only part of the GAN that “sees” the data is the discriminator. Consequently, the generator can be trained with regular stochastic gradient descent. The results from training GANs with DP-SGD have been mixed so far. A 14 variable dataset was generated which performed well on the, albeit limited, assessments that they implemented (Frigerio et al., 2019). A GAN was trained to generate a sequence of blood pressure and heart rate measurements of comparable quality to the original measurements (Beaulieu-Jones et al., 2019). In another example, a GAN was trained to generate a high dimensional dataset of binary variables from electronic health records (Xie et al., 2018). Their results show that the model had somewhat learned the data distribution, although there was a noticeable deterioration in quality, even for relatively large values of ϵ . Lin et al. (2019) found that implementing DP-SGD training destroyed the temporal correlations of the data, in comparison to regular stochastic gradient descent.

Private aggregation of teacher ensembles (PATE) GAN implements a differentially private mechanism that does not rely on DP-SGD (Jordon et al., 2019). In PATE GAN, the discriminator consists of an ensemble of “teacher” networks and a single “student” network. The teacher networks are trained to discriminate on a disjoint subset of the real labelled data, while the student network learns to discriminate from a noisy aggregation of the labels that are output by the teachers. Training the teachers on disjoint subsets and training both the student and generator on noisy labels from the teachers, ensures that each query of PATE GAN is differentially private with a known privacy cost.

Stadler et al. (2022) carried out a comparison of the disclosure risk of several implementations of differentially private GANs. They discovered that the disclosure risk of data that was generated by the models was higher than one would expect, given the differential privacy guarantees. Upon investigation, they found that the models required metadata, such as the ranges of variables, to run. For convenience, they automatically extracted this from the data but this was outwith the differentially private mechanism. Consequently,

the models were not actually differentially private. Stadler et al. (2022) patched the models and found that the disclosure risk was significantly reduced, however, the quality of data generated was also reduced.

4.4 Sequential modelling

Under the sequential modelling approach, synthetic data is generated one variable at a time from a sequence of conditional models. This approach is analogous to the popular multivariate imputation by chained equations (MICE) approach to handling missing data (van Buuren & Groothuis-Oudshoorn, 2011). We begin with a general overview of sequential modelling, before we outline the specifics of the sequential modelling step for both regression and decision tree models in Sections 4.4.2 and 4.4.3.

Let y_j^o and y_j^s be the observed and synthetic vectors for the j^{th} variable to be synthesised, and θ_j be the model parameters for f_j , the model from which the j^{th} variable will be synthesised,

$$\begin{aligned} Y^o &= (y_1^o, \dots, y_p^o), \\ Y^s &= (y_1^s, \dots, y_p^s), \\ \theta &= (\theta_1, \dots, \theta_p). \end{aligned}$$

The parameterised joint data distribution in Equation (4.6) is factorised into the product of univariate conditional distributions,

$$\begin{aligned} P(Y^s, \theta | X, Y^o) &= P(y_1^s, \dots, y_p^s | X, Y^o, \theta) P(\theta | X, Y^o), \\ &\approx \prod_{j=1}^p P(y_j^s | X, Y^o, Y_{-j}^s, \theta_j) P(\theta_j | X, Y^o), \end{aligned}$$

where Y_{-j} denotes the first through $(j-1)^{\text{th}}$ columns of the matrix Y (T. E. Raghunathan et al., 2001).

Draws are approximated from the joint distribution by sequentially drawing from the univariate conditional distributions of each variable $j = (1, \dots, p)$:

1. Draw $\hat{\theta}_j \sim P(\theta_j | X, Y_{-j}^o)$.
2. Draw $y_j^s \sim f_j(y_j^s | X, Y^o, Y_{-j}^s, \theta_j = \hat{\theta}_j)$.

A sequential modelling step consists of training a model on the observed data and then drawing values of a synthetic variable conditional on previously synthesised variables (T. E. Raghunathan et al., 2001). For fully synthetic data, model parameters are drawn from the PPD. In the case of partially synthetic data, this is unnecessary. As such, step 1 can

be skipped (Reiter & Kinney, 2012) and y_j^s is drawn conditional on the posterior modes or maximum likelihood estimates of θ_j .

4.4.1 Order of synthesis

Sequential modelling requires us to specify the order that variables are synthesised. There is some evidence that the quality of synthetic data is affected by this choice (El Emam et al., 2021; Goncalves et al., 2020; G. M. Raab et al., 2017). Several strategies for deciding synthesis order are suggested and discussed in the literature. However, Caiola and Reiter (2010) note the lack of any mathematical theory to suggest a particular ordering strategy. In this section, we discuss the following commonly utilised strategies for deciding the order of synthesis:

- a) Empirical optimisation,
- b) strongest predictors first,
- c) high cardinality variables last, and
- d) using prior knowledge of the relationships between predictors.

An obvious strategy is to choose an ordering that optimises some metric for utility or privacy, although this is problematic for several reasons. There is the difficulty of specifying a good metric for either utility or privacy. We discuss this in Chapters 5 and 6, so for now let's assume that we have a good metric to score synthetic data. The other problem is the computational cost of optimising the order of synthesis. Not only does the computation time of sequential synthesis increase with the number of variables. As the number of variables increases, the number of permutations for synthesis order increases exponentially. As such, it very quickly becomes impossible to search through more than a fraction of possible orderings. El Emam et al. (2021) explore a particle swarm optimisation approach for selecting the order of synthesis and find that it achieves the optimal order faster than a random search. The particle swarm approach is promising, however, their study is limited to a comparison of two strategies for choosing order of synthesis. Further research is necessary to compare the utility of particle swarm optimisation with other strategies.

For datasets where optimisation is computationally infeasible, strategies **b)** through **d)** select the order of synthesis without repeatedly synthesising datasets. G. M. Raab et al. (2017) recommends improving utility by synthesising the variables with the strongest variables together and near the beginning of the sequence. In theory, this allows them to have the greatest affect on each other, before the synthesis models are diluted by other variables. Caiola and Reiter (2010) follow a similar reasoning when they suggest generating

incompletely synthetic data based on the number of replacement values. They presume that quality of the synthetic variables will be highest towards the start of the sequence, although there has been little research to verify whether this improves utility. G. M. Raab et al. (2017) noted that the variable ‘occupation’ had a strong effect on the utility of their synthetic data, so they experimented with synthesising occupation first. This improved the utility of other variables that were synthesised early in the sequence. However, the utility of variables synthesised towards the end of the sequence decreased.

For data that is synthesised with decision trees, synthesising the variables with many categories at the end of the sequence will speed up synthesis (Caiola & Reiter, 2010). Recall that the computational complexity of partitioning unordered categorical variables scales exponentially with the number of categories (see Section 2.2). When we sequentially synthesise data, we generate each variable by conditioning on all previously synthesised variables. Consequently, we can expect to partition variables that are synthesised earlier in the sequence more often than those that are synthesised later. Therefore, by synthesising variables with many categories at the end of the sequence, we can drastically reduce the computational complexity of fitting the entire sequence of models.

There is some speculation that synthesising variables with low cardinality first will also improve the quality of synthetic data (El Emam et al., 2021). Goncalves et al. (2020) find that, when data is synthesised with logistic regression and low cardinality variables are synthesised first, the correlation matrices are more similar to the original data. However, they do not find any improvement from synthesising low cardinality variables first for CART synthesised data.

The final strategy we will discuss is to select a synthesis order that is logically consistent with our prior knowledge of the relationships between variables. For example, Reiter (2005a) opts to synthesise alimony and child support payments after synthesising education and marital status. This strategy encourages the use of synthesis models that are closer to the underlying mechanisms that cause the real data. While conceptually this is an appealing strategy, we are not aware of any research demonstrating that it improves the quality of synthetic data. Furthermore, there will often be ambiguity when deciding on the logical synthesis order. For example, Reiter (2005a) follows the logical order of synthesising marital status before alimony payments, but should child support be synthesised before or after alimony? One possible solution to deal with such ambiguity is to use strategy **b)** or **d)** as a tiebreaker.

In summary, there is a lack of literature on the topic of choosing the order of synthesis. Several common strategies for selecting synthesis order are justified based on their hypothesised benefits or the results of limited experimentation. So further research into understanding how synthesis order affects the quality of synthetic data is necessary. Ideally we would treat the order as a parameter to optimise, however, that can be time consuming

when there are many permutations. It is unclear from current research how much utility is gained by changing the order of synthesis. If the gains are small, then it may be better to spend computation time on optimising other modelling parameters.

4.4.2 Synthesising with regression models

Regression models for the sequential modelling step can be fit with maximum likelihood or Bayesian methods. Bayesian methods lend themselves well to fully synthetic data, as new values can be directly drawn from the posterior predictive distribution. In the case that a linear regression model was fit with maximum likelihood, Rubin (1987, p. 167) presents a method for drawing β and σ^2 from their posterior predictive distributions, which we describe in the next section. In the literature, this method is also referred to as the Bayesian method. Fitting the model with maximum likelihood is equivalent to using a uniform prior. As such, it may not meet some statistician's definitions of Bayesian.

In the partially synthetic case, new synthetic values can be drawn conditional on the maximum likelihood estimates,

$$Y^s \sim P(Y_j^s | X, Y_{-j}^s, \hat{\theta}_j),$$

where $\hat{\theta}_j$ are the maximum likelihood estimates for parameters of the j^{th} model (Reiter & Kinney, 2012).

Synthesis using linear models: The Bayesian method

For a linear regression model, we assume the j^{th} synthetic variable belongs to the distribution

$$y_i \sim \mathcal{N}(\mathbf{x}_i^T \beta, \sigma^2),$$

where the design matrix of predictor variables

$$\mathbf{x}_i^T = (\mathbf{1}, x_{i1}, \dots, x_{ip}),$$

includes the $(j - 1)$ variables that were already synthesised and any additional design variables such as stratification indicators.

Recall the least squares estimates given in Equation (2.16). The population statistics β and σ^2 are these least squares estimates given via the relationships

$$\beta \sim \mathcal{N}(\hat{\beta}, (\mathbf{X}^T \mathbf{X})^{-1} \sigma^2),$$

and

$$(n - p - 1) \hat{\sigma}^2 \sim \sigma^2 \chi_{n-p-1}^2.$$

For the detailed mathematics of this, see (Hastie et al., 2009, p. 47).

For generating fully synthetic data, the linear regression is fit to the n^o observations of data, and we then sample \mathbf{y}^s from the population distribution. Rubin (1987, p. 167) describes the procedure for this as follows:

1. Draw a random variable $g \sim \chi_{n-p-1}^2$ and let

$$\sigma_*^2 = \frac{\hat{\sigma}^2(n-p-1)}{g}.$$

2. Draw $p-1$ independent random variables $\mathbf{z} \sim \mathcal{N}(0, 1)$ and let

$$\boldsymbol{\beta}_* = \hat{\boldsymbol{\beta}} + \sigma_*(\mathbf{X}^T \mathbf{X})^{-\frac{1}{2}} \mathbf{z},$$

where $(\mathbf{X}^T \mathbf{X})^{-\frac{1}{2}}$ is the lower triangular matrix from the Cholesky factorisation of $(\mathbf{X}^T \mathbf{X})^{-1}$.

3. Draw new synthetic values for $i = (1, \dots, n^s)$

$$z_i \sim \mathcal{N}(0, 1),$$

$$y_i^s = \mathbf{x}_i^T \boldsymbol{\beta}_* + z_i \sigma_*.$$

For partially synthetic data, the distribution of the original sample is being synthesised rather than the distribution of the population to which the original data belongs. Therefore, $\boldsymbol{\beta}_*$ and σ_*^2 can be replaced with the least squares estimators $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$ and skip to step 3.

In practice it is rare that continuous data follows a normal distribution. However, as in many other statistical contexts, linear regression is fairly robust to deviations from normality and transformations of y_i can help to achieve normality. In the case that y_i can not be transformed to a normal distribution, a generalised linear regression may be appropriate.

Synthesis using generalised linear models: The Bayesian method

Data can be synthesised from generalised linear models (Section 2.1.1) with a similar procedure to linear regression. Rubin (1987, pp. 169–170) describes the generation of imputations from a logistic regression model using the procedure which assumes that the sampling distribution of model parameters are normally distributed with covariance equal to the inverse Fisher information. We know that the sampling distribution of the model

parameters will converge to this for large sample sizes. However, convergence is not guaranteed for the finite sample sizes that are dealt with in practice. Brand (1999, pp. 93–95) and T. E. Raghunathan et al. (2001, Appendix) describe variations for generating imputations from multinomial and Poisson distributions, respectively. Kleinke and Reinecke (2013) describe variations for synthesising from negative binomial, zero-inflated Poisson and zero-inflated negative binomial distributions. Rubin (1987) describes the procedure for generating synthetic data from a generalised linear regression with the Bayesian method as follows:

Let $\hat{\boldsymbol{\beta}}$ be the maximum likelihood estimates for the parameters of a generalised linear regression model

$$\begin{aligned} \mathbf{Y} &\sim p(\mathbf{Y}|\mathbf{X}, \boldsymbol{\beta}), \\ g(\mathbb{E}[Y_i]) &= g(\mu_i) = \mathbf{x}_i^T \boldsymbol{\beta}, \end{aligned}$$

where

- $\mathbf{y}_i^o = (y_1, \dots, y_{n^o})$ are real observations of the j^{th} variable,
- $\mathbf{X}^o = (\mathbf{x}_1^o, \dots, \mathbf{x}_n^o)^T$ is the $(n^o \times p)$ design matrix of real predictor variables — which includes the real observations of all previously synthesised variables \mathbf{Y}_{-j}^o and any additional design variables such as stratification variables,
- $\mathbf{X}^s = (\mathbf{x}_1^s, \dots, \mathbf{x}_n^s)^T$ is the $(n^s \times p)$ design matrix of synthetic predictor variables — which includes all previously synthesised variables \mathbf{Y}_{-j}^s and any additional design variables such as stratification variables.

The steps for synthesising data from this model are as follows:

1. Draw p independent random variables $\mathbf{z} \sim \mathcal{N}(0, 1)$ and let

$$\boldsymbol{\beta}_* = \hat{\boldsymbol{\beta}} + (\mathbf{X}^{oT} \mathbf{X}^o)^{-\frac{1}{2}} \mathbf{z},$$

where $\mathbf{A}^{-\frac{1}{2}}$ is the lower triangular matrix from the Cholesky factorisation of a matrix \mathbf{A} .

2. Draw n^s values of y_i^s :

$$y_i^s \sim p(y|\mathbf{x}^s, \boldsymbol{\beta}_*).$$

The steps for which are:

- (a) Draw n^s values of μ_i as

$$\mu_i = g^{-1} \left((\mathbf{x}_i^s)^T \boldsymbol{\beta}_* + z_i \right),$$

where $z_i \sim \mathcal{N}(0, 1)$ for $i = (1, \dots, n^s)$.

- (b) Generate random draws from the probability distribution of y_i^s , conditional on the parameter μ_i .

For example:

- If y_i^s has a Poisson distribution,

$$y_i^s \sim \text{Poi}(\mu_i).$$

- If y_i^s has a Bernoulli distribution (Section 2.1.1),

$$y_i^s \sim \text{Bernoulli}(\mu_i).$$

- If \mathbf{y}_i^s has a categorical distribution (Section 2.1.1),

$$\mathbf{y}_i^s \sim \text{Categorical}(\boldsymbol{\mu}_i),$$

where $\boldsymbol{\mu}_i$ contains the predicted probabilities of each category.

For partially synthetic data, we can replace $\boldsymbol{\beta}_*$ with $\hat{\boldsymbol{\beta}}$ and skip directly to step 2.

4.4.3 Synthesising with decision tree models

Classification and regression tree (CART) models can generate synthetic variables within the sequential framework by doing the following. First, traversing the fitted tree according to the values of the already synthesised variables for an observation until a terminal node is reached and then second, randomly sampling a new value from the terminal node (Reiter, 2005b).

Let \mathbf{y}_j^o be the real observations of the variable to be synthesised and let \mathbf{Y}_{-j}^o and \mathbf{Y}_{-j}^s be the $(n^o \times (j - 1))$ real and $(n^s \times (j - 1))$ synthetic matrices of variables that were synthesised at the previous $j - 1$ steps.

To generate synthetic values for the j^{th} variable \mathbf{y}_j^s from a decision tree:

1. Fit a decision tree that partitions the non-synthetic data into T disjoint subsets,

$$f(y_{i,1}^o, \dots, y_{i,(j-1)}^o) \longrightarrow \{R_t : t = 1, \dots, T\},$$

such that

$$\mathcal{Y}_t = \{y_{i,j}^o : f(y_{i,1}^o, \dots, y_{i,(j-1)}^o) = R_t\},$$

are the subsets that are mapped by f to R_t .

2. For $i = 1, \dots, n^s$;

- (a) Find the subset to which the i^{th} synthetic observation is mapped by f ,

$$R_i = f(y_{i,1}^s, \dots, y_{i,(j-1)}^s).$$

- (b) Uniformly draw a synthetic observation from the corresponding set of observations

$$y_{i,j}^s \sim \mathcal{U}(\mathcal{Y}_i).$$

Random forests

Caiola and Reiter (2010) show how random forests can be used to generate synthetic categorical data by drawing synthetic observations from the combined set of leaves that the trees in the forest predict. We summarise their explanation now. To generate synthetic values for the j^{th} variable \mathbf{y}_j^s from a random forest:

1. Fit a random forest model containing B trees

$$f_{\text{rf}}(y_{i,1}^o, \dots, y_{i,(j-1)}^o) \longrightarrow \left\{ \bigcup_{b=1}^B f_b(y_{i,1}^o, \dots, y_{i,(j-1)}^o) \right\},$$

where the b^{th} tree partitions the non-synthetic data into T_b disjoint subsets

$$f_b(y_{i,1}^o, \dots, y_{i,(j-1)}^o) \longrightarrow R_{t_b},$$

$$\mathcal{Y}_{t_b} = \{y_{i,j}^o : f_b(y_{i,1}^o, \dots, y_{i,(j-1)}^o) = R_{t_b}\},$$

for $t_b = 1, \dots, T_b$.

2. For $i = 1, \dots, n^s$;

- (a) Find the set of real observations to which the i^{th} synthetic observation is mapped by f_{rf} ,

$$\mathcal{Y} = \left\{ \bigcup_{b=1}^B \mathcal{Y}_{t_b} : f_b(y_{i,1}^s, \dots, y_{i,(j-1)}^s) \in R_{t_b} \right\},$$

- (b) and uniformly draw a synthetic observation from that set,

$$y_{i,j}^s \sim \mathcal{U}(\mathcal{Y}).$$

In Section 2.2.1 we discuss general hyperparameter tuning of random forests, but there is limited research into hyperparameter tuning in the context of synthetic data generation. Caiola and Reiter (2010) do not carry out any hyperparameter tuning and synthesise data

from random forests with 500 trees. Their justification is that 500 trees is a common choice when fitting random forests. Shah et al. (2014) investigate the effect of the number of trees parameter in the closely related context of missing data imputation. They find that the quality of categorical variable imputations is similar for forests with 10 and 100 trees, and the bias of parameter estimates for continuous variables increases with the number of trees in the forest. However, these results are based on a single analysis of a single study, so it's not clear whether they will generalise to missing data imputation of other datasets and it's even less clear whether they will generalise to data synthesis. Ideally, hyperparameters should be chosen through hyperparameter tuning. But the computational cost of synthesising large datasets with random forests can be high, so the additional cost of a hyperparameter search may be prohibitive.

Reducing disclosure risk of decision tree synthesised data

Decision tree models tend to be strong predictors. This helps to produce high-quality data but it can lead to overfitting, which can be a problem from a disclosure standpoint. Reiter (2005b) suggests several methods synthesisers can use to control the disclosure risk of the decision tree models. These methods include reducing the complexity through pruning, enforcing some minimum number of observations per terminal node, and smoothing the leaves of trees before drawing samples of numeric variables

$$\hat{Y}_t = s(\mathcal{Y}_t) \forall t, \quad (4.7)$$

where s is some smoothing function.

The R package `synthpop` (Nowok et al., 2016) applies smoothing to the entire set of synthetic values after they have been drawn,

$$\hat{\mathbf{y}}_j^s = s(\mathbf{y}_j^s). \quad (4.8)$$

By default, `synthpop` smooths with cubic splines (James et al., 2013, p. 277), although kernel density estimation and rolling average methods are available. In the case of random forest models, smoothing can be applied to the set of observations to which the forest maps.

Smoothing the data allows the use of smoothing functions that have a minimum value requirement greater than the minimum node size and it is more computationally efficient than smoothing the leaves. Intuitively, smoothing the data rather than the leaves will better protect outliers, especially for trees with a small minimum node size. We are not aware of any research into the differences in data quality or disclosure risk from smoothing the leaves or the data.

Disclosure risk can also be managed by controlling the size of the tree. Larger trees are thought to generate higher quality data with a greater risk of disclosure, and vice

versa for smaller trees (Reiter, 2005b). Tree size can be controlled by either specifying a stopping condition for the recursive splitting algorithm, or pruning trees until some disclosure criteria have been satisfied.

Generating fully synthetic data from decision trees using the bootstrap

The methods for generating synthetic data from decision tree models described in the previous section are also appropriate for generating partially synthetic data. They do not account for the uncertainty of sampling the observed data from the population. Therefore, they are not appropriate for making inferences about the entire population. van Buuren and Groothuis-Oudshoorn (2011) describes the bootstrap procedure which incorporates this uncertainty by sampling with replacement from the observed data and fitting models to the resampled dataset. Mathematically, it is written as follows. The observed data \mathbf{X}_{-j}^o , which is fixed, is replaced with a non-parametric, non-informative multinomial prior

$$\tilde{\mathbf{X}}_{-j} \sim \text{Multinomial}(n_o, \hat{\omega}),$$

where each unique row $\mathbf{x}_i^o \in \mathbf{X}_{-j}^o$ is an event with probability equal to the observed proportion of rows in \mathbf{X}_{-j}^o that are equal to \mathbf{x}_i^o (Hastie et al., 2009, pp. 271–272).

This bootstrapping procedure can also be applied to many other data synthesis procedures. For example, it can be used as an alternative to the regression synthesis methods described in Section 4.4.2.

4.4.4 Examples of sequential modelling

The strength of the sequential approach lies in its flexibility. A model is specified for each variable, which allows for easy handling of data that contains a mixture of data types. The data synthesiser can select models that are tailored to the complexities of each variable and incorporate specific domain expertise.

T. E. Raghunathan et al. (2001) describes how, when imputing missing data, the regression model for each variable could be chosen to match its distribution. For example, a linear regression on a log transformed variable would be appropriate for a continuous variable that was right skewed; or a logistic regression for a binary variable; or a Poisson regression for a count variable. Drechsler, Bender, et al. (2008) and Reiter (2005a) synthesise variables with a combination of logistic, multinomial logistic and linear regressions. Kinney et al. (2011) use linear regression models with a one year lag to sequentially generate values for a variable that was measured repeatedly over several years. Pistner et al. (2018) use quantile regression to synthesise variables with heavy tails. Sakshaug and Raghunathan (2010) use a hierarchical regression model to synthesise spatial data. To introduce hierarchical structure, they stratify the data and independently fit regression

models to each stratum. Then the coefficients from the stratified regressions are modelled as multivariate normal, conditioning on group level covariates.

Non-parametric models, such as classification and regression trees (CART) and support vector machines (SVMs), have also been successfully applied within the sequential framework (Drechsler, 2010; Drechsler & Reiter, 2010; Reiter, 2005b). In contrast to regression approaches for data synthesis, there are not many design considerations when specifying a CART synthesis model. For example, to generate synthetic data with regression, an appropriate distribution must be specified for each variable. This can be time-consuming and, for some variables, a distribution that fits well may not even exist. G. Raab et al. (2016) found that regression models were inappropriate for the synthesis of some variables. The real dataset contained two age variables, measured at 10 year intervals, and there was also an interaction effect between sex and marital status. When neither the 10 year difference of the age variables or the relationship between sex and marital status were explicitly specified, the CART data was able to replicate both relationships, while the regression synthesised data failed to replicate either.

The results of the few studies that compare non-parametric models for synthesis favour CART. Drechsler and Reiter (2011) replace the quasi-identifying variables in samples from a census dataset with synthetic values that are generated from CART, bagging, random forest and SVMs. The quality of data generated from each method is evaluated by comparing the distributions of descriptive statistics and the coefficients of regression models. We discuss this method of synthetic data evaluation in Section 5.4.1. Of the four methods, the 95% confidence intervals for the CART synthesised datasets have the highest coverage probability, followed closely by SVM, with a large drop off in the coverage probability for both of the ensemble methods of synthesis. Caiola and Reiter (2010) carry out a similar comparison, replacing the quasi-identifying variables in a census dataset with synthetic values that are generated from CART and random forest models. They evaluate the quality of the synthetic datasets by fitting regression models to each and comparing which model's coefficients are closer to the non-synthetic data. Results are inconclusive, some coefficients are closer for the random forest data and others for the CART data. The results of both studies are somewhat surprising, given that we would expect the ensemble methods to outperform the single CART model at a standard prediction task. Given that these are only two examples, and both evaluate the models in the context of incompletely synthesised data, we can't be sure that the results are generalisable to either completely or incompletely synthesised datasets. Furthermore, neither study carries out any parameter tuning for the synthesis models, which may improve the performance.

Stratification can be implemented as a means of improving the quality of synthetic data or reducing the computational complexity of the synthesis model. Alfons et al. (2011) and Drechsler and Reiter (2011) stratify their data before synthesising each strata inde-

pendently, but they do not compare the results of stratification and non-stratification. A stratification approach is also applied by Yu et al. (2017), they replace geographical indicator variables with synthetic values that are generated from increasingly complex synthesis models. The utility of synthetic data from each model is evaluated by comparing the % difference and 95% confidence interval overlap for sample statistics and the coefficients of regression models. Synthetic data performance for an inference task (see Section 5.6) is evaluated by comparing the goodness of fit statistics for regression models. The results of these evaluations demonstrate that, by first stratifying patients based on race, ethnicity and cancer stage, data utility can be improved without major compromises to privacy. G. M. Raab et al. (2017) stratifies the data by a high cardinality categorical variable in order to reduce the computational complexity of sequential synthesis.

There are several software packages available that can generate synthetic data with a sequential approach. These include, the standalone software *IVEware* (T. Raghunathan et al., 2016), the R package *mice* (van Buuren & Groothuis-Oudshoorn, 2011) and the R package *synthpop* (Nowok et al., 2016).

Chapter 5

Literature Review: assessing the utility of synthetic data

In this chapter, we consider methods for assessing the utility of synthetic data.

Definition 5.1 (Utility). Consider a synthetic dataset that is substituted in place of real data for some purpose. A useful synthetic dataset is one that can be substituted with minimal difference in the process or outcomes. The smaller the difference, the higher the utility of the synthetic dataset.

Snoke et al. (2018) group methods for assessing the utility of synthetic data into two broad categories, general and specific utility. General utility measures aim to quantify the similarity between the distributions of synthetic and real data. Synthetic data with high general utility may be relatively close in “distance” to real data. However, there is no guarantee that the measure of distance captures the aspects of the data that correspond to being useful. Specific utility methods focus on particular properties or potential uses of data that a user would find desirable or useful. Examples of methods of specific utility include sample statistics and determining how the synthetic data performs as a substitute for real data in a particular task.

Ultimately, the differences between general and specific utility methods are fairly arbitrary and distinctions between the two often can be blurred. We do not find this classification of general or specific utility to be particularly helpful. As such, in our review of utility assessment methods we categorise them by methods that are similar. In this chapter, we review utility assessments from the literature of the following categories, measures of distributional similarity (Section 5.1), model based discriminators (Section 5.2), plots (Section 5.3), sample statistics (Section 5.4), qualitative feedback (Section 5.5), and task performance (Section 5.6).

5.1 Measures of distributional similarity

We can view utility (Definition 5.1) through the lens of information loss, where the utility of a synthetic dataset is inversely related to the information loss from approximating the real data distribution with the synthetic data distribution. Various measures are used to quantify this information loss or equivalently, the statistical distance between the two distributions. Collectively, we will refer to these as measures of distributional similarity. In the synthetic data literature, the Kullback-Leibler divergence is probably the most widely used of these measures, see e.g., (Debnath et al., 2021; Goncalves et al., 2020; Karr et al., 2006).

Let P and Q be continuous random variables with probability density functions p and q . Then, the KL divergence from P to Q is defined as

$$D_{KL}(P \parallel Q) = \int_{\mathbb{R}^k} p(\mathbf{x}) \log \left(\frac{p(\mathbf{x})}{q(\mathbf{x})} \right) d\mathbf{x}. \quad (5.1)$$

The KL divergence can be difficult to calculate for larger k . Although, if the data is multivariate Gaussian then a closed-form solution exists, see (Duchi, 2014).

Other measures of distributional similarity, which have been applied to synthetic data are earth mover’s distance (Wiese et al., 2019) and log-likelihood (Collier et al., 2021). However, these are also difficult to compute for higher dimensional data. Given the difficulty of computing these distributional similarity metrics for multivariate data, it is common to instead compute them for the marginal distributions (Collier et al., 2021; Debnath et al., 2021; Goncalves et al., 2020; Wiese et al., 2019). While this does significantly ease computation, it comes at the cost of losing all information about the multivariate relationships in the data.

When similarity metrics do not consider dependency structures in the data, analysts can instead take a pragmatic approach and use plots to verify that dependency structures are preserved in multivariate synthetic datasets. In Section 5.3, we describe examples of plots for checking multivariate and conditional relationships in synthetic datasets. Alternatively, there are simple measures, such as *pairwise correlation difference*, that compare aspects of the multivariate relationships between datasets.

Pairwise correlation difference (PCD) is a single number summary, which measures the similarity of the correlations of two datasets (Goncalves et al., 2020). They define PCD as follows:

$$pcd(\mathbf{X}_o, \mathbf{X}_s) = \|\text{cor}(\mathbf{X}_o) - \text{cor}(\mathbf{X}_s)\|_F, \quad (5.2)$$

where $\|\cdot\|_F$ is the Frobenius Norm (Golub & Van Loan, 1996, p. 55) and

$$\text{cor}(\mathbf{X}) = \begin{bmatrix} \rho_{11} & \cdots & \rho_{1p} \\ \vdots & \ddots & \vdots \\ \rho_{p1} & \cdots & \rho_{pp} \end{bmatrix},$$

is the matrix of Pearson correlation coefficients for pairs of columns in the $(n \times p)$ matrix of continuous variables \mathbf{X} ,

$$\rho_{ij} = \frac{\text{Cov}(\mathbf{x}_i, \mathbf{x}_j)}{\sigma_i \sigma_j}. \quad (5.3)$$

PCD can quickly highlight large differences in the correlations of synthetic data and real data. Information is lost when the correlation matrix is distilled down to a single number. Furthermore, its use as a utility measure is limited because the Pearson correlation, Equation (5.3), is not appropriate for categorical variables or non-linear relationships between variables.

5.2 Model based discriminators

The next class of utility assessments use models that try to distinguish between real and synthetic data. The idea is that when the synthetic data has a similar distribution to the real data, the discriminator model should struggle to distinguish between the two.

The use of model-based discriminators is commonplace in the machine learning literature. The Inception score method (Salimans et al., 2016) uses the Inception classifier (Szegedy et al., 2015) to assess image-generating GANs. The GANs are scored on the confidence and diversity of the Inception classifier’s (Szegedy et al., 2015) predictions. Debnath et al. (2021) and X. Zhang et al. (2018) both applied a model-based discriminator to the problem of evaluating the quality of images generated by GANs. In each paper, a neural network was trained to classify a sample of real and GAN-generated images as real or synthetic. They formulate this as a hypothesis test by calculating the Jensen-Shannon divergence between the distribution of discriminator predictions and a Bernoulli(0.5) distribution.

One of the first examples of applying model-based discriminators to synthetic data is propensity score (Woo et al., 2009). The propensity score is used throughout the synthetic data literature, see, e.g., (Bowen & Snoke, 2020; Oganian, 2014; Oganian & Domingo-Ferrer, 2017; Pistner et al., 2018; Snoke et al., 2018). Propensity scores are calculated by training a model to discriminate between real and synthetic data and then scoring the synthetic data based on how well the model discriminates.

Following the definitions of (Bowen & Snoke, 2020; Oganian, 2014; Oganian & Domingo-Ferrer, 2017; Pistner et al., 2018; Snoke et al., 2018), the matrix that contains $n = n_o + n_s$

rows \mathbf{x}_i is written as

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_o & \mathbf{X}_s \end{bmatrix}^T,$$

where

$$\mathbf{i} = \begin{bmatrix} i_1 & \cdots & i_{n_o} & i_{n_o+1} & \cdots & i_{n_o+n_s} \end{bmatrix}^T,$$

is a vector of indicators for whether observations are real or synthetic

$$i_i = \begin{cases} 0, & \mathbf{x}_i \in \mathbf{X}_o, \\ 1, & \mathbf{x}_i \in \mathbf{X}_s. \end{cases}$$

In addition, $f : \mathbf{X} \rightarrow \mathbf{i}$ is called the *discriminator*, a binary classifier that outputs a probability that each \mathbf{x}_i is synthetic

$$p_i = P(i_i = 1 | \mathbf{X}).$$

The propensity scores for each observation are the predicted probabilities from the trained model

$$\hat{p}_i = f(\mathbf{x}_i),$$

from which the single number summary for evaluating the utility of a synthetic dataset is calculated as

$$\text{pMSE}(\mathbf{X}_o, \mathbf{X}_s) = \frac{1}{n} \sum_{i=1}^n \left(\hat{p}_i - \frac{n_s}{n} \right)^2. \quad (5.4)$$

If a discriminator model can distinguish between the original and synthetic data, then the propensity scores will be close to zero or one and pMSE will be larger. On the other hand, if a discriminator model struggles to distinguish between the real and synthetic observations, then propensity scores will be close to n_s/n and pMSE will be smaller. It is important to note that, while pMSE will be low when the original and synthetic datasets have high similarity, this can also occur when the discriminator is poorly specified and unable to detect differences in the data. For example, pMSE is trivially optimised by the discriminator

$$\mathbf{i} \sim \text{Bernoulli} \left(\frac{n_s}{n} \right).$$

Woo et al. (2009) suggest improving the quality of the discriminator by including second-order and third-order terms of variables. For any dataset with a reasonable number of variables, this may require fitting an extremely large regression model and still may not result in a good discriminator. Bowen and Snoke (2020) note that different discriminators will consider different aspects of distributional similarity and recommend comparing the results from multiple discriminator models.

Small pMSE values can indicate high utility. However, interpreting “small” pMSE val-

ues is problematic for several reasons. First, the definition of small is relative. Second, pMSE is minimised when the original and synthetic data are identical which would generally be highly undesirable for synthetic data. To address difficulties with interpreting pMSE, Snoke et al. (2018) propose normalising by dividing the pMSE by a null value

$$\text{pMSE}_{\text{ratio}}(\mathbf{X}_o, \mathbf{X}_s) = \frac{\text{pMSE}(\mathbf{X}_o, \mathbf{X}_s)}{\text{pMSE}_{\text{null}}(\mathbf{X}_o)}. \quad (5.5)$$

If the discriminator is a logistic regression model with $k - 1$ degrees of freedom, then the null pMSE has a chi-squared distribution

$$\text{pMSE}_{\text{null}} \sim \frac{(n_o/n)^2 (n_s/n)}{n} \chi_{k-1}^2, \quad (5.6)$$

with expected value

$$\mathbb{E}[\text{pMSE}] = \left(\frac{k-1}{n}\right) \left(\frac{n_o}{n}\right)^2 \left(\frac{n_s}{n}\right),$$

and standard deviation

$$\text{SD}[\text{pMSE}] = \frac{\sqrt{2(k-1)}}{n} \left(\frac{n_o}{n}\right)^2 \left(\frac{n_s}{n}\right).$$

For discriminator models where the number of fitted parameters is not fixed, such as CART, Bowen and Snoke (2020) suggests the following. First, estimate the mean and standard deviation by repeatedly generating pairs of data by sampling the original data with replacement. Then, calculate the pMSE, as given in Equation (5.4), for those pairs.

Algorithm 1 Bootstrapped $\text{pMSE}_{\text{null}}$ estimate (Bowen & Snoke, 2020)

```

1: Inputs  $n_r, \mathbf{X} = [\mathbf{x}_1 \ \dots \ \mathbf{x}_n]^T$ 
2: for  $r : 1, \dots, n_r$  do
3:    $\mathbf{X}_1, \mathbf{X}_2 \leftarrow \text{sample}(\mathbf{X})$ 
4:    $\text{pMSE}_{(r)} \leftarrow \text{pMSE}(\mathbf{X}_1, \mathbf{X}_2)$ 
5: end for
6:  $\mathbb{E}[\text{pMSE}_{\text{null}}] \leftarrow \frac{1}{n_r} \sum_r \text{pMSE}_{(r)}$ 
7:  $\text{SD}[\text{pMSE}_{\text{null}}] \leftarrow \sqrt{\frac{1}{n_r} \sum_r (\text{pMSE}_{(r)} - \mathbb{E}[\text{pMSE}_{\text{null}}])^2}$ 
8: return  $\mathbb{E}[\text{pMSE}_{\text{null}}], \text{SD}[\text{pMSE}_{\text{null}}]$ 

```

pMSE ratios that are greater than one indicate that the discriminator model has better performance discriminating between \mathbf{X}_o and \mathbf{X}_s than between two samples of real data. pMSE values are only comparable if the discriminator model is the same. For a CART discriminator, this means that the complexity parameter is fixed. As such, Bowen and Snoke (2020) use cross-validation to choose the complexity parameter of the CART model used for all pMSE and $\text{pMSE}_{\text{null}}$ values.

5.3 Plots of data distributions

Another method for assessing utility is to compare the univariate distribution variables from the real and synthetic datasets. This comparison can highlight when the distributions of synthetic variables are wrong. For numerical variables, histograms (Choi et al., 2017), cumulative distribution function (CDF) plots and boxplots (Alfons et al., 2011; Lin et al., 2019; Templ & Alfons, 2010) have been used to compare distributions. For categorical variables one can compare the proportions of each level in the real and synthetic data (Choi et al., 2017; Drechsler, 2010; Kinney et al., 2011).

For multivariate data, univariate plots do not provide a complete picture. Hence, conditional distributions are frequently plotted to determine whether relationships between variables have been preserved. Alfons et al. (2011) show how relationships between pairs of variables are preserved in synthetic data by using mosaic plots and plotting the relative differences between coefficients of contingency tables. Note that both can be difficult to interpret for more than a few variables. Beaulieu-Jones et al. (2019) plot pairwise Pearson correlation coefficients to check whether strong relationships between pairs of variables are preserved. However, the interpretability of these plots also scales poorly with the number of variables. Hu and Hoshino (2018) and Hu and Savitsky (2021) plot the densities of the differences in relative frequencies for all one-way, two-way, and three-way cross-tabulations. The plots show larger peaks at zero when the relative frequencies between the real and synthetic cross-tabulations are close. Manrique-Vallier and Hu (2018) also plot comparisons of cross-tabulations. However, they consider uncertainty by comparing the mean 95% interval coverage for the three-way proportions.

For time series data, plotting the real and synthetic variables against time can show how well the data synthesis model captures temporal variations (Frigerio et al., 2019; Kinney et al., 2011). Lin et al. (2019) use auto-correlation plots to show that their synthesis model captures both weekly and annual variations in the data. In addition, they plot the distribution of sequence length to show that their model can capture that aspect of the data.

Choi et al. (2017) introduce dimension-wise prediction (DWP), a flexible method for assessing how well conditional relationships between binary variables are preserved in synthetic data. Each variable is regressed on all others in the dataset and the prediction scores between the real and synthetic data were compared using F1 Score. Slight variations of the same approach have been implemented elsewhere. For example, Xie et al. (2018) score predictions using the area under the receiver operating characteristic (AUROC) curve instead of F1 score. Sometimes these approaches are referred to as *feature prediction*. Extension to other data types requires the specification of an appropriate loss function for that data type.

We can formally describe the general procedure of DWP for arbitrary variable distri-

butions as follows.

Definition 5.2 (Dimension-wise prediction). Let \mathbf{O} be the $(n_o \times p)$ original dataset, \mathbf{T} be a $(n_t \times p)$ dataset such that

$$\mathbf{O}, \mathbf{T} \subset \mathcal{D} : \mathbf{O} \cap \mathbf{T} = \emptyset,$$

are disjoint samples from the population \mathcal{D} and

$$\mathbf{S} = f(\mathbf{O}),$$

is the synthetic dataset, generated with the model $f: \mathbb{R} \rightarrow \mathbb{R}$.

Identical prediction models are trained to predict the j^{th} variable on the original and synthetic datasets respectively

$$g_j: \mathbb{R}^{n_o \times (p-1)} \rightarrow \mathbb{R}^{n_o},$$

and

$$h_j: \mathbb{R}^{n_s \times (p-1)} \rightarrow \mathbb{R}^{n_s}.$$

Test error is evaluated for each model

$$\begin{aligned} c_{oj} &= l_j(\mathbf{t}_j, g_j(\mathbf{T}_{\setminus j})), \\ c_{sj} &= l_j(\mathbf{t}_j, h_j(\mathbf{T}_{\setminus j})), \end{aligned}$$

where l_j is an appropriate loss function. c_{oj} and c_{sj} are plotted against each other for all p . Variables that are similarly well predicted by both models will lie close to the line of equality.

We have already discussed some examples of loss functions for categorical variables. For non-categorical variables, possible loss functions include root mean squared error for continuous variables or Poisson log-likelihood for count variables.

Similar to model based discriminator methods, DWP also relies on a well specified prediction model. As such, good performance on DWP can indicate well-synthesised data, but it can also indicate a prediction model that cannot model the differences between the conditional distributions. Torfi and Fox (2020) repeat DWP with four prediction models, which increases the likelihood that at least one prediction model is well specified. However, this is by no means perfect. It is entirely possible that all of the prediction models are poorly specified. Note, the flexibility of the models that can be specified for DWP helps to address several of the limitations of PCD (see Section 5.1).

Comparing plots is inherently qualitative, which brings both advantages and disadvantages. They do not allow for direct and objective comparison of datasets such as measures

of distributional similarity (Section 5.1). Many of those measures average over the data, making identifying poorly synthesised aspects of the data difficult. In contrast, the plots can display a lot of information in a format that is easy to interpret. This is invaluable information, especially during the development of the synthesis model. We believe that many of the measures of distributional similarity would be more useful if they were not averaged over all variables, but were instead plotted.

5.4 Sample statistics

Comparison of the sample statistics for real and synthetic datasets is used as a basic verification in countless examples; most commonly, means or regression coefficients are compared (Drechsler, 2011a, 2018; Drechsler, Bender, et al., 2008; Drechsler, Dundler, et al., 2008; Drechsler & Reiter, 2010; Reiter, 2002; Yu et al., 2017). In another example, standard deviations, tail values, and the percentage of zeros are compared for different stratum (Alfons et al., 2011).

There are endless potential choices of statistics that capture important aspects of the data. Many interesting examples are contained in the synthetic survey data literature. Sakshaug and Raghunathan (2010) synthesise spatial data and compare the means and standard deviations of variables within each spatial unit. For each variable, the observed point estimates for each real and synthetic spatial unit are regressed against each other. A slope close to zero indicates a strong linear correspondence between the point estimates of that variable for each area. In another example of synthesising spatial data, distributions of the geographical indicators for respondents with different combinations of gender, income, and age are compared (Hu & Savitsky, 2021). Other interesting examples from synthetic survey literature include “the % of people who are divorced and have child support or social security payments” (Reiter, 2005a), “the percentage of households with an income over \$200,000” (Reiter, 2005b), “the proportion of households with two workers” (Hu et al., 2016), “incidence rates (of cancer) by race/ethnicity” (Yu et al., 2017), and “percentage of patients with systolic blood pressure above a target value and that had a medication added” (Beaulieu-Jones et al., 2019).

The choices of statistics should reflect the complexities of the data. For example, whether trends are preserved in synthetic time series data and in an assessment of synthetic stock data, authors compared the values of the DY metric, ACF and leverage scores, which are measures of the expected returns and correlation of returns over time (Wiese et al., 2019). In another example of synthesising time series data, the synthetic and real data were compared by computing the mean squared error of the difference between the auto-correlations of each sample (Lin et al., 2019).

5.4.1 Sample statistics: methods of comparison

Various methods have been developed for comparing synthetic data sample statistics with real data sample statistics. The naïve approach of comparing the differences in point estimates does not contextualise the scale or account for the uncertainty. Methods that account for the scale include the relative percentage difference of point estimates (Oganian & Domingo-Ferrer, 2017) and the ratio of point estimates (Taub et al., 2020). Both are undefined when the denominator is zero and neither accounts for uncertainty.

To account for uncertainty while comparing sample statistics, uncertainty interval overlap is the standard approach (Karr et al., 2006). Let (l_{ok}, u_{ok}) and (l_{sk}, u_{sk}) be the uncertainty interval bounds of the k^{th} statistical quantity that is calculated on the original and the synthetic data respectively. For synthetic datasets, proper computation of the uncertainty intervals will require an appropriate variance estimator (see Section 4.2.2).

The overlap of these intervals, (l_{ik}, u_{ik}) , has the lower and upper bounds

$$l_{ik} = \max(l_{ok}, l_{sk}) \quad \text{and} \quad u_{ik} = \min(u_{ok}, u_{sk}),$$

where $u_{ik} \geq l_{ik}$.

The relative interval overlap for k is

$$J_k = \frac{1}{2} \left(\frac{u_{ik} - l_{ik}}{u_{ok} - l_{ok}} + \frac{u_{ik} - l_{ik}}{u_{sk} - l_{sk}} \right), \quad (5.7)$$

where $J_k = 1$ when the confidence intervals are equal and $J_k = 0$ when they are disjoint. Interpreting this for more than a few variables is complicated. A possible single-number summary for a synthetic dataset is

$$J = \frac{1}{K} \sum_{k=1}^K J_k.$$

As with the distributional similarity measures, averaging does lose information. Another option is to plot the overlaps for all k .

Uncertainty interval overlaps do not account for the scales of the estimates themselves. This can be an issue when the uncertainty of an estimate is small in comparison to the estimate. In such cases, the relative interval overlap may be low despite the close point estimates. When this occurs, comparing the ratio of point estimates may be more appropriate. While it does not account for uncertainty, that is far less problematic when the uncertainty is small.

One critique of uncertainty interval overlaps is that they can be hard to interpret when the intervals overlap due to becoming negative (Barrientos et al., 2023). Negative interval overlaps arise from computing Equation (5.7) when $u_{ik} < l_{ik}$. In actuality, the interval

overlap should be undefined for $u_{ik} < l_{ik}$, although one could reasonably argue that the overlap should be zero. While we do not wholly agree with this premise, interpreting the interval overlap for non-overlapping intervals is difficult. The real problem with interpreting non-overlapping intervals is that there is no way to differentiate between two sets of non-overlapping intervals. Irrespective of how close the point estimates may be, both overlaps will be undefined or zero.

We are not aware of any examples of methods in the synthetic data literature that compare sample statistics and consider both their uncertainty and scale. There are methods that compute the uncertainty for a ratio of two sample statistics that could be explored. These include bootstrapping, Bayesian simulation, and Fieller's theorem (Fieller, 1954).

5.5 Human feedback

Human feedback refers to any utility assessment that relies on quantitative feedback. These assessments are beneficial when utility is easy to see but difficult to measure. Examples of such situations include identifying whether an image is clear or looks like a real thing and identifying that a sentence is semantically correct. Examples of using the eye test as a measure of quality can be found throughout the literature for image, video and text generation (Guo et al., 2017; Saito et al., 2020; Sauer et al., 2021).

Flaws in tabular data may not be as easily perceptible as in text, images, or video, but expert opinion can still provide a measure of quality. Beaulieu-Jones et al. (2019) and Choi et al. (2017) both ask doctors to rate the realism of real and synthetic samples. While this type of quality check may not pick up on the details of complicated relationships between variables, for some uses of synthetic data not all relationships need to be well preserved. For example, if the data were to be used as a teaching aid for medical students who need to see what patient databases look like, the eye test may be an acceptable quality check.

5.6 Task performance

The final approach for assessing the utility of a dataset that we discuss in this chapter is task performance. In this approach the performance of the synthetic data at a relevant task is compared to the performance of the real data at the same task. An exceedingly common example of this type of assessment in the synthetic data literature is to fit a regression model that is known to work well on the real data to both the real and synthetic data. Then compare the 95% confidence interval overlap in Equation (5.7) of all regression coefficients (Caiola & Reiter, 2010; Drechsler, 2010, 2011a, 2018; Drechsler, Bender, et al., 2008; Drechsler, Dundler, et al., 2008; Drechsler & Reiter, 2010; Hu & Hoshino, 2018; Hu et al., 2014; Hu & Savitsky, 2021; Karr et al., 2006; Reiter, 2002; Sakshaug & Raghunathan,

2010; Taub et al., 2020; Yu et al., 2017).

Both confidence interval overlap and relative percentage difference can give an idea of how the inference results are similar. However, neither can tell you with certainty that the results of hypotheses tests on real and synthetic data will give the same results. In an attempt to address this problem, Taub et al. (2020) use *severity ratings*, which are a more subjective measure that compare whether conclusions made from an inference on real data would still be made if the inference had been carried out on a particular synthetic dataset.

In the machine learning literature, the synthesised datasets are often associated with a prediction task. The standard approach is to train a prediction model on both synthetic and real datasets and then compare the predictive performance of each of those models on a real test set (Beaulieu-Jones et al., 2019; Frigerio et al., 2019; Jordon et al., 2019; Lin et al., 2019; Torfi & Fox, 2020). This approach, sometimes called “train on synthetic, test on real”, mirrors a common motivation for using synthetic data for many analysts and researchers. They are interested in making predictions about the real world but they are not able to access the data that is necessary to train a predictive model. For time series data, predictive performance can be assessed by interpolating observations and predicting future steps or entire sequences (Che et al., 2018; Debnath et al., 2021). Frequently, comparisons are repeated for multiple prediction models. This repetition can guard against the prediction model biasing the results in favour of a particular data synthesis method.

X. Zhang et al. (2018) compare the performance of an image classifier trained on a small set of real images and the performance of the same classifier trained on a larger dataset comprised of a small amount of real and a large amount of synthetic images. They encountered some difficulties with training the model on the larger dataset. However, by slowly introducing the synthetic images, they saw a significant improvement of the classifier, especially when the sample of real images was small. Our concern with introducing synthetic data in this manner is that because the image classifier’s training data includes synthetic images that it labelled itself, the training data will end up containing a higher proportion of labels than the classifier already tends to predict. Presumably, the labels that the classifier prefers to predict tend to be common in the real data, in which case the classifier’s predictions could become biased towards the most common labels. X. Zhang et al. (2018) assess the classifier proportion with accuracy, which could be improved by predicting the common labels. Recent research has found that introducing generated data into the training data for generative models can reduce the quality of generated data (Shumailov et al., 2023). While not all of their reasons for this phenomenon apply to the training of an image classifier, the synthetic samples are an imperfect approximation of the real data.

While most prediction assessments for synthetic data test performance on real-world data, there are some scenarios where predictive performance on synthetic data is the goal,

such as synthesising data to be used in challenges or to teach students. In these cases, it may be preferable to compare the predictive performance of a model trained and tested on real data with one trained and tested on synthetic data. For such a comparison, the rankings of the prediction models can be more important than the actual scores of each model. Jordon et al. (2019) and Lin et al. (2019) both assess data in this manner. They use Spearman's rank correlation coefficient to score how well the predictions of each model align when trained on the synthetic and real datasets.

With many utility assessments, it is not clear that good performance on the assessment will correspond to high utility. Directly testing the performance of a dataset in the scenario in which it is intended to be used is a very effective method of evaluating utility. Performance evaluation tends to be difficult for more complex tasks, so task based assessments will often be more qualitative.

Chapter 6

Literature Review: Assessing disclosure risk of synthetic data

In this chapter, we consider the assessment of disclosure risk for synthetic and SDC datasets. Expanding on the disclosure concepts that we introduced in Chapter 3, we discuss methods for assessing the three types of disclosure risk when a motivated intruder (Definition 3.1) is attempting to learn information about targets in a synthetic dataset.

6.1 Identity disclosures

Recall from Section 3.1 that identity disclosures are not applicable to completely synthetic data due to the lack of a clear mapping from the synthetic observations to individuals in the population. However, despite the focus on completely synthetic data in this thesis, identity disclosure risk is frequently assessed for incompletely synthesised data and other SDC methods preserving the identity of real observations.

Reiter and Mitra (2009) introduce a general framework that uses a Monte Carlo approach for estimating the probability of an identity disclosure. They establish a prior distribution for assumptions of the intruder in order to account for uncertainty in the assumptions. We give a specific example of this approach for incompletely synthetic data in Example 6.1. For other examples, see, e.g., (Caiola & Reiter, 2010; Drechsler, 2011a; Drechsler & Reiter, 2010; Hu & Hoshino, 2018).

Example 6.1. *Hu and Savitsky (2021) assess the risk of identity disclosures by assuming that an intruder, with prior knowledge of a subset of variables for all observations in the original data, will attempt to use that subset to match the original observations with observations in the incompletely synthetic data. They calculate three values. First, they calculate the probability that an intruder will get a correct match. Then, they calculate proportion of original observations with a unique and correct match in the synthetic data.*

Finally, they calculate the proportion of original observations with a unique but incorrect match in the synthetic data.

After calculating these values, Hu and Savitsky (2021) calculate these statistics for two non-synthetic baselines. They do this in order to contextualise the identity disclosure risk of their synthesis method. The first non-synthetic baseline is a masking procedure and serves as a “worse” utility, SDC procedure. Specifically, they replace the geographical variable with random values sampled from a uniform distribution. The second non-synthetic baseline is the unchanged original data, which serves as the “best” utility but the worst disclosure case.

Comparison to these baselines can show one of two results. One result is that synthetic data offers similar disclosure risk and better utility than alternative SDC procedures. The other result is that synthetic data offers similar utility but better disclosure risk than the real data.

6.2 Membership disclosure

A membership inference attack models the scenario in which an attacker has black-box or white-box access to a model and attempts to predict whether certain samples of data have been used to train the model. Two examples of membership inference attack models include Shokri et al. (2017) and Hayes et al. (2018). Shokri et al. (2017) presents a membership inference attack against models for when the attacker has black-box access to a model. Hayes et al. (2018) present a method for carrying out membership inference attacks that uses GANs and can be applied when an attacker has either black or white box access to the target model. Differential privacy and strong L2 regularisation are effective counters to MI attacks (Choquette-Choo et al., 2020).

Choi et al. (2017) implement a simple procedure to qualitatively assess the risk of membership disclosure. For their procedure, they assume that the intruder’s goal is to predict membership. In addition, they assume that the intruder’s prior knowledge consists of the entire synthetic dataset and a subset of k variables for real observations. Now, write \mathbf{X}_s for the synthetic dataset and

$$\mathbf{X}_r = \begin{bmatrix} \mathbf{X}_o & \mathbf{X}_t \end{bmatrix}^T$$

for the intruder’s prior knowledge. Choi et al. (2017) define the *indicator vector* as a vector consisting of 0’s and 1’s. These 0’s and 1’s denote whether a real observation in \mathbf{X}_r is a part of the training set or the test set. The indicator vector is written as

$$\mathbf{y} = \begin{bmatrix} y_1 & \cdots & y_{2n} \end{bmatrix}^T.$$

The intruder attempts to determine membership by finding observations in the real dataset,

\mathbf{X}_r , that are close to observations in the synthetic dataset \mathbf{X}_s . They do this by classifying each observation as part of the training set if the observation is within the distance threshold, \hat{d} , or classifying it as part of the test set if the observation is outwith \hat{d} . Mathematically, the prediction for membership is written as follows:

$$\hat{y}_i = \begin{cases} 1, & \exists \mathbf{x}_{sj} \in \mathbf{X}_s : d(\mathbf{x}_{ri}, \mathbf{x}_{sj}) < \hat{d}, \\ 0, & \nexists \mathbf{x}_{sj} \in \mathbf{X}_s : d(\mathbf{x}_{ri}, \mathbf{x}_{sj}) < \hat{d}, \end{cases}$$

where $i = 1, \dots, 2n$ and

$$d : \mathbb{R}^{(2 \times k)} \rightarrow \mathbb{R}^+$$

where d is the measure of distance between observations.

Both the real and synthetic datasets of Choi et al. (2017) consist entirely of binary variables. Therefore, they calculate the distance between observations using Hamming distance. Note that any appropriate measure of the distance between two observations from a dataset can be used. They score their membership disclosure predictions for the distance threshold, \hat{d} , with precision and recall, written as,

$$\text{precision}_{\hat{d}}(\mathbf{y}, \hat{\mathbf{y}}) \text{ and } \text{recall}_{\hat{d}}(\mathbf{y}, \hat{\mathbf{y}}).$$

They calculate the precision and recall for each \hat{d} . Then, they plot these precision and recall scores and qualitatively compare the precision-recall curves to see the disclosure risk.

6.3 Attribute disclosure

One example of assessing attribute disclosure is explored in Reiter (2005a) and is discussed in the following example, Example 6.2.

Example 6.2. *Reiter (2005a) assess the attribute disclosure risk of a household income variable. To do this, first, they calculate the prediction error of using the synthetic income values instead of the observed income values. Then, they compare the prediction error with the standard deviation of the observed income values. They find that the two values are similar. Hence, they conclude that the attribute disclosure risk of releasing the synthetic income values is comparable to the release of the average household income. In addition, they point out that, from a sample of approximately 50,000, this would be reasonable information to release.*

Another example of assessing attribute disclosure is given in Yu et al. (2017) and is discussed in the following, Example 6.3.

Example 6.3. *Yu et al. (2017) assess the attribute disclosure risk of an incompletely synthesised database of cancer patients. In their synthetic data, they replace geographical indicators with synthetic values. They do not change variables containing patient demographic information or information about their cancer.*

Yu et al. (2017) consider two scenarios. In one scenario an intruder obtains a single replication of synthetic data. In the other scenario an intruder obtains multiple synthetic replications and treats the most common replacement as the truth. They calculate two measures to assess the risk of these scenarios. One measure is the % of subjects for which the synthesised variable is the wrong value. The other measure is the % of correct guesses when the intruder guesses that the attribute value for a subject is the most common value in the synthetic replications. Then they compare these measures for various synthetic data generation models and other SDC methods. Note that their methodology relies on the existence of a one-to-one correspondence between real and synthetic data. As such, it is only relevant for incompletely synthesised data.

Hu and Savitsky (2021) implement a similar assessment to the one described in Example 6.3. Their approach is discussed in Example 6.4.

Example 6.4. *Hu and Savitsky (2021) study the attribute disclosure risk of an incompletely synthetic dataset where County labels are synthesised and other demographic variables remain unchanged. We have several critiques of their methodology.*

Our first critique is that they treat all correct predictions of all geographic regions as equal despite the fact that we would expect more populous regions to be easier to guess. “Census tracts generally have a population size between 1,200 and 8,000 people, with an optimum size of 4,000 people” (U.S. Census Bureau, 2022). This is a considerable range that needs to be accounted for, but we do not know the specifics of the regions in either study so perhaps they were all close in population size and the assumption is reasonable. They also do not consider the possibility that the intruder is aware that the synthesised variables are synthetic and may be wrong. Perhaps an intruder aware of this possibility would also know that the synthesis model may take into account the spatial relationships in the data and would widen their predictions to include regions that are nearby to the values in the synthetic dataset.

Another significant issue with the methodology of Hu and Savitsky (2021) is that they assessed attribute disclosure risk for the geographic region. In fact, the geographic region is a quasi-identifier, not a sensitive variable in the data. The risk posed by a correct prediction of geographic region is that it could lead to a successful linkage attack where the intruder learns truly sensitive information. A better assessment of the disclosure risk of sensitive variables would be to directly evaluate the intruder predictions for those sensitive variables. A motivated intruder could know some demographic information about their targets. As such, it would have been a reasonable scenario to evaluate the probability of

predicting either the cancer diagnostic factors or household income from the demographic information and synthesised geographic region variable.

Choi et al. (2017) assumes that a potential intruder, who has access to a completely synthesised dataset and subset of variables from the original data, will predict attribute values using a k -nearest neighbours model trained on the synthetic dataset. They score intruder predictions using precision and recall scores. In addition, they explore the effect of different numbers of neighbors and synthetic dataset sizes on the quality of predictions. The size of the synthetic dataset does not have any effect on the risk of an attribute disclosure and $k = 1$ is the most effective strategy for the intruder. The precision and recall scores do not provide much information about the disclosure risk in a vacuum, although a qualitative assessment of different synthetic or SDC datasets could be made through a comparison of their precision and recall curves.

Notice that there is a general lack of consistency in the choices for intruder prior knowledge, prediction methods, and the methods of evaluating those predictions. Furthermore, that all of these implementations assess the disclosure risk for each sensitive attribute independently. Both of these issues are noted by Reiter (2023), who criticises the “ad hoc” nature and failure to quantify risk of most assessments of attribute disclosure risk.

6.3.1 A Bayesian approach

Reiter et al. (2014) introduces a general framework for assessing the risk of attribute disclosure in both completely and incompletely synthesised data. They formalise the predictions of the sensitive attributes for a subject \mathbf{x}_i , by a hypothetical intruder, as a Bayesian posterior distribution

$$P(X_i = \mathbf{x}_i | Z, A, S) \propto P(Z | X_i = \mathbf{x}_i, A, S) P(X_i = \mathbf{x}_i | A, S),$$

conditioned on the synthetic data Z , and the intruder’s prior beliefs for the real data A and the synthesis model S .

Their conservative prior assumptions are that the intruder knows the synthesis model and all sensitive values for all observations in the real data except for their target

$$A = X_{-i} = \bigcup_{j \neq i} \mathbf{x}_j.$$

These assumptions are unrealistically strict. However, the resulting posterior distribution can be considered a worst case for the probability of the intruder correctly predicting the sensitive attributes. In theory, this approach is ideal for the assessment of attribute disclosure risk. While the need to make intruder assumptions is not eliminated, the uncertainty of these assumptions can be reflected in the prior and likelihood distributions.

In practice, calculation of the posterior distribution is often infeasible. If there are too many sensitive attributes, the support of X_i quickly becomes too large to search. If the number of observed combinations \mathbf{x}_i is too large or the synthesis model too slow to fit, then the computation of

$$P(Z|X_i = \mathbf{x}_i, X_{-i}, S), \quad (6.1)$$

is computationally prohibitive for all i .

The issues computing the likelihood in Equation (6.1) can be overcome for Bayesian synthesis models. This is achieved with importance sampling and restricting the sample space X_i to only combinations that are close to the truth (Hornby & Hu, 2021; Hu et al., 2014). However, Manrique-Vallier and Hu (2018) found that the estimator is still too unstable to be sampled from and variance of the posterior is very high.

In theory, the framework of Reiter et al. (2014) addresses all of the criticisms that Reiter (2023) makes of other examples of attribute disclosure assessment in the literature. However, the current limitations are quite prohibitive. If the goal is to develop a standardised approach for disclosure assessment, then it stands to reason that it must be widely applicable. If the framework is to become the standardised approach to disclosure risk, then the inability to apply it to high-dimensional data and many synthesis models must be overcome.

6.3.2 An empirical approach

Elliot (2014) introduces procedures for assessing the attribute disclosure risk of numeric and categorical variables. Taub et al. (2018) formally describes these procedures for categorical variables as follows.

We assume that the intruder knows the values of some quasi-identifying variables for their target, and will attempt to predict the value of some sensitive attribute. Let \mathbf{S} be the synthetic dataset consisting of n rows, where

$$\mathbf{s}_i^T = (\mathbf{k}_{si}^T, t_{si}),$$

is a row that contains the quasi-identifying variables \mathbf{k}_{si} and the target variable t_{si} . Similarly, let \mathbf{O} be the original dataset, consisting of m rows, where the row for the intruder's j^{th} target is written

$$\mathbf{o}_j^T = (\mathbf{k}_{oj}^T, t_{oj}).$$

Then, the *correct attribution probability* (CAP) for the record \mathbf{o}_j is the empirical probability of its target variables given the key variables. We write this

$$\text{CAP}(\mathbf{o}_j, \mathbf{S}) = \frac{\sum_{i=1}^n [t_{si} = t_{oj}, \mathbf{k}_{si} = \mathbf{k}_{oj}]}{\sum_{i=1}^n [\mathbf{k}_{si} = \mathbf{k}_{oj}]}, \quad (6.2)$$

where the square brackets are Iverson brackets.

For numerical attributes, Elliot (2014) describes a slightly different method that they call “General empirical differential privacy procedure assuming a continuous target variable”. However, we refer to it as *correct attribution error* (CAE). Specifically, the correct attribution error is defined as,

$$\text{CAE}(\mathbf{o}_j, \mathbf{S}) = \left| t_{\mathbf{o}_j} - \frac{\sum_{i=1}^n t_{s_i} [\mathbf{k}_{s_i} = \mathbf{k}_{\mathbf{o}_j}]}{\sum_{i=1}^n [\mathbf{k}_{s_i} = \mathbf{k}_{\mathbf{o}_j}]} \right|. \quad (6.3)$$

Note that when the dataset \mathbf{S} contains no observations that match the target observation \mathbf{o}_j we have,

$$\sum_{i=1}^n [\mathbf{k}_{s_i} = \mathbf{k}_{\mathbf{o}_j}] = 0.$$

As such, CAP and CAE (see Equation (6.2) and Equation (6.3)) are both undefined.

Elliot (2014) proposes several suggestions to address non-matches, but none are particularly satisfactory. We will go through each suggestion and discuss the issues that we have. For observations with undefined CAP scores, Elliot (2014) suggests scoring non-matches as zero. In effect, a score of zero represents an intruder with zero probability of making a disclosure. Although, the only scenario that an intruder has zero probability of being correct, is if they do not make a prediction. Perhaps this is a reasonable assumption. If there was a time or financial cost to the intruder for making a prediction then, in the case of a non-match, it would not be worthwhile for them to make a prediction. However, if the intruder would still make a prediction in the case of a non-match then scoring non-matches as a zero will underestimate the risk of disclosure.

Another suggestion for dealing with non-matches during CAP score calculations is to ignore them (Elliot, 2014). However, this can then overestimate the risk of disclosure. Consider, that we measure the CAP scores for two datasets and find that both are equal. In the first dataset, we are able to find at least one match for every observation, whereas the second dataset contains many non-matched observations which we choose to ignore. According to the CAP scores, the disclosure risk of both datasets is equal. However, this is only because we discounted many observations that presumably have a lower disclosure risk than the average observation in either dataset. Consequently, by ignoring non-matches we have overestimated the disclosure risk for the second dataset.

For non-matches that are encountered when calculating during CAE, Elliot (2014) suggests predicting the mean value of the variable. Of the three suggestions, this is the most sensible because it does not assume that the intruder immediately gives up in the face of adversity. Furthermore, it does utilise some information that is available to the intruder. However, assuming that the intruder knows the values of multiple quasi-identifying variables, they could have instead made a guess that utilised a smaller set of

the quasi-identifiers. That would be a more informed guess than the average value.

In addition to defining the CAP and CAE measures for quantifying risk of attribute disclosure, Elliot (2014) introduces *empirical differential privacy* or *differential correct attribution probability (DCAP)*. DCAP is a method for assessing whether attribute disclosure scores represent an acceptable risk of disclosure, by comparing them against some acceptable baseline of disclosure risk. “A dataset is differentially confidential in respect of a given target and key if on average there is no difference in the CAP score for a record whether the record is in the original dataset or not” (Taub et al., 2018). In the original work of Elliot (2014), this is tested as follows.

Let f be a data synthesis model, that is used to synthesise the consecutive years of a survey dataset, \mathbf{O}_1 and \mathbf{O}_2 ,

$$\mathbf{S}_1 = f(\mathbf{O}_1) \quad \text{and} \quad \mathbf{S}_2 = f(\mathbf{O}_2).$$

For each of the synthetic datasets, the mean CAP score is calculated across all observations $\mathbf{o}_{1j} \in \mathbf{O}_1$. That is,

$$c_1 = \frac{1}{m} \sum_{j=1}^m \text{CAP}(\mathbf{o}_{1j}, \mathbf{S}_1), \quad (6.4a)$$

and

$$c_2 = \frac{1}{m} \sum_{j=1}^m \text{CAP}(\mathbf{o}_{1j}, \mathbf{S}_2). \quad (6.4b)$$

Then, the mean scores are compared. If they are indistinguishable, then \mathbf{S}_1 is differentially confidential with respect to the quasi-identifying variables.

Taub et al. (2018) propose a method of formally testing for differential confidentiality. Rather than averaging the CAP scores (as in Equation (6.4)), they instead use Welch’s t-test to test the difference between the two sets of m CAP scores. If the t-test shows that the disclosure risk is significantly lower for \mathbf{S}_2 , then we conclude that \mathbf{S}_1 is not differentially confidential.

Remark. In our opinion, a paired sample t-test would be a more appropriate choice than Welch’s t-test. When testing the difference between two sets of m samples,

$$\begin{aligned} & \{\text{CAP}(\mathbf{o}_j, \mathbf{S}_1) : j = 1, \dots, m\}, \\ & \{\text{CAP}(\mathbf{o}_j, \mathbf{S}_2) : j = 1, \dots, m\}. \end{aligned}$$

it is easy to see that this can be constructed as a set of paired observations,

$$\{(\text{CAP}(\mathbf{o}_j, \mathbf{S}_1), \text{CAP}(\mathbf{o}_j, \mathbf{S}_2)) : j = 1, \dots, m\}.$$

In which case, a paired t-test better reflects the paired nature of the data.

On the surface, testing the difference using t-tests is preferable to an informal comparison of mean values. However, in practice, proving differential confidentiality using a t-test is highly unlikely. This is due to the fact that differential confidentiality is the null hypothesis of the t-test. As such, the realistic best case scenario will be a failure to reject the null. In theory, it is possible to carry out a test where the CAP score of \mathcal{S}_1 is lower than \mathcal{S}_2 . We expect this to be difficult to achieve for synthetic datasets with reasonably high utility. If that were to be the result of a hypothesis test, we would have concerns that it was a type I error. Another problem with t-tests, is that they do not scale well when there are multiple synthetic or SDC datasets to be compared. While p-values can be adjusted for multiple comparisons, interpreting the results of large numbers of tests can be difficult. In the scenario that we compare multiple synthetic datasets that have multiple sensitive attributes, the number of comparisons could become very large.

The methods described above, assess disclosure risk by comparing the attribute disclosure risk of an observation being included or excluded from the synthetic data training set. As we have discussed, we are unlikely to prove that the disclosure risk of an observation is lower when included than excluded. As an alternative method, Taub et al. (2018) proposes the comparison of CAP scores to a baseline that represents a reasonable disclosure risk. Their choice of baseline, written as

$$\text{CAP}_{\text{base}}(\mathbf{o}_j) = \frac{1}{n} \sum_{i=1}^n [t_{oi} = t_{oj}],$$

is the risk of disclosure when the an intruder knows the average value for the sensitive attribute. As they and others have pointed out, summary statistics such as the mean are routinely presented for sensitive datasets (Reiter, 2005a; Taub et al., 2018). Therefore, if disclosure risk of synthetic data is equivalent to the release of summary statistics, than presumably it is acceptable.

It is arguable whether CAP or CAE reflects a realistic assessment of disclosure risk. Unlike other methods, e.g, Reiter et al. (2014), that make unrealistically strict assumptions about the capabilities of intruders, we are concerned that CAP and CAE do not meet the criteria of a motivated intruder.

While Elliot (2014) does not formally outline the approach of the hypothetical intruder, we can infer from Equation (6.2) and Equation (6.3) that their methods are as follows. For an observation \mathbf{o}_j^T , the intruder will identify the subset of observations in a synthetic dataset \mathcal{S} that match on all quasi-identifiers \mathbf{k} . Then, for categorical variables, they will either randomly sample or predict the most common value of t_{si} from the matching subset. While, for numeric variables, they will predict the sample mean of t_{si} for the matching subset. For the remainder of this thesis, we refer to this method of attribute prediction as

empirical matching.

Empirical matching does not utilise information about ordering or distance between values the quasi-identifying variables. As such, there are surely better methods for attribute prediction when quasi-identifying variables are numeric or ordinal variables. Even in the case that all quasi-identifiers are categorical, given the many prediction methods that are available, it is hard to believe that the empirical matching is the best prediction method. As such, it is hard to justify that the methods of Elliot (2014) and Taub et al. (2018) represent a motivated intruder.

In contrast to the Bayesian attribute disclosure risk assessment of Reiter et al. (2014), the empirical attribute disclosure risk assessment of Elliot (2014) and Taub et al. (2018) has some clear issues. First, the disclosure risk of each attribute is measured individually. This complicates the risk assessment when there are many sensitive attributes. Furthermore, all assumptions about the intruder’s prior knowledge or methods are treated as fixed. Although it is entirely possible to change the quasi-identifiers or implement a different intruder prediction method, this choice will still be fixed. While it is reasonable to argue that the prior knowledge of the intruder in Elliot (2014) is too restrictive, it is also true that we can never be certain of the intruder’s methods or prior knowledge (see discussion of Example 3.2). Therefore, representing this prior knowledge as a probability distribution allows us to reflect that uncertainty.

Elliot (2014) and Taub et al. (2018) describe a framework for quantifying the attribute disclosure risk of synthetic data. Despite the disadvantages that we have described, their framework has a significant advantage in computability over the framework of Reiter et al. (2014). As such, we believe that is currently the best available approach for a standardised method of assessing the risk of disclosure for completely synthetic data.

6.4 Disclosure risk of outliers

In Section 3.1.1 we discussed how outliers are generally considered to be more vulnerable to information disclosures than other data points. Statistical disclosure control methods combat vulnerabilities of outliers by rounding rare categories, truncating outlying values, or synthesising outliers (Reiter, 2003). In addition, some disclosure risk assessments specifically consider the disclosure risk of outliers.

Whether protecting outliers or assessing their disclosure risk, the first step is the identification of the outliers. This is often done on an ad hoc basis where values of variables that are infrequent enough to be considered outliers are identified. There are also procedural methods of identifying outliers, such as those implemented by algorithms for achieving k -anonymity (Definition 3.2).

Taub et al. (2018) calculate disclosure metrics for the entire set of training data and for

two subsets of outliers. The first subset, called *uniques*, consists of all training observations that have a unique combination of values for identifying variables. The second subset, called *special uniques*, consists of observations for which the combination of identifying variables is in the top 10% of special unique detection algorithm (SUDA) scores. SUDA is an algorithm that identifies observations that are unique due to an unusual combination of a small set of variables (Elliot et al., 2002). We highlight the fact that SUDA does not consider the order or the distance between values of a variable. Therefore, SUDA does not work well with non-categorical variables so alternative methods such as local outlier factor (LOF) should be used (Breunig et al., 2000).

6.4.1 Local outlier factor

An alternative method to find outliers, local outlier factor (LOF), is implemented by Ichim (2009). This method attempts to identify observations in low density regions that may be vulnerable to membership disclosure. LOF only requires that a measure of the distance between observations can be specified. Hence, LOF is more widely applicable than SUDA. In particular, LOF is applicable to datasets containing mixtures of data types.

LOF identifies outliers by comparing the local density (or reachability) of a point with the local density of neighbouring points (Breunig et al., 2000). Outlying points will have significantly lower local densities than their neighbours. The distance that encapsulates the neighbourhood of an observation can be defined in one of two manners.

Definition 6.1 (*k*-distance of O_i). For any positive integer k , the k -distance of an observation, O_i , is written as $d_k(O_i)$ and is defined as the distance $d(O_i, O_j)$ between O_i and an observation $O_j \in \mathcal{O}$ such that:

- i. for at least k objects $O'_j \in \mathcal{O} \setminus O_i$ it holds that $d(O_i, O'_j) \leq d(O_i, O_j)$, and
- ii. for at most $k - 1$ objects $O'_j \in \mathcal{O} \setminus O_i$ it holds that $d(O_i, O'_j) < d(O_i, O_j)$.

This definition can be problematic when there are k or more identical observations. This is because a neighbourhood can consist of solely identical observations which will be infinitely unreachable from other observations. In such cases, considering the neighbourhood size in terms of k -distinct distance is preferred. This definition is given in Breunig et al. (2000) and is as follows:

Definition 6.2 (*k*-distinct distance of O_i). For any positive integer k , the k -distinct distance of an observation O_i , written as $dd_k(O_i)$, and defined as the distance $d(O_i, O_j)$ between O_i and an observation $O_j \in \mathcal{O}$ such that:

- i. for at least k objects $O'_j \in \mathcal{O}_u \setminus O_i$ it holds that $d(O_i, O'_j) \leq d(O_i, O_j)$, and
- ii. for at most $k - 1$ objects $O'_j \in \mathcal{O}_u \setminus O_i$ it holds that $d(O_i, O'_j) < d(O_i, O_j)$,

where \mathcal{O}_u is the set of all unique observations

$$\mathcal{O}_u = \{\mathcal{O} : \forall i \neq j, O_i \neq O_j\}.$$

In short, $dd_k(O_i)$ is the distance between O_i and its k^{th} -nearest unique neighbour. For sets of data that contain no duplicates, k and k -distinct distance are equivalent. Subsequent definitions are written with the neighbourhood size defined by k -distance but are also valid for k -distinct distance.

For more information on the following definition, see (Breunig et al., 2000).

Definition 6.3 (Reachability distance of O_i w.r.t O_j). For any positive integer k , the reachability distance of an observation O_i with respect to an observation O_j is written as

$$rd_k(O_i, O_j) = \begin{cases} d_k(O_j), & \text{for } O_i \in N_k(O_j), \\ d(O_i, O_j), & \text{for } O_i \notin N_k(O_j), \end{cases}$$

where $N_k(O_i)$ is the set of k -nearest neighbours for the observation O_i .

Reachability distance ensures that all observations within a neighbourhood are treated as equidistant from the observation at the center. Note that it is whether O_i is in the neighbourhood of O_j that is relevant, rather than the reverse.

For more information on the following definition, see (Breunig et al., 2000).

Definition 6.4 (Local reachability density of O_i). The local reachability density of an observation O_i is defined as

$$\text{lrd}_k(O_i) = \frac{|N_k(O_i)|}{\sum_{O_j \in N_k(O_i)} rd_k(O_i, O_j)}. \quad (6.5)$$

Local reachability density is a measure of how outlying an observation is. Observations that are in the neighbourhoods of most of their neighbours will tend to have lrd_k close to one. Those that are not will have smaller values, with the value tending towards zero, as the distance from those neighbours increases.

Again, see Breunig et al. (2000) for more information on the following definition.

Definition 6.5 (Local outlier factor of O_i). The local outlier factor of an observation O_i is defined as

$$\text{lof}_k(O_i) = \frac{\sum_{O_j \in N_k(O_i)} \frac{\text{lrd}_k(O_j)}{\text{lrd}_k(O_i)}}{|N_k(O_i)|}. \quad (6.6)$$

LOF scales the local reachability densities given in Equation (6.5) with respect to each other. Inlying observations will have similar reachability densities to their neighbours. Therefore, they will have an LOF close to one. Outlying observations will have local

reachability densities that are less than neighbours. Consequently, their LOF will be greater than one, with greater differences in local reachability density corresponding to a larger LOF (Breunig et al., 2000).

LOF is sensitive to the choice of k . Choosing k that is too small can lead to problems where larger clusters are not differentiated from much smaller clusters. In contrast, choosing k too large can lead to a failure to identify outliers. Breunig et al. (2000) advises against choosing k values smaller than 10 since lof_k fluctuates wildly when k is small. They suggest calculating the lof_k over a range of k values and then, for each observation in a dataset, selecting the maximum lof_k score. They provide guidance on choosing upper and lower bounds. The lower bound of the range should be the minimum number of observations that could be considered a cluster, so that other nearby observations can be considered local outliers to that cluster rather than within. The upper bound should be the maximum number of nearby observations that can potentially be outliers to a cluster, rather than their own separate cluster. Ichim (2009) make a loose comparison between the k values of LOF and k -anonymity. In addition, they choose $k = 5$ which is smaller than the minimum value recommended previously. It is true that, in both contexts, k can be seen as some minimum number of observations that should be similar to an observation in order for it to be protected from disclosure. However, the requirement that at least k observations should be identical to an observation and the requirement that at least k observations should be in the neighbourhood of an observation are not comparable restrictions of privacy. Intuitively, k neighbours seems much less restrictive than k identical observations, so $k = 5$ feels like a low privacy criteria, but it is not clear how the size of k for lof_k affects the eventual classification of a point as an outlier.

Once lof_k has been calculated for each observation, a threshold to classify observations as outlier should be selected. Ichim (2009) makes three suggestions. The first is to classify observations within the top $\alpha\%$ of lof_k scores as outliers. However, the choice of α is not always obvious and will vary between datasets. A more data driven method is to plot lof_k in ascending order and select the elbow of the curve as the threshold to classify outliers. This procedure can be automated by fitting a structural change model (Zeileis et al., 2003) with a single break point to the sorted lof_k scores and using the break point as the threshold to classify outliers.

Specifically, the LOF score of each observation, \mathbf{x}_i , was calculated over a range of $k \in [10, 50]$ and choose the maximum lof_k score for each,

$$\text{lof}(\mathbf{x}_i) = \max_{k=10}^{50} \text{lof}_k(\mathbf{x}_i). \quad (6.7)$$

The outliers were classified according to a threshold which was found with a structural change model (Zeileis et al., 2002). LOF scores were sorted into ascending order and a

structural change model with a single break point was fit.

$$y_i \sim \mathcal{N}(\beta_j, \sigma^2) \quad \text{for} \quad y_n \geq y_{n-1} \geq \dots \geq 1,$$

where $(i = i_{j-1} + 1, \dots, i_j, j = 1, 2)$, j indicates the segment index, i_1 is the single break point, $i_0 = 0$ and $i_2 = n$.

The least squares estimate for the break point is the value i_1 that minimises

$$\hat{i}_1 = \min_{i_1} \left\{ \sum_{i=1}^{i_1} (y_i - \hat{\beta}_1)^2 + \sum_{i=i_1+1}^n (y_i - \hat{\beta}_2)^2 \right\},$$

where $\hat{\beta}_1 = \sum_{i=1}^{i_1} y_i / i_1$ and $\hat{\beta}_2 = \sum_{i=i_1+1}^n y_i / (n - i_1)$. Any observations with LOF greater than the break point is classified as an outlier

$$\mathcal{Y}_{\text{outlier}} = \left\{ i : y_i > \hat{i}_1 \right\}. \quad (6.8)$$

6.5 Summary of the literature review

We have reviewed an extensive and diverse range of utility assessments and disclosure risk assessments for synthetic data and other SDC methods. It is clear that there is a lack of consensus on how to assess utility and disclosure risk. The development of a broader consensus would aid in the comparison of synthetic data generation methods with other synthetic or non-synthetic privacy preserving techniques. This would then lead to more efficient development of new methods.

We have identified several obstacles that make assessment of synthetic data utility challenging. The first is that utility is quite difficult to describe. The qualities of a useful dataset will depend on the dataset itself and the datasets purpose.

This difficulty is further magnified by the privacy-utility trade-off that exists for synthetic data. If the perfect utility metric was to be defined it would not be enough to optimise that, because one must also be concerned about disclosure risk. It is not always clear, for some utility assessment methods, how good performance corresponds to the actual utility for a user of the data.

Unfortunately, these issues are not easy to solve. The best solutions that we have seen are those that implement a diverse suite of multiple evaluations (e.g. Beaulieu-Jones et al. (2019), Bowen and Snok (2020), Goncalves et al. (2020), Hu and Savitsky (2021), and Lin et al. (2019)). Rather than attempt to use the single ‘‘best’’ utility assessment, these papers accept that ‘‘best’’ does not exist, and instead consider a multiple utility assessments.

The obstacles to establishing a consensus for assessing the disclosure risk of synthetic datasets are very different. The solutions to these obstacles are also clearer. Completely

and incompletely synthesised data are not at risk from the same types of disclosure. While attribute disclosures can occur for either, identity disclosures can only occur for incompletely synthesised datasets. The solution to this problem seems to be clear. We should prefer to directly assess the disclosure scenario that is most concerning. Attribute disclosure represents the disclosure scenario that is generally most concerning and this can be assessed for both completely and non-completely synthesised data.

We identified two promising attribute disclosure risk assessments in the literature, the Bayesian framework of Hu et al. (2014) and Reiter et al. (2014) and the empirical attribute disclosure risk framework of Elliot (2014) and Taub et al. (2018). The Bayesian framework is appealing in theory but, in practice, it is difficult to implement, at least, outside of smaller examples and a few synthesis methods. It requires further development before it can be practically applied to a wide range of datasets and synthetic data generation methods. In contrast, the empirical framework establishes the foundations of an attribute disclosure assessment framework that can be applied to a diverse array of SDC datasets, but the intruder assumptions are fixed and do not reflect a motivated intruder (Definition 3.1).

Chapter 7

Methods for assessing synthetic data

In Chapters 5 and 6, we reviewed many of the commonly used methods for assessing the utility and disclosure risk of synthetic data. In this chapter, we establish our methodology for the assessment of synthetic data which we implement in Chapters 8 and 9. This includes methods that we have developed to address some of the issues that we encountered in our review of methods for the assessment of synthetic data.

This chapter is structured as follows. In Section 7.1, we focus on methods for assessing the utility of synthetic data. Section 7.2 concerns the disclosure risk assessment of synthetic data. Specifically, this chapter contains our novel framework for assessing disclosure risk of synthetic data.

7.1 Assessing the utility of synthetic data

In our review of utility assessment methods from the literature, we found that there is a lack of agreement within the statistics and ML communities for methods of utility assessment. Furthermore, we highlighted the difficulty of specifying assessments that encapsulate genuine utility and determined that there was no single “best” utility assessment. We found that the most effective examples of evaluating the utility of synthetic data were those that implemented several utility assessments (see, e.g., Beaulieu-Jones et al. (2019), Bowen and Snoke (2020), Goncalves et al. (2020), Hu and Savitsky (2021), and Lin et al. (2019)). As such, we mirror this approach for our utility assessments in Chapter 8 and Chapter 9. We consider a broad range of utility assessments that we found to be most effective in the literature. Specifically, we found that single number summaries can mask underlying problems with synthetic data (see Section 5.1), whereas, if many synthetic datasets are compared, the results of qualitative assessments can be difficult to interpret. We view the weaknesses of these single number summaries and qualitative assessments as complementary. Consequently, we incorporate both into our suite of utility assessments. We discuss some of those utility assessments in the following sections. Additional details of

the specifics of utility assessments for each example are described in the methods sections of each chapter (Section 8.2.3 and Section 9.2.3).

7.1.1 Plots

We use a variety of plots to assess our synthetic data, including univariate distribution plots, quantile plots and dimension-wise prediction (DWP) plots. For our DWP assessment, we use root mean squared error as the loss for continuous variables, Poisson log-likelihood for count variables, and area under the ROC curve for categorical variables. Furthermore, we follow the approach of Torfi and Fox (2020) and assess DWP with three prediction models: generalised linear regression, CART, and XGBoost.

7.1.2 Task performance

As discussed in our literature review in Section 5.6, evaluation of task performance is an invaluable method of assessment for synthetic datasets, especially those that have a specific intended use case. We generally associate the datasets that we synthesise in the later chapters of this thesis with prediction tasks. However, a prediction task would have considerable overlap with DWP. Therefore, we evaluate performance using statistical inference tasks. When carrying out our research, we were conscious of introducing potential researcher bias into our choices of inference models. In other words, if we chose the synthesis models and the inference task models, the data may perform better than if someone else chose the inference task models. Consequently, for each dataset, we evaluate performance on the inference task by replicating the statistical inference methods of others, such as Strack et al. (2014) and Vehtari et al. (2022).

Throughout our utility assessments we utilise the simple sample variance estimator (Equation (4.5a)) (G. Raab et al., 2016) to calculate variance estimates for synthetic data (see Section 4.2). This choice is appropriate as all synthetic data is completely synthetic and generated with Little’s approach (Section 4.1.2).

7.2 Assessing the disclosure risk of synthetic data

This section contains our methodology for evaluating disclosure risk. In Section 3.1, we concluded that, when considering the release of completely synthetic datasets containing sensitive, subject-level data, attribute disclosure is the most relevant type. Furthermore, we found that, based on current capabilities, the attribute disclosure methods described by Elliot (2014) and Taub et al. (2018) represent the best option for a standardised disclosure assessment framework. Our methodology for assessing the risk of disclosure builds on the

work of Elliot (2014) and Taub et al. (2018) and addresses several of the issues that we identified with their methods.

Now, we introduce our framework for assessing disclosure risk on synthetic data. We describe formal testing for whether attribute disclosure risk is acceptable and how information about outlying observations can be incorporated into an analysis of disclosure risk. To begin with, we outline four key elements of any effective framework for disclosure risk assessment.

1. The prior knowledge of the intruder,
2. the method(s) the intruder will use to predict sensitive information,
3. a metric to quantify the risk of disclosure, and
4. the criteria for the risk of disclosure to be deemed acceptable.

Most disclosure assessment methods in the literature establish several elements of this framework, however, a proper assessment of disclosure risk requires all four. Furthermore, the choices for these elements should reflect a reasonable disclosure scenario where, at the absolute minimum, a reasonable scenario is one that reflects a motivated intruder. As we discussed in our literature review in Sections 3.2.1 and 3.3.1, current and future intruder capabilities are unknown. As such, there is an argument to be made for choosing more conservative assumptions. Carrying out a comprehensive evaluation into the potential prior information and prediction methods for a motivated intruder will aid the design of an effective disclosure assessment.

We now discuss our choices for the four elements of the disclosure risk assessment that we implement in Chapters 8 and 9.

7.2.1 The prior knowledge of the intruder

In Chapters 8 and 9, we assume that the intruder is attempting to disclose multiple sensitive attribute values for the original data. For each subject in the original data, the intruder knows a set of quasi-identifying attributes which we will identify in each chapter. Furthermore, we assume that the intruder has full access to the synthetic dataset that is being assessed. We will generate multiple synthetic replications of each dataset; however, we assess the disclosure risk of each replication independently of all other replications.

All of these assumptions are quite standard (see, e.g., Choi et al. (2017), Elliot (2014), and Taub et al. (2018)) and reflect a reasonable motivated intruder. That said, our methods could be modified for a much stricter scenario. For example, we could modify by assuming that the intruder knows all attributes in the original data except for the sensitive attribute that they are predicting.

7.2.2 Intruder prediction method(s)

We emulate the scenario of a motivated intruder by training models to predict sensitive attributes conditional on the quasi-identifying variables. These models are trained on the synthetic data before predicting the sensitive attributes for each observation in the original data. In order to reflect uncertainty about the intruder’s choice of prediction method, we consider multiple prediction models:

- generalised linear regression,
- classification and regression trees (CART),
- random forest,
- XGBoost, and
- Modified empirical matching.

Additionally, we implement a modification of the empirical matching methods of Elliot (2014), which we introduce below. Recall that the empirical matching methods are the attribute prediction methods implied by the CAP and CAE methods of Section 6.3.2.

Modification to methods of Elliot (2014)

From Section 6.3.2, recall the issues we identified with the proposed solutions for undefined CAP and CAE scores that occur in the case of non-matches. Specifically, the proposals fail to utilise all of the information that is available to the intruder and, depending on the chosen solution, disclosure risk will be either overestimated or underestimated. To address these problems, we propose a modification to the prediction methods of Elliot (2014) and Taub et al. (2018) that better utilises the information that would be available to the intruder. We assume that, rather than giving up after a failed match, the intruder will continue to search for matches on subsets of the quasi-identifiers. Following the notation used in Section 6.3.2, our modification is defined as follows:

Definition 7.1 (Procedure for estimating undefined correct attribution probabilities). Let \mathbf{S} be the synthetic dataset consisting of n rows, where

$$\mathbf{s}_i^T = (\mathbf{k}_{si}^T, t_{si}),$$

is a row that contains the quasi-identifying variables \mathbf{k}_{si} and the target variable t_{si} . Similarly, let \mathbf{O} be the original dataset, consisting of m rows, where the row for the intruder’s j^{th} target is written as

$$\mathbf{o}_j^T = (\mathbf{k}_{oj}^T, t_{oj}).$$

Let $\text{CAP}_{\mathbf{k}}(\mathbf{o}_j, \mathbf{S})$ be the correct attribution probability given in Equation (6.2), where \mathbf{k} is the set of quasi-identifiers that the intruder will match on.

For \mathbf{o}_j where no match with \mathbf{S} exists we say that $\text{CAP}_{\mathbf{k}}(\mathbf{o}_j, \mathbf{S})$ is undefined.

We initialise the variable $k = 0$, which defines how many values in \mathbf{k} will be used to match and proceed as follows:

1. Identify all combinations of length $p - k$ vectors that are within \mathbf{k} . We denote the set of $\binom{p}{p-k}$ combinations as

$$\{\mathbf{k}_c \subset \mathbf{k} : |\mathbf{k}_c| = p - k\},$$

where $c = 1, \dots, \binom{p}{p-k}$.

2. For each combination \mathbf{k}_c , calculate $\text{CAP}_{\mathbf{k}_c}(\mathbf{o}_j, \mathbf{S})$.
3. If at least one $\text{CAP}_{\mathbf{k}_c}(\mathbf{o}_j, \mathbf{S})$ score is defined then we use the average as our estimate for $\text{CAP}(\mathbf{o}_j, \mathbf{S})$, ignoring undefined values,

$$\text{CAP}(\mathbf{o}_j, \mathbf{S}) = \frac{\sum_c \text{CAP}_{\mathbf{k}_c}(\mathbf{o}_j, \mathbf{S}) [\text{CAP}_{\mathbf{k}_c}(\mathbf{o}_j, \mathbf{S}) \in \mathbb{R}]}{\sum_c [\text{CAP}_{\mathbf{k}_c}(\mathbf{o}_j, \mathbf{S}) \in \mathbb{R}]}.$$

If not, increment k by one and return to step 1.

Note that, when $k = 0$, the steps define the original CAP score. When $k = p$, \mathbf{k}_c is empty and the modified CAP score is equivalent to the empirical probability of t_{si} ,

$$\text{CAP}(\mathbf{o}_j, \mathbf{S}) = \frac{1}{n} \sum_{i=1}^n [t_{si} = t_{oj}].$$

The modification to CAE is almost identical but with all instances of CAP replaced with CAE. That is, when the subset \mathbf{k}_c is empty,

$$\text{CAE}(\mathbf{o}_j, \mathbf{S}) = \left| t_{oj} - \frac{1}{n} \sum_{i=1}^n t_{si} \right|,$$

which is equivalent to the solution that Elliot (2014) proposed if CAE was undefined. The modification we have described is a more realistic scenario for how an attacker would proceed if they do not find an initial match. However, other problems with CAP and CAE still have not been addressed. Specifically, that the ordering of quasi-identifiers is ignored and all unique values are treated as distinct, so, for numerical or ordinal variables, it will be more appropriate to use other attribute prediction methods.

7.2.3 Quantifying the risk of disclosure

Our methodology for scoring predictions is different in Chapter 8 than it is in Chapter 9. For the dataset in Chapter 8, we score categorical variables using log-loss and numerical variables with mean squared error. We chose these loss functions to reflect our belief that the risk of disclosure decreases very quickly as an intruder’s prediction is further from the truth. We demonstrate this in the following example.

Example 7.1. *Suppose that we have a person with diabetes, who we denote P , and we have three intruders, who we denote I_1 , I_2 , and I_3 . Each intruder independently predicts whether P has diabetes, their predictions are as follows:*

1. I_1 believes that P has diabetes with probability $p_1 = 1$,
2. I_2 believes that P has diabetes with probability $p_2 = 0.75$, and
3. I_3 believes that P has diabetes with probability $p_3 = 0.5$.

If we rank these predictions in order of the severity of disclosure, it is p_1 then p_2 then p_3 . However, if we were to assign some numerical values to the severity, what would it look like? Consider the following two possibilities:

- (1) *The difference in severity from p_1 to p_2 and then p_2 to p_3 is about the same,*

$$p_1 > p_2 > p_3.$$

- (2) *The difference in severity from p_1 to p_2 is much larger than p_2 to p_3 ,*

$$p_1 \gg p_2 > p_3.$$

Possibility (1) tells us that the severity of a disclosure decreases roughly linearly as the prediction is further from the truth. On the other hand, possibility (2) tells us that the severity of a disclosure decreases rapidly as the prediction is further from the truth.

Since we believe that possibility (2) reflects the reality of most disclosure scenarios, we chose to penalise with log-loss and mean squared error. This is different from the methods of Elliot (2014), where he scores numerical predictions with mean absolute error and categorical predictions with the probability of the intruder predicting the true value. Under these choices, the loss increases linearly as the prediction moves further from the truth, so they reflect a belief in possibility (1). Neither approach is correct or incorrect, they simply quantify the disclosure risk differently.

That said, in Chapter 9, our scoring method for categorical variables changes, and we score with the probability of the intruder predicting the true value. This change is due to difficulties we encountered during the evaluation of the disclosure risk scores for both datasets, but more acutely for the dataset in Chapter 9. We explain the details of these problems in Section 9.2.4.

7.2.4 Evaluating acceptability of disclosure risk

In Chapters 8 and 9, we compare the risk of disclosure for several methods of synthetic data generation. We evaluate the risk of disclosure by comparing the synthetic datasets against each other and against disclosure risk baselines. We carry out these evaluations by fitting inference models to the disclosure risk scores (see Section 7.2.3). Specifically, we wish to answer these questions:

Questions 7.2.1.

- a. Which intruder attribute predictions methods result in the highest risk of disclosure?
- b. What is the relationship between method of synthetic data generation and risk of disclosure?
- c. How does the risk of disclosure for synthetic data compare with the test partition and k -anonymised baseline?

In this section, we describe the baselines and then introduce the inference modelling approach that we implemented.

Baselines

As in other examples of the literature (see, e.g, (Hu & Savitsky, 2021; Taub et al., 2018)), the following baselines represent acceptable levels of disclosure risk:

1. the test sample of data, and
2. the k -anonymous data (Definition 3.2), which is the training data with k -anonymisation conditioned on the quasi-identifiers.

The k -anonymised data (2) is a privacy baseline that we chose in order to reflect the masking methods described in Section 3.2.2. Recall that k -anonymised data is vulnerable to several types of disclosure attacks. Given these vulnerabilities, it does not reflect a realistic masking method by current standards. However, our intruders do not attempt any of the disclosure attacks that k -anonymised data is specifically vulnerable to so that will not impact our analysis.

The test data is independent and identically distributed to the training data. Therefore, if there was no risk of disclosure from releasing the test partition, it would be the best possible alternative to releasing the training data. Note that the risk of attribute disclosure will generally be higher for datasets that are more closely distributed to the training data (Palley & Simonoff, 1987). As such, the disclosure risk scores for the testing data are effectively as large as possible without any leakage of information from the training data.

Modelling risk of disclosure

In our disclosure risk evaluation, we consider the disclosure risk of multiple sensitive attributes (Section 7.2.1) and several intruder prediction methods (Section 7.2.2). Furthermore, we consider multiple synthetic data generation methods and multiple baselines. These choices were made to reflect uncertainty in the prior assumptions but they significantly complicate the risk evaluation.

In the literature review, we noted the difficulty of interpreting the results of attribute risk assessments for multiple attributes (Section 6.5). In our case, this difficulty is further complicated by our consideration of multiple sets of prior assumptions. With so many comparisons, it becomes infeasible to evaluate risk with t-tests. As such, we model the disclosure risk scores with regression models.

The regression framework is well suited to multiple comparisons. Model coefficients can describe the similarities and differences in between each set of prior assumptions. We consider the disclosure risk for multiple replications of the same synthesis method, and are able to incorporate this repeated measures design into our inference model. Furthermore, disclosure risks can be modelled for individual subjects, allowing for the possibility of including subject level details, such as whether an observation is an outlier.

During our disclosure assessments in Chapter 8 and Chapter 9 we encounter several challenges that require different modelling approaches to overcome. We describe these challenges and the specific details of the inference models that we implement in Section 8.2.3 and Section 9.2.4. After modelling our disclosure scores, we carry out model inference to answer the questions in posed Questions 7.2.1.

Membership disclosure

Earlier in this thesis we explained why we believe that, in the majority of cases, attribute disclosures are the most relevant disclosure type for synthetic data (Section 3.1). Consequently, our membership disclosure evaluations are far less rigorous. In the case that membership disclosures are the most relevant, the disclosure assessment that we discussed in Section 7.2 can be modified to handle membership disclosure. In fact, this would significantly simplify the process of fitting models for disclosure risk evaluation since we would no longer need to model the risks for several sensitive attributes.

The membership disclosure risk of each replication of synthetic data was evaluated following Choi et al. (2017), an approach we discussed in Section 6.2. The datasets in our examples contain a mixture of variable types, so we measure distance with Euclidean distance, we one-hot encode categorical variables, and standardise all variables by the mean and standard deviation of the training partition. We compute precision-recall curves at increasing distance thresholds and smooth the curves for each replication. Finally, we plot the smoothed precision-recall curves for each replication for comparison.

Chapter 8

Generating and assessing synthetic Pima data

In this chapter, we use the Pima Indians data as a simple example to demonstrate how synthetic data can be generated and evaluated. We use a variety of models to synthesise Pima data and subsequently evaluate the synthetic data following the methods described in Chapter 7.

Understanding some risk factors associated with diabetes is useful for determining the order in which we synthesise data. As such, we begin by briefly summarising the two types of diabetes and their risk factors, as well as the relevance of diabetes to the Pima dataset.

8.1 Diabetes and the Pima dataset

Diabetes is a serious and lifelong condition where a person's body is unable to regulate its blood sugar levels. There are two main types of diabetes, type 1 and type 2. Of the people in the UK that have diabetes, fewer than 10% have type 1, while 90% have type 2 (Diabetes UK, 2017b, 2019b). Several other types of diabetes also exist, most notably gestational diabetes, which affects 4-5% of pregnant women and develops during pregnancy (Diabetes UK, 2017c, 2017d; NHS, 2022).

The risk factors that are associated with developing diabetes differ for each type. The majority of people with type 1 diabetes are diagnosed as children. Additionally, there is a slightly increased risk of developing type 1 diabetes if another family member has it, (Diabetes UK, 2017a, 2017b).

According to Diabetes UK (2019b, 2023) and the NHS (2023), the factors that are associated with an increased risk of type 2 diabetes include, but are not limited to,

- being overweight or obese,
- being older,

- a family history of diabetes,
- high blood pressure,
- gestational diabetes during a pregnancy (NHS, [2022](#)),
- being of African Caribbean, Black African, South Asian or Chinese ethnicity, and
- having a history of high blood pressure (hypertension), heart attacks, strokes or severe mental illness.

According to Diabetes UK ([2017d](#)) and the NHS ([2022](#)), the factors that are associated with a greater risk of developing gestational diabetes are

- being overweight or obese,
- being older,
- having a family history of diabetes,
- gestational diabetes during prior pregnancies,
- previously giving birth to very large babies (over 4.5kg), and
- being of African Caribbean, Black African, South Asian or Middle Eastern ethnicity.

The Pima Indians Diabetes dataset is a dataset frequently used to demonstrate classification methods. The dataset contains personal information and measurements for women ($N = 768$) from the Pima Indian population near Phoenix, Arizona (Smith et al., [1988](#)). Participants were aged 21 years or older and did not have diabetes when they entered the study. The dataset has nine variables, they are:

1. age,
2. number of pregnancies,
3. plasma glucose concentration¹ (Eyth et al., [2023](#)),
4. diastolic blood pressure,
5. triceps skin fold thickness,
6. 2-hour serum insulin,
7. body mass index (BMI),

¹Subjects with plasma glucose concentration ≥ 200 mg/dl were specifically excluded from the original data. This is because the authors were interested in predicting whether the subjects would develop diabetes in the future and that is the threshold to be classified as diabetic (CDC, [2019](#)).

8. diabetes pedigree function (DPF)², and
9. whether a subject developed diabetes within the next five years.

8.2 Methods

Our data pre-processing steps mirror those of Ripley (1996, p. 14). We remove the insulin variable (48% missing) and any incomplete cases. Then, we divide the data into training ($n = 200$) and test ($n = 332$) splits³. By removing observations that contain missing values instead of imputing them, the variance and the covariance of variables in the data are biased to be smaller than the true values. Unless the values are missing completely at random, the means will also be biased. To correctly account for missingness, data should be multiply imputed. This can be incorporated into synthesis by following the two-step procedure that we described in Section 4.2.3. However, in the interest of keeping the example simple and easy to follow, we forgo imputation during synthetic data generation.

8.2.1 Synthesising Pima data

Order of synthesis

We synthesise Pima data using sequential synthesis. As such, we must determine the order in which we will synthesise the data (see Section 4.4.1). For the Pima dataset, there are well-established associations between some of the variables, so we opt to select a synthesis order that is logically consistent with those relationships. Notice that the seven variables we focus on can be sorted into two groupings. The first grouping is for a variable that describes a medical test result or risk factor for diabetes. That is, the first grouping consists of variables (1) through (8). The second grouping is the outcome variable, that is, variable (9) in the above list. Since the outcome variable is conditional on all other variables, the logical choice for the order of synthesis is to synthesise the variables in the first grouping and then synthesise the variable in the second grouping. As for the order of synthesis within the first grouping, we make several assumptions about the dependence between variables. These assumptions are informed by risk factors and pathophysiology of diabetes, see Figure 8.2.

First, we assume that age and DPF are independent of other variables. As such, we synthesise these variables first. Then, we synthesise number of pregnancies, which we assume depends on age. Next, notice that skin thickness and BMI are measures for whether a person is overweight. In fact, the prevalence of this increases with age and

²DBF is a metric developed by Smith et al. (1988). It uses a person's family history of diabetes to estimate the genetic influence on their chances of developing diabetes.

³These splits can be found in the R package, MASS (Venables & Ripley, 2002)

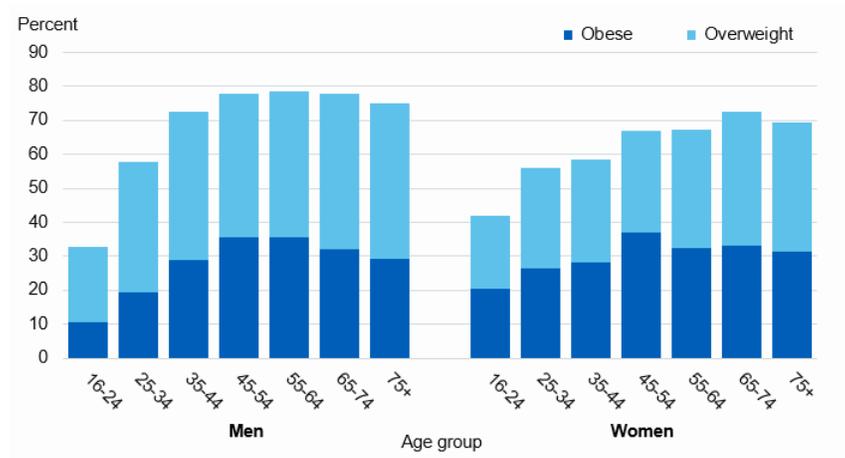


Figure 8.1: Prevalence of obesity in the UK for different age groups and sexes (NHS Digital, 2019).

in postpartum women (see, Figure 8.1 and Hollis et al. (2017), respectively). Therefore, we synthesise these variables next. Then, we assume that plasma glucose concentration is conditional on all risk factors of diabetes. To summarise, we synthesise the variables in the following order: age, DPF, number of pregnancies, skin thickness, BMI, plasma glucose concentration, and the outcome variable, whether a subject developed diabetes within five years.

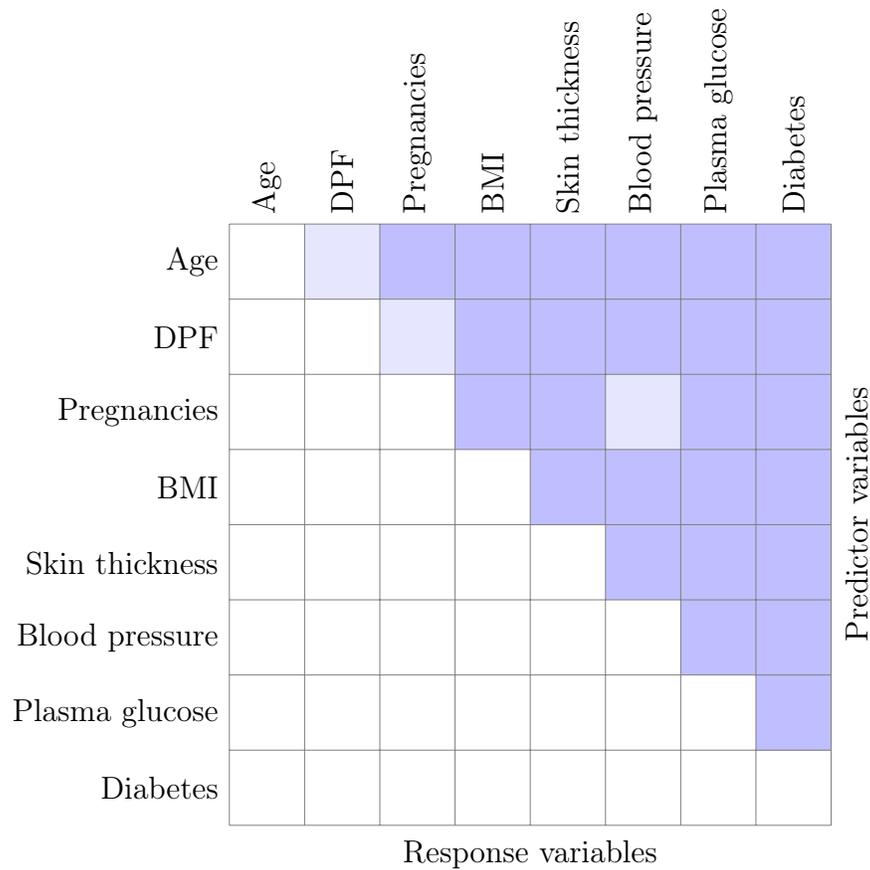


Figure 8.2: The variables of the Pima Indians dataset, arranged from left to right in order of synthesis. Shaded squares indicate predictor variables for each response, and darker squares indicate conditional relationships that were assumed when deciding the order of synthesis.

Synthesis models

We use Little’s framework (defined in Section 4.1.2) to generate 100 replications of completely synthetic Pima data for eight variations of synthesis models. Two regression variations (Model 8.1), three classification and regression trees (CART) variations (Model 8.2), and three random forest variations (Model 8.3). For all variations, we will keep the order of synthesis constant, see Section 8.2.1. With the exception of age, we generate all variables using the models we list below. As the first variable to be synthesised, age is generated by sampling with replacement.

For the generalised linear models, we will synthesise data from both penalised and non-penalised regression models.

Model 8.1 (Regression synthesis model). For the first variation, regression models for each variable are fit with no penalty. Whereas, for the second variation, the regression models for each variable are fit with an elastic net penalty, where $\alpha = 0.5$ is the mixing parameter (see Equation 2.17) and the regularisation penalty is fixed to $\lambda = 10^{-5}$. These

hyperparameters were chosen to enforce a small degree of shrinkage and variable selection.

We will model the diabetes variable with a Bernoulli distribution, as in Equation (2.7). Then, we will model the eight numeric variables as normally distributed, see Equation (2.2). However, during the exploratory analysis of the data we noted that some numeric variables are right skewed. We will address this by transforming those variables pre-synthesis, and then reversing those transformations post-synthesis. The variables age, BMI, number of pregnancies, 2-hour serum insulin and number of pregnancies are log transformed. As the number of pregnancies variable contains zeroes, we will increment by one before the log transformation.

Now, for the CART synthesis models, we will explore three approaches of smoothing the numeric variables with cubic splines.

Model 8.2 (CART synthesis model). For each numeric variable, data is generated according to one of the following strategies, no smoothing, smoothing the leaves and smoothing the data (see Equation (4.7) and Equation (4.8)). For all variables, the minimum node size is set to 5.

For the random forest synthesis models, we will also explore three approaches of smoothing the numeric variables with cubic splines.

Model 8.3 (Random forest synthesis model). For each numeric variable, data is generated according to one of the following strategies, no smoothing, smoothing the leaves and smoothing the data (see Equation (4.7) and Equation (4.8)). Minimum node size, the number of variables considered at each split and the number of trees were set to the default values for the randomForest package (see Section 2.2.1).

Ideally we would choose the hyperparameter values for each regression, CART and random forest model through hyperparameter tuning. If the hyperparameters for the models of individual variables are optimised, the models will fit the data better and this should improve the quality of the synthetic data. However, hyper-parameter tuning would increase the computational complexity of the synthesis models and so we leave this for future work.

8.2.2 Assessing the utility of synthetic Pima data

Recall the methods for assessing the privacy and utility of synthesised data from Chapter 7. In this section, we describe our method of implementation for the Pima dataset.

Baseline datasets

Throughout this chapter, we will compare the utility and privacy of the synthesised datasets with each other, but also against three baseline datasets.

1. The training sample of Pima data,
2. the test sample of Pima data, and
3. the 5-anonymous data, which is the training data with k -anonymisation ($k = 5$), where the quasi-identifiers are age and number of pregnancies.

The training data (1) is a “high” utility baseline, so we will compare the synthetic data against that for utility assessments. The k -anonymised data (3) is a “high” privacy baseline. Therefore, we are interested in whether the synthetic datasets have similar or higher utility while maintaining a similar or lower risk of disclosure. The test partition (2) is both a “high” utility and “high” privacy baseline. The test data is independent and identically distributed to the training. Therefore, if there was no risk of disclosure from releasing the test partition, it would be the best possible alternative to releasing the training data. Any differences in the utility of the training and testing partitions can be attributed to sampling variance. The disclosure risk scores for the testing data are effectively as large as is possible, without any leakage of information from the training data.

Dimension-wise prediction (DWP)

We will use dimension-wise prediction (Definition 5.2) to check conditional relationships in the synthetic datasets. All 8 variables will be predicted with each of the following models.

Model 8.4 (Dimension-wise prediction).

- a. Penalised generalised linear regression with an elastic net penalty,
- b. random forest, and
- c. the gradient boosting algorithm XGBoost (Chen & Guestrin, 2016).

Predictions for the diabetes variable will be scored using area under the receiver operating characteristic (AUROC), while the numeric variables will be scored using root mean squared error (RMSE).

Some hyperparameter tuning is carried out using 10-fold stratified cross validation, however, to keep the run time reasonable we choose fixed values for other parameters. For penalised regression, the weight penalty λ is optimised and the elastic net penalty α is set at 0.5 to allow for both shrinkage and variable selection. The number of trees in the random forest is set to 128, which is reasonable for a dataset that contains 200 observations. More trees may have improved predictions, however, the ensemble models are the bulk of the total runtime for dimension-wise prediction, and using fewer trees shortens the runtime of the random forests. For the remaining random forest hyperparameters, we use the

default values that were described in Section 2.2.1. For XGBoost, the depth of each tree is optimised with a grid search and models are trained with early stopping. A maximum of 128 rounds of boosting are carried out, but the algorithm is terminated if there is no improvement in validation loss for 15 consecutive boosting rounds. Default values were used for all other XGBoost hyperparameters (XGBoost developers, 2023).

Propensity scores

We will also implement the propensity scores assessment given in Section 5.2. We explore two discriminator models, a logistic regression that includes first order terms, and CART.

Initially, the complexity parameter for the CART discriminator is chosen to be the average of the best values (according to 10 fold CV) for all synthetic datasets. However, in contrast to Bowen and Snoke (2020), we note that there are large differences in the best complexity parameters across the datasets. Furthermore, for some datasets, the average of the best complexity parameters leads to trees that have a single node, which is equivalent to a random Bernoulli draw. As we noted in Section 8.2.2, a Bernoulli model will optimise pMSE, regardless of the data. Consequently, we instead fix the complexity parameter of the CART discriminator individually for each synthesis model. We choose this value to be the average of the best values (according to 10 fold CV) for all 100 replications of that synthesis model. In addition, we find the minimum value for the discriminator to have at least one split for all replications of a synthesis model. Then, we ensure that the complexity parameter for each model is greater than that minimum value.

For each synthetic dataset, we use pMSE ratio to assess predictions from each discriminator, see Equation 5.5. Notice that we calculate the null distribution of the CART discriminator (Algorithm 1) individually for each synthesis model so that the complexity parameters match. This ensures that pMSE values are comparable.

Task performance: inference

We will be assessing the utility of the synthetic Pima data in a teaching scenario, by replicating an inference that was previously carried out on the dataset by Vehtari et al. (2022). Our goal is to compare the conclusions of a person that has access to the synthetic data, with those of a person that has access to the original data.

Vehtari et al. (2022) do not partition the data or remove the serum insulin variable, and they remove all observations with missing data. However, we prefer to use the same training, test and synthetic data for all utility and privacy assessments. So we will carry out the real data inferences on the training partition (see Section 8.2). As discussed in Sections 4.2.3 and 8.2, the removal of observations will bias the data. The missing observations were removed before the synthesis models were fit, therefore both the training and synthetic datasets should be similarly biased. Since we are interested in whether inferences

align between the synthetic and training data, the biased data is not as problematic as it would be if the main goal was a valid inference of the original Pima data. Although, it is still possible that the bias could affect the results for the comparison of inferences. For example, the conditional distributions of the biased training data may be simpler than the true conditional distributions, which might benefit a poor synthesis model that would have struggled to accurately model the true distribution.

We fit inference models to each dataset, with diabetes within the next five years modelled as a Bernoulli distributed variable

$$y_i \sim \text{Bernoulli}(p_i).$$

Note that we scale and centre all covariates prior to fitting.

The first inference model that we fit is a standard logistic regression with weakly informative priors.

Model 8.5 (Pima inference, normal prior).

$$p_i = (1 + \exp(-\alpha - \mathbf{x}_i^T \boldsymbol{\beta}))^{-1},$$

$$\alpha \sim \mathcal{N}(0, 2.5) \quad \text{and} \quad \boldsymbol{\beta} \stackrel{\text{iid}}{\sim} \mathcal{N}(0, 2.5),$$

where \mathbf{x}_i^T contains all covariates in the Pima data.

We then fit a more sparse model, a logistic regression with a regularised horseshoe prior, see Equation (2.10).

Model 8.6 (Pima inference, horseshoe prior).

$$p_i = (1 + \exp(-\alpha - \mathbf{x}_i^T \boldsymbol{\beta}))^{-1},$$

$$\alpha \sim \mathcal{N}(0, 2.5), \quad \beta_j | \lambda_j, \tau, c \sim \mathcal{N}(0, \tau^2 \bar{\lambda}_j^2), \quad \lambda_j \sim \text{C}^+(0, 1),$$

$$\tau | \sigma \sim \text{C}^+(0, \tau_0^2), \quad \tau_0 = \frac{2}{7} \frac{\sigma}{\sqrt{n}}, \quad c^2 \sim \text{Inv-Gamma}(2, 12.5),$$

where \mathbf{x}_i^T contains all covariates in the Pima data and σ^2 is the estimated psuedo-variance for the Bernoulli distribution

$$\sigma^2 = \frac{1}{\bar{y}(1 - \bar{y})}.$$

Finally we fit a generalised additive mixed model (GAMM), which was the best fitting model in the original inference (Vehtari et al., 2022).

Model 8.7 (Pima inference, generalised additive mixed model).

$$p_i = \left(1 + \exp\left(-\sum_{j=1}^4 f_j(x_{ij})\right) \right)^{-1},$$

where $\mathbf{x}_i^T = (x_{i1}, x_{i2}, x_{i3}, x_{i4})$ are the covariates, glucose, age, BMI and DPF, and f_j is the low rank thin-plate regression spline basis for the j^{th} covariate.

This model is reparameterised to the mixed model representation (see Section 2.1.4),

$$g(\mu_i) = \tilde{\mathbf{x}}_i^T \boldsymbol{\beta} + \mathbf{Z}_i \mathbf{b},$$

where g is the logit function.

Weakly informative priors are selected for the coefficients of the fixed effects

$$\boldsymbol{\beta} \stackrel{\text{iid}}{\sim} \mathcal{N}(0, 2.5),$$

the coefficients of the random effects

$$\mathbf{b} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}/\lambda),$$

and the smoothing penalty

$$\lambda \sim \text{Exponential}(1).$$

We evaluate the fit of each model using expected log predictive density (ELPD), see Equation (2.9). Then, we check which model has the “best” fit for each dataset, where we have two definitions for “best”. Under our first definition, we consider the model with the highest ELPD for a replication to be the best. However, in practice, if multiple models have virtually indistinguishable fits then we would not rule out the marginally worse model. In fact, if the difference is not significant, we may prefer a more sparse model that is slightly worse fitting. Consequently, our second definition of “best” is that all models with an ELPD within 1 standard deviation of the model with the highest ELPD are the best for that replication. In the original analysis by Vehtari et al. (2022), the ELPD of the GAMM (Model 8.7) was significantly larger than either of the generalised linear models (Model 8.5 and Model 8.6). That is to say, the GAMM was the best fitting model under both definitions.

We will also compare the 90% credible intervals interval overlaps of the synthetic and training datasets, see Equation (5.7). For each replication of data and coefficient, we will average over all coefficients to obtain an interval overlap per replication. However, by averaging over all coefficients, we lose information about the percent overlaps of individual variables. Therefore, we will also average the overlap for a coefficient over all replications. Note, it is only possible to compare interval overlaps for the coefficients of the generalised linear models, Model 8.5 and Model 8.6, as the smooths of Model 8.7 will change across datasets.

As mentioned previously, we are interested in whether the conclusions of the synthetic and original data align. Therefore, we also check whether the hypothesis test results of these 90% credible intervals align with each other. For each coefficient, we count the number of synthetic data replications that the hypothesis test result matches the result

for the original data. Let us introduce some formal notation to clarify this.

Consider a posterior sample $\tilde{\beta}$, and let $\tilde{\beta}_{[p]}$ denote the p^{th} percentile of that posterior sample. Then the 90% credible interval of $\tilde{\beta}$ is

$$\left(\tilde{\beta}_{[5]}, \tilde{\beta}_{[95]} \right).$$

For our analysis, we use a two-sided hypothesis test. As such, there are three possible outcomes for the hypothesis test result. Without loss of generality, the null hypothesis of our hypothesis test is that $\beta = 0$, and the three possible outcomes are

$$H(\tilde{\beta}) = \begin{cases} H_{A+}, & \text{if } \tilde{\beta}_{[5]} > 0, \\ H_{A-}, & \text{if } \tilde{\beta}_{[95]} < 0, \\ H_0, & \text{otherwise.} \end{cases}$$

For us to consider the hypothesis of a synthetic data to match the original data, the result must match. Lets clarify this with an example of a non-match. Suppose the original data hypothesis $H(\tilde{\beta}_o)$ and the synthetic data hypothesis $H(\tilde{\beta}_s)$ are equal to H_{A+} and H_{A-} , respectively. Despite both results indicating that the respective parameter is significant, the hypotheses of the original and synthetic data *do not* match.

8.2.3 Assessing disclosures risks of synthesised Pima datasets

In this section, we assess the risk of attribute disclosure by a motivated intruder (see Definition 3.1). The potential quasi-identifiers for the Pima dataset are age and the number of pregnancies. We assume that the intruder knows these quasi-identifying variables for all observations in the training data, that the intruder has access to a synthetic dataset, and that they will attempt to predict the values of the attributes BMI, blood pressure, plasma glucose concentration and diabetes for observations in the training partition. As we are unsure which prediction methods the intruder will use, or what the most effective methods are, we compare the 5 methods discussed in Section 7.2.2.

Model 8.8 (Attribute prediction). Attributes will be predicted with the following prediction models

- a. CART,
- b. random forest,
- c. generalised linear regression,
- d. empirical matching (see Section 7.2.2), and

e. XGBoost (Chen & Guestrin, 2016).

So, our intruder will be training their chosen prediction method on the synthetic dataset, before predicting the attribute values for each observation in the real data. We introduce some notation for this scenario. We denote the intruder's target observation i , intruder's prediction method j , the attribute they will predict t , and the r^{th} replication from the s^{th} synthesis model to be rs . Also, let

$$f: \mathbf{k} \rightarrow t,$$

be a prediction method, which takes the two quasi-identifying variables \mathbf{k} as input and predicts an attribute.

So, the intruder will proceed as follows. First, they will use the synthetic data, rs , to fit their prediction method, which we denote f_{jrst} . Then, they will use their fitted prediction method to predict values for observations in the training partition,

$$f_{jrst}(\mathbf{x}_{i,\mathbf{k}}).$$

We use a prediction scoring function to evaluate the risk of disclosure for each prediction. However, our target attributes are not all the same data type. Consequently, we must use different scoring functions for different data type.

$$y_{ijrst} = l_t(f_{jrst}(\mathbf{x}_{i,\mathbf{k}}), x_{i,t}), \quad (8.1)$$

where l_t is the attribute disclosure scoring method for the t^{th} attribute. We want to choose scoring functions that penalise predictions more severely as they are further from the truth. Consequently, l_t is the logistic likelihood for categorical variables, and root mean squared error for numeric variables.

We have now described the process that we use to model a scenario where a motivated intruder utilises a synthetic dataset and some background knowledge to disclose sensitive attribute information. Before we describe the methods for analysing the results of this scenario, let us consider the questions that we would like to answer.

We had several questions of interest when evaluating the attribute disclosure scores.

Questions 8.2.1.

- a. Which attribute predictions methods result in the highest risk of disclosure?
- b. What is the relationship between method of synthetic data generation and risk of disclosure?
- c. Does smoothing or regularisation during data synthesis affect the risk of disclosure?

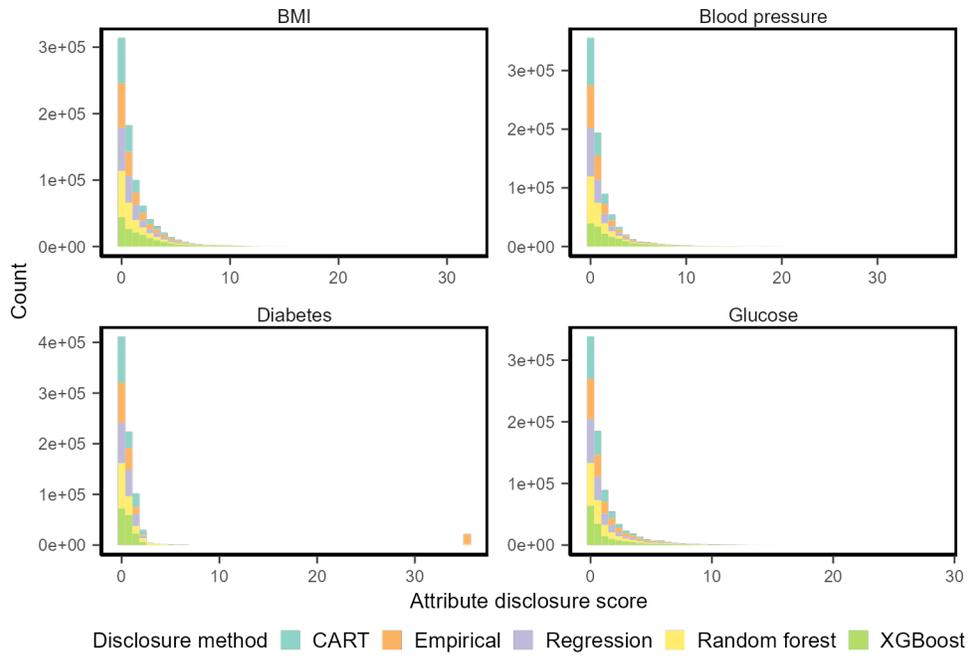


Figure 8.3: Histogram of attribute disclosure scores.

- d. How does the risk of disclosure for synthetic data compare with the test partition and k -anonymised baseline?
- e. Is the risk of disclosure higher for outliers?

We can answer all of these questions by fitting an inference model to the attribute disclosure data, see Equation 8.1. Let us carry out an exploratory analysis of the attribute disclosure data, to determine if that is feasible.

The first thing to note is that there are 200 subjects in the training data, 5 attribute prediction methods, 4 target variables, and 100 replications of 8 synthesis models. Consequently, the dataset is quite large (3,200,000) and contains 16,000 repeated measures per subject. If we were to fit a simple model, 3,200,000 observations would not be a problem. However, at minimum, we need a hierarchical model to handle the repeated measures. Furthermore, we are about to discuss some other challenging aspects of the data distribution that we must also model. In short, a simple inference model is not going to work.

Consider the histogram of attribute disclosure data (Figure 8.3). The distribution is heavily right skewed, so we require an inference model that can handle heavily skewed data, such as log-normal or gamma. Alternatively, we can consider modelling the data as normally distributed, if we can find a transformation to address the skew. However, the data also contains a large number of zeros, see Table 8.1. This complicates things, as neither log transformations, log-normal models or gamma models can handle zeros.

Another issue with the attribute disclosure data (Figure 8.3) is that the distribution

Disclosure method	Percentage zeros	Target variable	Percentage zeros
CART	0.056%	Blood pressure	0.505%
Empirical	1.195%	BMI	0.293%
Regression	0.000%	Diabetes	0.000%
Random forest	0.000%	Glucose	0.202%
XGBoost	0.000%		

Table 8.1: The percentage of subsets of attribute disclosure scores y_{ijrst} equal to zero.

of y_{ijrst} is multi-modal. While the numeric variables BMI, blood pressure and glucose all display similar peaks at zero and long tails, the distribution of the categorical variable diabetes has a much shorter tail. It will be challenging to find an inference model that can fit both the numeric and categorical scores. The simplest solution would be to model them separately. Although, the effects of the covariates for the two models would then be completely independent of each other.

We could also consider a *distributional model* which is a model that allows for specification of predictor terms for all parameters, rather than just the location or shape parameter (Bürkner, 2017). This relaxes the regression assumption that errors are independently and identically distributed. Instead, we assume that the errors are conditionally independent, and depend on some covariates. With a distributional model, the effects of the covariates on disclosure risk are shared for all attributes.

The final issue to note from our exploratory analysis is a worrying spike of extremely high values around 35. Further investigation reveals that these high values occur for every observation that the prediction model is wrongly certain about a prediction. More formally, as the predicted probability of the true value tends towards 0, the log-loss function converges to infinity. It will be very difficult to find a model that fits the extreme values without modelling them as separate processes. However, we believe that a large enough sample of Pima data would contain a value that contradicts any certain prediction. In other words, the values of exactly zero or 35 are the result of a lack of precision rather than a different process. Due to the difficulties that we would face in modelling these extreme values, we need to average over some dimensions of the data.

Let's consider two options for averaging data. Averaging over the subjects in the training data and averaging over each replication. By averaging over the subjects

$$y_{jrst} = \frac{1}{n_i} \sum_{i=1}^{n_i} y_{ijrst}, \quad (8.2)$$

each y_{jrst} is now the average attribute disclosure score for all subjects in the training partition. This reduces the size of the data to 16,000 observations. We can see that averaging over all of the subjects has removed the extreme values at both 0 and 35 (Figure 8.4).

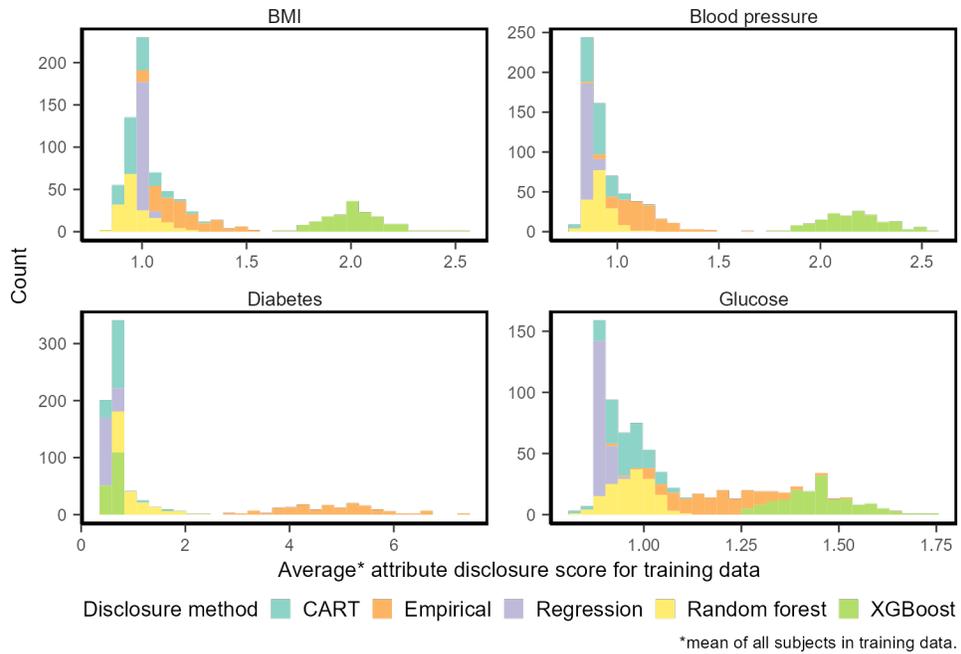


Figure 8.4: Histograms of y_{jrst} for each combination of target variable and disclosure method, see Equation 8.2.

However, we are no longer able to model the effect of the subject on disclosure risk. This does simplify the modelling, but we would not be able to explore whether outliers were at greater risk of disclosure (see Question 8.2.1e).

Instead, if we were to average over the replications

$$y_{ijst} = \frac{1}{100} \sum_{r=1}^{100} y_{ijrst}, \quad (8.3)$$

each y_{ijst} is now the average disclosure risk for the i^{th} subject, we would reduce the size of the data to 32,000 observations and has removed the extreme values at both 0 and 35 (Figure 8.5). However, we are still able to model the effect of subject on disclosure risk, which was not possible with Equation (8.2). Note, by averaging over the replications, we underestimate the variance in disclosure risk for a single replication of synthetic data.

We will attempt to model the attribute disclosure scores (Equation (8.2) and Equation (8.3)) by fitting increasingly complex regression models. Our goal is to find the simplest regression model that explains the attribute disclosure and use that model to answer the questions posed in Questions 8.2.1. However, let us first begin where almost all statistical inference starts, a linear regression model. More specifically, we will fit the implementation that allows us to specify priors on R^2 , see Model 2.1 for more details.

Recall, linear regression assumes that the errors are independent and identically distributed. Given the multimodalities and repeated measures that we discussed in the ex-

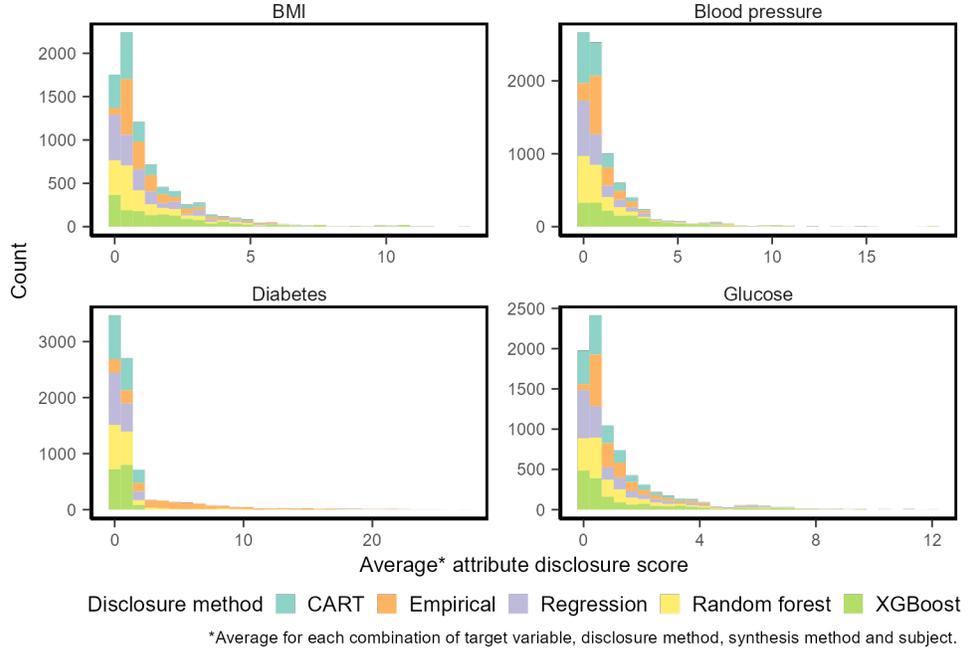


Figure 8.5: Distribution of y_{ijst} for each target variable and disclosure method, see Equation 8.3.

ploratory analysis, this is not the case for our data. Consequently, we will further simplify the attribute disclosure scores

$$y_{jst} = \frac{1}{n_r n_i} \sum_{r=1}^{100} \sum_{i=1}^{n_i} y_{ijrst}. \quad (8.4)$$

Note the distribution of y_{jst} , see Figure 8.6. Due to differences in type and scoring function for our attributes (Equation 8.1), this data is not identically distributed. We address this in the models defined in this chapter, see Model 8.9 through Model 8.16.

It is difficult to draw any conclusions about the distribution of y_{jst} , due to the small number of observations but there is some evidence of a right skew. Consequently, we will log transform the scores before fitting the linear regression model.

Model 8.9 (Full linear model $\log y_{jst}$).

$$\log y_{jst} | R^2 \sim \mathcal{N}(\alpha + \beta_{1,j} + \beta_{2,s} + \beta_{3,t} + \beta_{4,(j,t)}, \sigma_\epsilon^2),$$

with a weakly informative prior

$$R^2 \sim \text{Beta}\left(\frac{1}{4}, \frac{1}{4}\right).$$

From our exploratory analysis, we found the synthesis model to be the weakest predictor. Therefore, we will also consider a further simplification of removing the synthesis

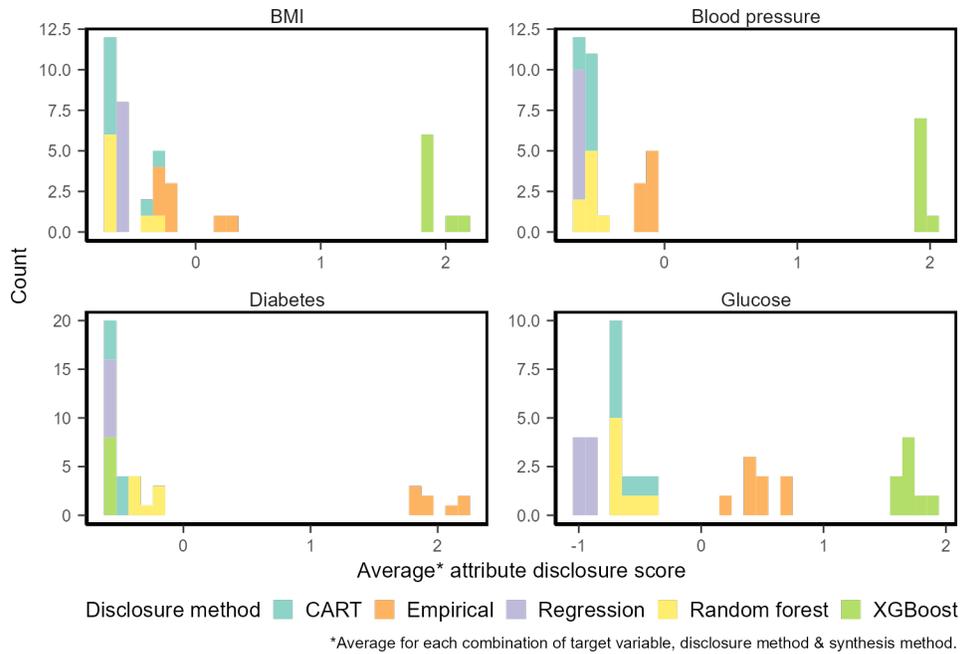


Figure 8.6: Histogram of simplified attribute disclosure scores, see Equation 8.4.

model coefficient⁴.

Model 8.10 (Simplified linear model $\log y_{jst}$).

$$\log y_{jst} \sim \mathcal{N}(\alpha + \beta_{1,j} + \beta_{3,t} + \beta_{4,(j,t)}, \sigma_\epsilon^2),$$

with a weakly informative prior

$$R^2 \sim \text{Beta}\left(\frac{1}{4}, \frac{1}{4}\right).$$

We run each of the linear regression models for 6 chains. Each chain has 5000 warm-up iterations followed by 5000 iterations of sampling from the posterior.

Next, we consider fitting distributional models to the data. As we discussed in our exploratory analysis, we prefer distributional models over separate models for each attribute. As there is no hierarchical component to these distributional models, they are not appropriate for the subject level data. Consequently, we will fit these models to y_{jrst} (Equation 8.2). The distribution of y_{jrst} contains modalities for target variable and disclosure method that have different variances (Figure 8.4). Furthermore, these distributions are right-skewed, with the degree of skewedness varying between modalities. Given the skewed data, we will model log-transformed y_{jrst} as normally distributed.

⁴We also considered other simplifications, including the removal of the interaction, however this was excluded for brevity.

Model 8.11 (Gaussian distributional).

$$\begin{aligned}\log y_{jrst} | \alpha, \boldsymbol{\beta}, \boldsymbol{\gamma}, \pi &\sim \mathcal{N}(\mu_{jst}, \sigma_{jst}^2), \\ \mu_{jst} &= \alpha + \beta_{1,j} + \beta_{2,s} + \beta_{3,t} + \beta_{4,(j,t)}, \\ \log \sigma_{jst} &= \pi + \gamma_{1,j} + \gamma_{2,s} + \gamma_{3,t} + \gamma_{4,(j,t)}.\end{aligned}$$

Additionally, we also consider modelling y_{jrst} with a log-normal and Gamma distributions.

Model 8.12 (log-normal distributional).

$$\begin{aligned}y_{jrst} | \alpha, \boldsymbol{\beta}, \boldsymbol{\gamma}, \pi &\sim \text{Lognormal}(\mu_{jst}, \sigma_{jst}^2), \\ \mu_{jst} &= \alpha + \beta_{1,j} + \beta_{2,s} + \beta_{3,t} + \beta_{4,(j,t)}, \\ \log \sigma_{jst} &= \pi + \gamma_{1,j} + \gamma_{2,s} + \gamma_{3,t} + \gamma_{4,(j,t)}.\end{aligned}$$

Model 8.13 (Gamma distributional).

$$\begin{aligned}y_{jrst} | \alpha, \boldsymbol{\beta}, \boldsymbol{\gamma}, \pi &\sim \text{Gamma}(k_{jst}, \theta_{jst}), \\ \log k_{jst} &= \alpha + \beta_{1,j} + \beta_{2,s} + \beta_{3,t} + \beta_{4,(j,t)}, \\ \log \theta_{jst} &= \pi + \gamma_{1,j} + \gamma_{2,s} + \gamma_{3,t} + \gamma_{4,(j,t)}.\end{aligned}$$

Exploratory analysis and initial model fitting show that the attribute scores are relatively small and variance between the models is not too extreme. As such, we will use the same weakly informative priors for all three models (Model 8.11, Model 8.12 and, Model 8.13).

$$\begin{aligned}\alpha &\sim \mathcal{N}(0, 2.5^2), \quad \boldsymbol{\beta} \stackrel{\text{iid}}{\sim} \mathcal{N}(0, 2.5^2), \\ \pi &\sim \text{student-t}_3(0, 2.5) \quad \text{and} \quad \boldsymbol{\gamma} \stackrel{\text{iid}}{\sim} \mathcal{N}(0, 1^2)\end{aligned}$$

Prior predictive draws from the models contained many samples that were orders of magnitude smaller and larger than the data. This indicates that the priors are not strongly informative and that, if necessary, we can specify more informative priors.

During early experimentation with distributional models, we find that it takes more than 24 hours to run Model 8.11 for 2000 iterations. Given that Models 8.12 and 8.13 will be slower to train, and estimation of ELPD can also require refitting the models multiple times, it is unrealistic to train all models on the entire dataset within a reasonable timeframe. As such, we save time by randomly sampling 20 of the 100 replications and only training the models on those. In our experimentation, we found that reducing the training data to 20 replications had a negligible effect on the coefficients of Model 8.11.

We run 4 chains with 1000 warm-up iterations and 1000 sampling iterations per model and find this to be sufficient for convergence.

With such flexible models, there is a danger of over-fitting. So we will consider reducing model complexity by removing parameters and enforcing regularising priors (Model 8.14d). Recall that, exploratory analysis showed that synthesis method was the weakest predictor. Therefore, we will introduce simplifications that remove the synthesis method for either the location or scale parameter, see Models 8.14a and 8.14b. Furthermore, recall that the distributions of BMI, blood pressure and glucose were similar, as were the distributions for variants of the same synthesis model (Model 8.1, Model 8.2 and, Model 8.3). So, we also consider simplifying the model by changing σ_{jst} to depend only on the synthesis model type, and whether an attribute is numeric or categorical (Model 8.14c).

Model 8.14 (Simplified Gaussian distributional regression). We consider four simplifications to the Gaussian distributional regression model in Model 8.11. These simplifications are as follows:

- a. μ_{jst} does not depend on synthesis method

$$\forall s \beta_{2,s} = 0.$$

- b. σ_{jst} does not depend on synthesis method

$$\forall s \gamma_{2,s} = 0.$$

- c. σ_{jst} only depends on whether the attribute is numeric or categorical, and does not depend on the specific variation of CART, random forest, or regression synthesis model.

$$\begin{aligned} \gamma_{3,t} : t = 1, \dots, n_t &\rightarrow \gamma_{3,t'} : t' = 1, 2; \\ \gamma_{2,s} : s = 1, \dots, n_s &\rightarrow \gamma_{1,s'} : s' = 1, \dots, 3; \text{ and} \\ \gamma_{4,(j,t)} : t = 1, \dots, n_t &\rightarrow \gamma_{4,(j,t')} : t' = 1, 2. \end{aligned}$$

- d. Placing a regularised horseshoe prior (see Section 2.1.2) on the synthesis method

coefficients for μ_{jst} ⁵.

$$\begin{aligned}\beta_{2,s} | \lambda_s, \tau, c &\sim \mathcal{N}(0, \tau^2 \bar{\lambda}_s^2), & \lambda_s &\sim \text{C}^+(0, 1), \\ \tau | \sigma &\sim \text{C}^+(0, \tau_0^2), & \tau_0 &= \frac{1}{5} \frac{\sigma}{\sqrt{n}}, & c^2 &\sim \text{Inv-Gamma}(2, 8), \\ \gamma &\stackrel{\text{iid}}{\sim} \mathcal{N}(0, 0.5^2) & \text{and} & \pi &\sim \mathcal{N}(0, 0.5^2).\end{aligned}$$

We do encounter issues with divergent transitions when fitting the Gaussian distributional model with a horseshoe prior (Model 8.14d). Recall that issues are common with the horseshoe prior due to the difficult posterior geometry that it can induce (Betancourt, 2021). We attempt to resolve this by experimenting with centered and non-centered parameterisations for the model coefficients, as this can improve the posterior geometry (Betancourt, 2021). We find that most parameterisations have no effect and that divergent transitions can not be completely eliminated. However, a non-centered parameterisation for $\beta_{2,s}$ and a centered parameterisation for $\gamma_{2,s}$ results in the fewest divergent transitions. Finally, we resolve to brute forcing the sampling from the posterior distribution. We drastically reduce the step size of the sampler and increasing the warm-up phase to 5000 iterations. This helped to reduce the number of divergent transitions, but there were still 6 divergent transitions in the final sample of 4000 draws. As such, posterior samples from Model 8.14d will be biased.

The final inference models that we will consider for the disclosure scores incorporate a hierarchical structure for the subjects. Recall, that we considered the alternative approach of averaging the attribute disclosure scores over all replications, see Equation (8.3). Furthermore, recall our question about the disclosure risk of outliers, see Question 8.2.1e. In general, we expect that some subjects in the data will be more vulnerable to disclosures than others. As such, we would prefer to model attribute scores at the subject-level and test this.

We calculate local outlier factor (see Equation (6.7)) for the observations in training data, as a measure of how outlying they are. An initial exploration of the relationship between y_{ijst} and $\log(\text{lof}_i)$ indicates a weak positive correlation (Figure 8.7). Therefore, we will fit a Gaussian hierarchical distributional regression to the data. We run this model for 6 chains with 6000 warm-up iterations and 5000 sampling iterations per chain.

⁵When using the regularised horseshoe prior, we specify narrower prior distributions for the scale coefficients as it was found to reduce the number of divergent transitions.

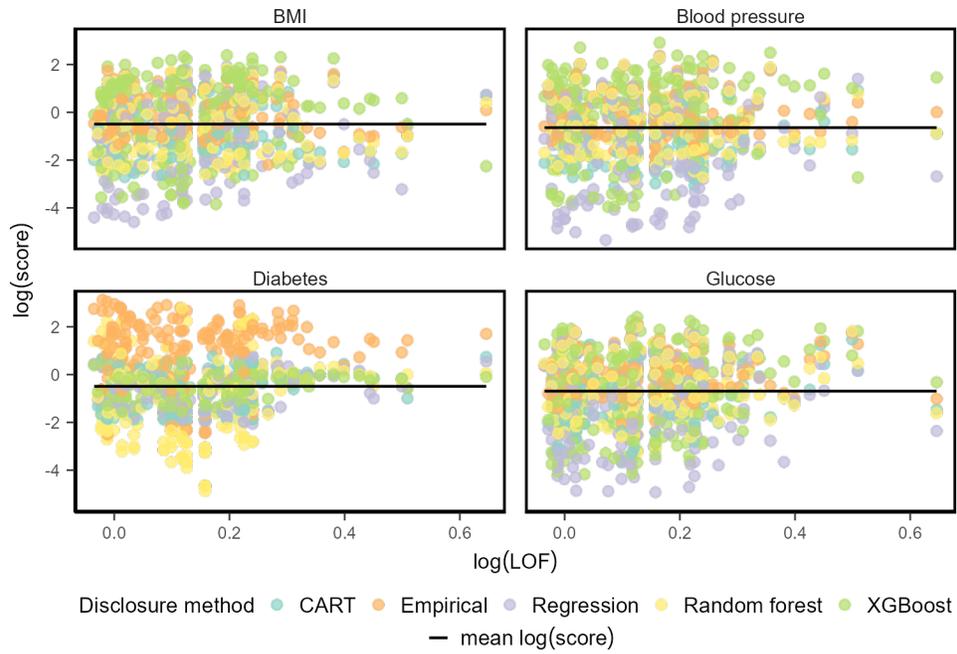


Figure 8.7: $\log(\text{lof}_i)$ against y_{ijst} for a small sample of the attribute disclosure scores.

Model 8.15 (Gaussian hierarchical distributional).

$$\begin{aligned} \log y_{ijst} | \alpha, \beta, \gamma, \pi, \sigma_b &\sim \mathcal{N}(\mu_{ijst} + b_{[i]}, \sigma_{jst}^2), \\ \mu_{ijst} &= \alpha + \beta_{1,j} + \beta_{2,s} + \beta_{3,t} + \beta_{4,(j,t)} + x_{5,i} \beta_5, \\ \log \sigma_{jst} &= \pi + \gamma_{1,j} + \gamma_{2,s} + \gamma_{3,t} + \gamma_{4,(j,t)}. \end{aligned}$$

With weakly informative parameters priors

$$\begin{aligned} \alpha &\sim \mathcal{N}(0, 2.5^2), \quad \beta \sim \mathcal{N}(0, 2.5^2), \\ \pi &\sim \text{student-t}_3(0, 2.5), \quad \gamma \sim \mathcal{N}(0, 1^2), \\ b_{[i]} &\sim \mathcal{N}(0, \sigma_b^2) \quad \text{and} \quad \sigma_b \sim \text{Exponential}(1). \end{aligned}$$

where $x_{5,i}$ is $\log(\text{lof}_i)$.

Note, these are narrower priors than for previous models but we find that they are necessary for convergence. We check the prior predictive distribution for the model and verify that the distribution is many orders of magnitude more varied than the data. In fact, we could improve training by selecting more informative priors. However, this is difficult for such a complex model so we prefer to keep our very weakly informative choices.

Given the complexity of Model 8.15 we will explore several choices for simpler models.

Model 8.16 (Simplified Gaussian hierarchical distributional regression). We consider the following simplifications to the Gaussian hierarchical distributional regression model

(Model 8.15).

- a. σ_{jst} only depends on whether the attribute is numeric or categorical, and does not depend on the specific variation of CART, random forest or regression synthesis model.

$$\begin{aligned} \gamma_{3,t} : t = 1, \dots, n_t &\rightarrow \gamma_{3,t'} : t' = 1, 2; \\ \gamma_{2,s} : s = 1, \dots, n_s &\rightarrow \gamma_{1,s'} : s' = 1, \dots, 3; \text{ and} \\ \gamma_{4,(j,t)} : t = 1, \dots, n_t &\rightarrow \gamma_{4,(j,t')} : t' = 1, 2. \end{aligned}$$

- b. attribute disclosure score does not depend on local outlier factor (LOF)

$$\beta_5 = 0.$$

- c. A hierarchical non-distributional regression model

$$\sigma_{jst}^2 = \sigma^2,$$

where

$$\sigma \sim \text{Exponential}(1).$$

Most of these simplifications are similar to the simplified Gaussian distributional regression (Model 8.14), so we will only discuss the notable changes. We removed the horseshoe prior simplification (Model 8.14d), which we found to be very difficult to sample from. The hierarchical model is more difficult to train, so these problems would only become worse.

Recall, that the exploratory analysis of LOF indicated that the effect on y_{ijst} was quite weak (see Figure 8.7). Furthermore, with the inclusion of a hierarchical term on each subject, LOF of the subjects may not provide much additional information. Therefore, we consider the simplification of removing the LOF term (Model 8.16b). Also, we compare the fit to a non-distributional hierarchical model (Model 8.16c) but we do not expect this to fit well.

Membership disclosure

We evaluate membership disclosure risk in the scenario where the hypothetical attacker knows the age and number of pregnancies of all training subjects, and also when they

know all variables for the training subjects. We calculate membership predictions for the entire training set, and also separately for the training set inliers and outliers. We plot precision-recall curves for all sets of membership predictions and we test for significant differences in the membership disclosure risk of outliers by fitting a regression model to the area under the precision-recall curves for both inliers and outliers.

Let y_{rst} be the area under the precision recall curve for the r^{th} replication of the s^{th} synthetic data generation method, calculated for the training set inliers ($t = 0$) and outliers ($t = 1$). We model y_{rst} with the hierarchical linear regression model:

$$y_{rst} \sim \mathcal{N}(\mu_{st}, \sigma_\epsilon^2), \quad (8.5)$$

where

$$\mu_{js} = \alpha + \alpha_{[s]} + \beta_t,$$

for

$$r = 1, \dots, N_r; \quad s = 1, \dots, N_s; \quad t = 0, 1; \quad \beta_0 = 0;$$

with weakly informative hierarchical priors

$$\begin{aligned} \alpha, \beta_1 &\sim \mathcal{N}(0, 2.5), \\ \alpha_{[s]} &\sim \mathcal{N}(0, \sigma_s^2), \text{ and} \\ \sigma_\epsilon, \sigma_s &\sim \text{Exponential}(1). \end{aligned} \quad (8.6)$$

8.3 Results

8.3.1 Results of utility assessments of synthetic Pima data

Consider the univariate plots of the Pima variables for 50 replications of the eight synthesis models (Figure 8.8). All synthetic data generation methods have accurately replicated the general distribution of most variables in the original data. However, the regression synthesised data (Model 8.1) fails to capture the bimodality of BMI and the slight right skew of glucose.

We explore the issue of skewed variables further by looking at quantile plots for DPF and pregnancies (Figure 8.9). These reveal that Model 8.1 generates more extreme values than we observe for the other datasets. For example, the regression data contain values such as 60, 100 and -1 pregnancies. The extreme values at the tails of the distributions are caused by a combination of two aspects of the regression synthesis approach. The first is that we synthesised the variable from a Gaussian distribution, which is unbounded. The second is that we applied the log transformation to right skewed variables. Reversing this transformation after synthesis causes the scale of the outliers at the upper end to be magnified. At the lower end, most inverse transformed variables are bounded by zero. However, recall that pregnancies contained zero values which we addressed by adding one

and then log transforming. As such, the inverse transformed values are lower bounded by negative one, which is why we observe a few -1 values.

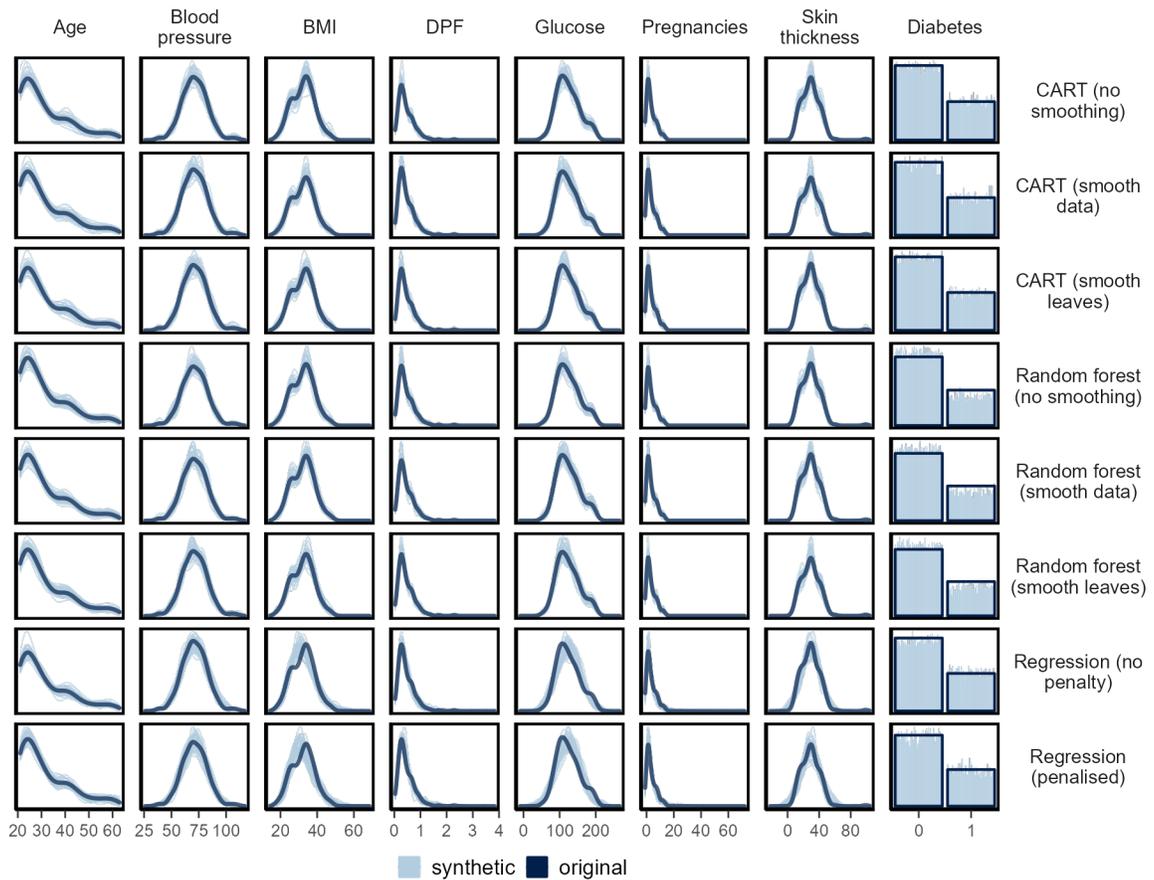


Figure 8.8: Univariate variable distributions for 50 replications of synthetic Pima data.

Recall that subjects with plasma glucose concentration ≥ 200 mg/dl were specifically excluded from the original data (see Section 8.1). However, plots of plasma glucose concentration show that the regression synthesised data contains subjects with plasma glucose concentration values that are above the 200 (Figure 8.10).

Overall, the distributions of the numeric variables for the regression synthesised data indicate a clear problem with model selection. The linear regression model is not able to model the tail distributions of the log transformed variables, and this problem is magnified on the original scale. Truncating the distributions of pregnancies and plasma glucose concentration would remove the impossible values, however, this would also introduce bias. These problems can be addressed in the future by choosing more appropriate models for variables with skewed, truncated and/or count distributions. Poisson regression, negative binomial regression, zero-inflated Poisson regression, zero-inflated negative binomial regression, CART and random forest models are all potential solutions (Kleinke & Reinecke, 2013).

We check whether relationships between variables are preserved in the synthetic data.

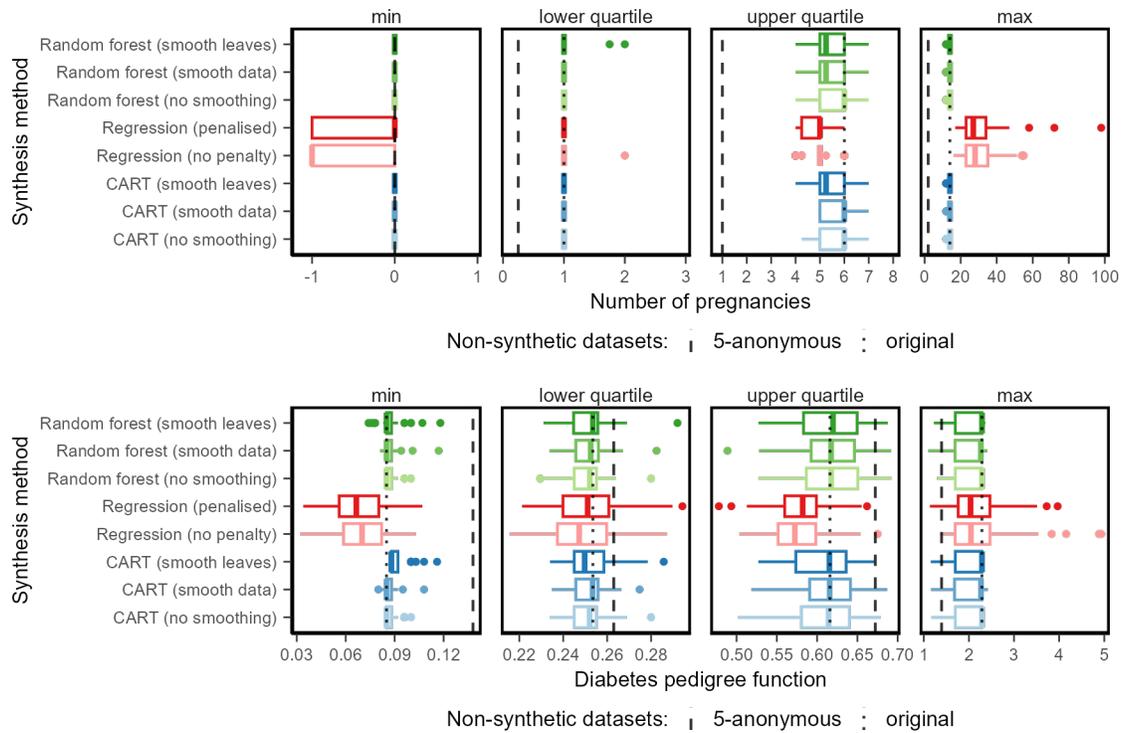


Figure 8.9: Synthetic data quantile distributions for pregnancies and diabetes pedigree function.

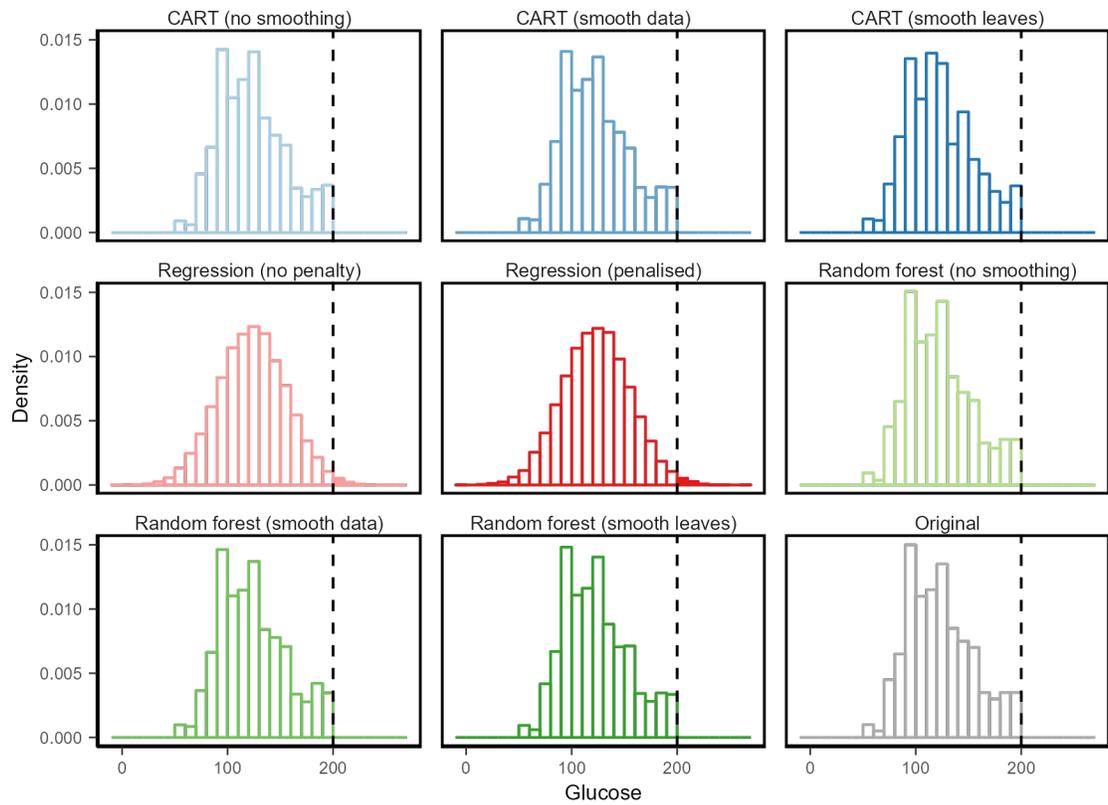


Figure 8.10: Distribution of synthesised glucose variables.

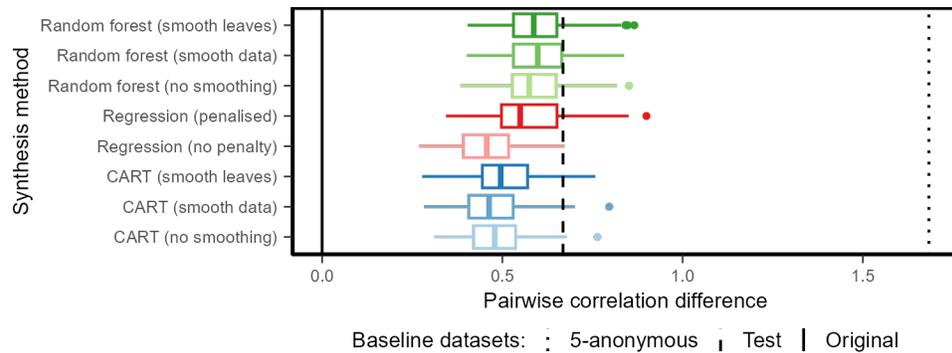


Figure 8.11: Pairwise correlation difference between original Pima data and 100 replications of data from eight synthesis methods.

The pairwise correlation differences (PCDs) for all eight data synthesis methods are significantly better than the 5-anonymised data. We can conclude that the removal of outlying values has weakened the linear associations between variables (Figure 8.11). In fact, PCDs for all eight data synthesis methods are slightly better than the test data, which has the same underlying distribution as the training data (see Section 8.2.2). As such, if the PCDs values were much lower, then we may have concluded that the models were overfit.

The dimension-wise prediction (DWP) results for all replications of synthetic data are very close to the the results for the original data (see Figure 8.12). This is the case for all three prediction models (see Models 8.4a, 8.4b and 8.4c). In comparison the results for the 5-anonymous data are significantly worse. We can conclude that the conditional relationships are better preserved in the synthetic data than the 5-anonymous data. However, recall that the near-perfect results can also be indicative of prediction models that are unable to model the differences between conditional distributions (see Section 5.3).

Now we consider the results of the propensity score assessment (see Section 8.2.2). Plots of pMSE ratio show weak but not conclusive evidence that the CART discriminator is able to distinguish the regression synthesised data from the training data (Figure 8.13a). The results for data synthesised from CART and random forest variants do not show any evidence that the discriminators can identify the synthetic data. However, recall that we would observe the same if the discriminators were poorly specified (see Section 5.2).

Task performance: inference

Now, we consider the model comparison results of the Pima inference task (Figure 8.14). Recall, the GAMM inference model (Model 8.7) fit the training data best. It was also the best fitting model for the majority of replications of all synthesis models. However, there are differences in how many replications it was best, and whether other methods were similarly well fitting. There were clearly more replications of CART synthesised data

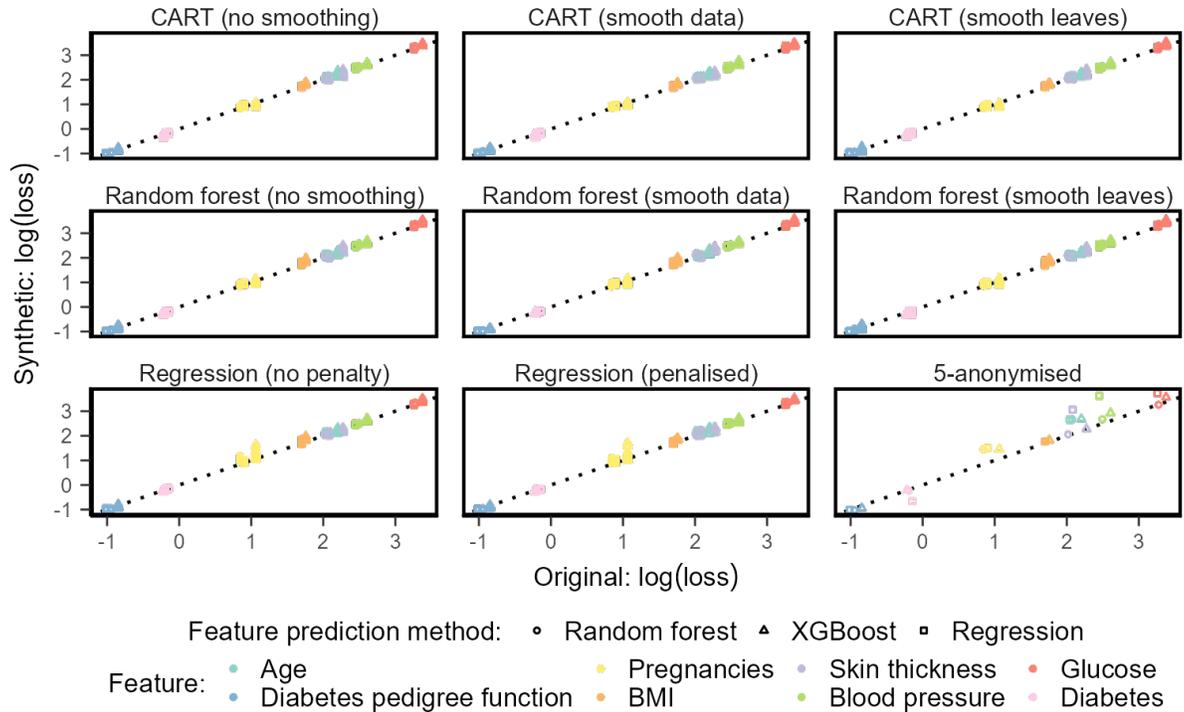
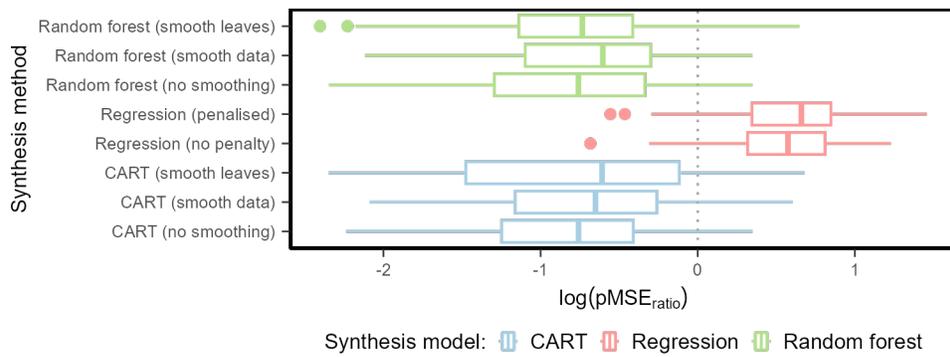


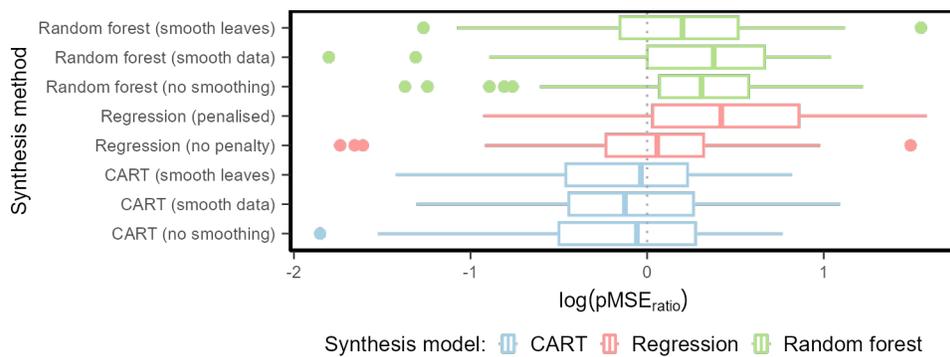
Figure 8.12: Dimension-wise prediction scores for 20 replications of synthetic datasets and 5-anonymised data. For three prediction models (see Models 8.4a, 8.4b and 8.4c).

(Model 8.2) that the GAMM was the best fitting model. Also, there were more replications and the GAMM was significantly better than the other inference models. Consequently, the CART synthesised data best aligns with the original data. Both variants of regression synthesised data (Model 8.1) performed particularly poorly on this comparison. In particular, penalised regression, for which there was no significant difference between any inference models in about 90% of replications. Such a result is perhaps not surprising. We would expect a logistic regression model to fit well to a variable that was synthesised with linear regression. In fact, it would be odd if the GAMM model was significantly better than the logistic regression models.

The 90% credible interval overlaps are similar for most synthesis methods. The notable exception is penalised regression, Model 8.1, which clearly has the least overlap for both inference models, see Figure 8.15 and Figure 8.16. The overlaps are slightly higher for the coefficients of the inference model with the horseshoe prior (Model 8.6). Our current theory for why this might be, is that the synthesis models replicate strong conditional relationships better than they replicate weak relationships. Therefore, the interval overlaps of the coefficients for the strongly related variables are larger than the weakly related variables. However, since the horseshoe prior shrinks weak coefficients towards zero, this improves the overlap. In comparison to the baseline datasets, the synthesis models all have lower overlap than the test baseline, on average. However, we do observe better overlaps

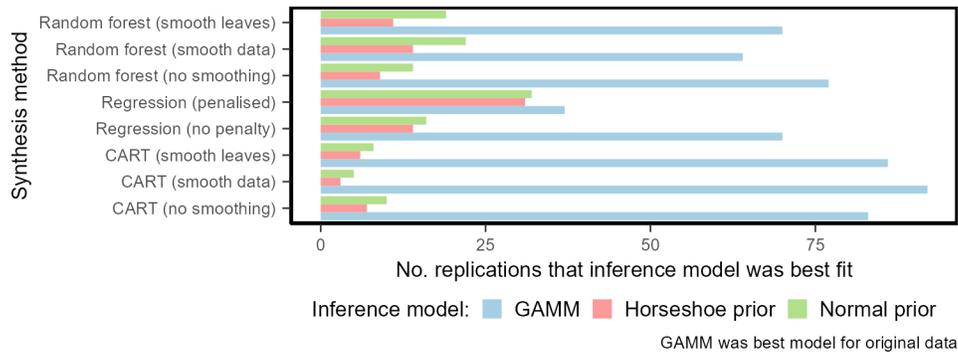


(a) CART discriminator

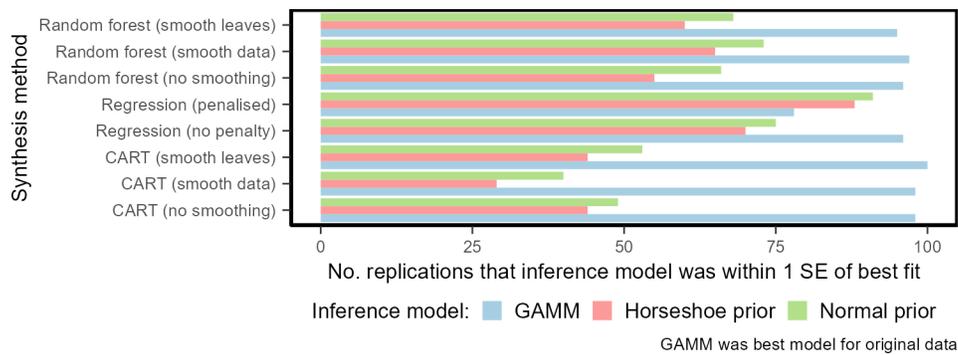


(b) Logistic regression discriminator

Figure 8.13: Boxplots of $\text{pMSE}_{\text{ratio}}$ scores for 100 replications of data generated with eight synthesis methods, calculated with both a CART and a logistic regression discriminator.



(a) Only the model with the largest ELPD.



(b) All models that were within 1 standard error of largest ELPD.

Figure 8.14: The number of synthetic Pima replications that each inference model was the best fit.

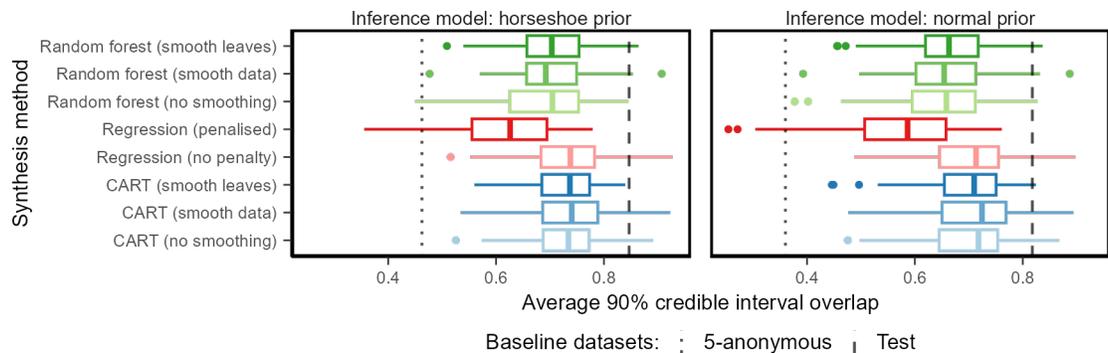


Figure 8.15: 90% credible interval overlap for Pima inference, Model 8.5 and Model 8.6, averaged over all coefficients.

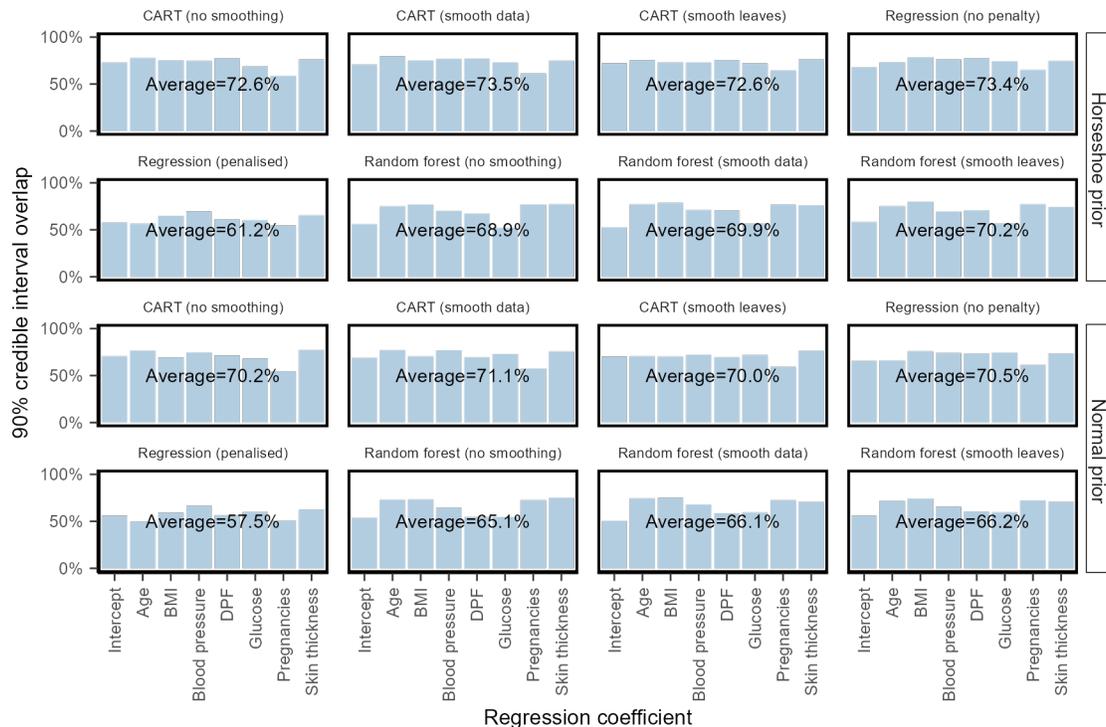


Figure 8.16: 90% credible interval overlap for Pima inference, Model 8.5 and Model 8.6, averaged over all replications.

for some replications, Meanwhile, the interval overlap is very low for the 5-anonymised data. In fact, it is significantly lower than even the penalised regression.

Let us now compare the results of hypothesis tests for the Pima inference models, see Figure 8.17. First, we consider the inference model with the normal prior (Model 8.5). In this case, the regression synthesised datasets (Model 8.1) are most aligned with the original data. In contrast, for the inference model with the horseshoe prior (Model 8.6), the CART synthesised data (Model 8.2) aligns most with the original data. However, we can see that there is another underlying pattern here. Let us consider the hypotheses tests for both inference models where we reject the null hypothesis. For tests that reject the null, the regression synthesised replications tend to align better. On the other hand, if the hypothesis test result is a failure to reject the null, then the CART synthesised replications tend to align better. The reason that we see the regression synthesised data performing better for the normal prior is because coefficients are more likely to be significant. Conversely, coefficients are less likely to be found significant for the horseshoe prior, so the CART synthesised data performs better.

These results have highlighted that the comparison of hypothesis test results is very sensitive to the hypothesis test. For our example, the conclusion of the comparison are strongly influenced by the choice of inference model. This represents a major flaw in the comparison of hypothesis test results. As such, it is also worth questioning if other utility

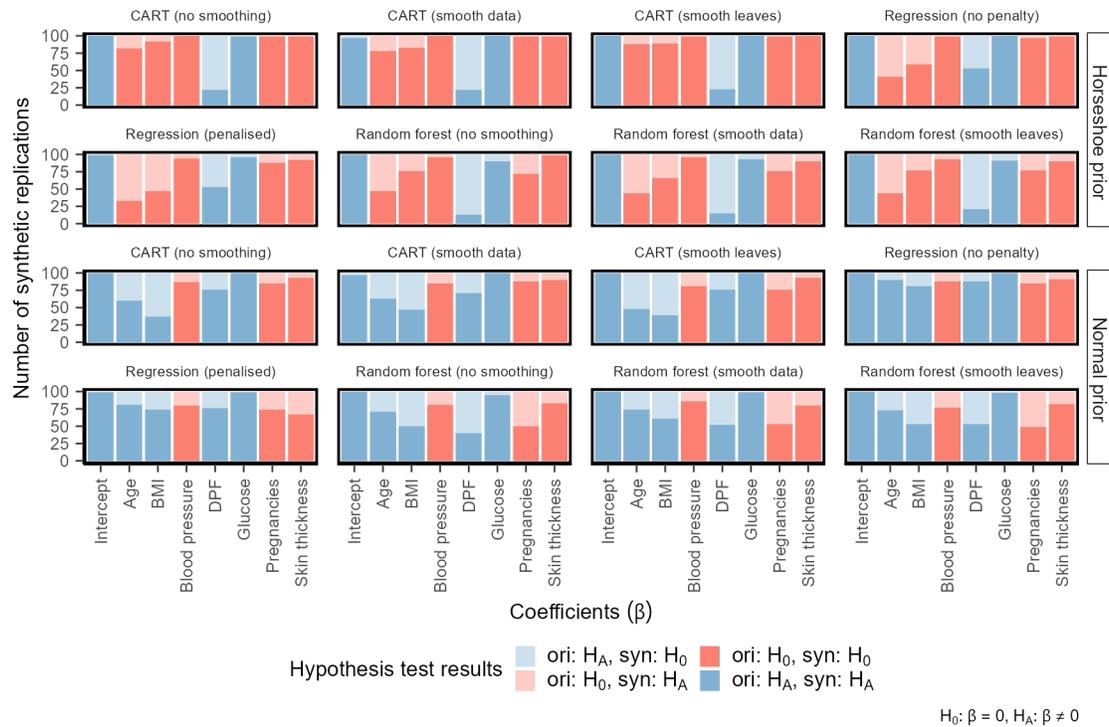


Figure 8.17: The number of replications of synthetic data for which the hypothesis test result aligns with the original data.

assessments that are based on hypothesis tests, such as severity ratings, are of value (Taub et al., 2020).

8.3.2 Results of disclosure risk assessments of synthetic Pima data

Risk of membership disclosure and memorisation

The results of the membership disclosure assessment show that the risk of membership disclosures for observations in the training data when the intruder has access to synthetic data depend on the synthetic data generation methods, the amount of prior knowledge the intruder has of the real observations, and whether the training observations are outliers. When the intruder knows all variables for the real observations, they are able to reliably classify the few training observations that are extremely close to the synthetic CART or random forest synthesised data. More often than not, the intruder can correctly classify observations as training at larger distance thresholds (> 1) (Figure 8.18). When the intruder only knows the quasi-identifying variables (age and number of pregnancies), they are still able to correctly classify training observations about as well as they were for all variables at thresholds larger than 1, but they can not classify nearly as reliably at very small thresholds. The risk of membership disclosures are significantly lower when the intruder attempts to match on regression synthesised data, regardless of whether they

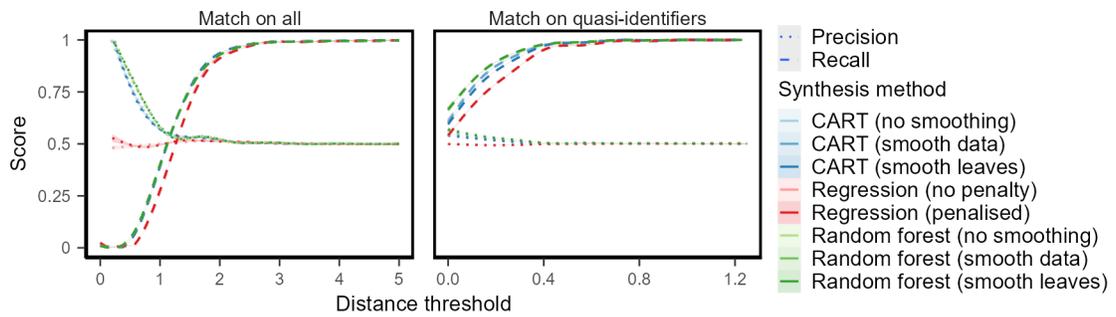


Figure 8.18: Precision-recall curves for changing intruder prior knowledge.

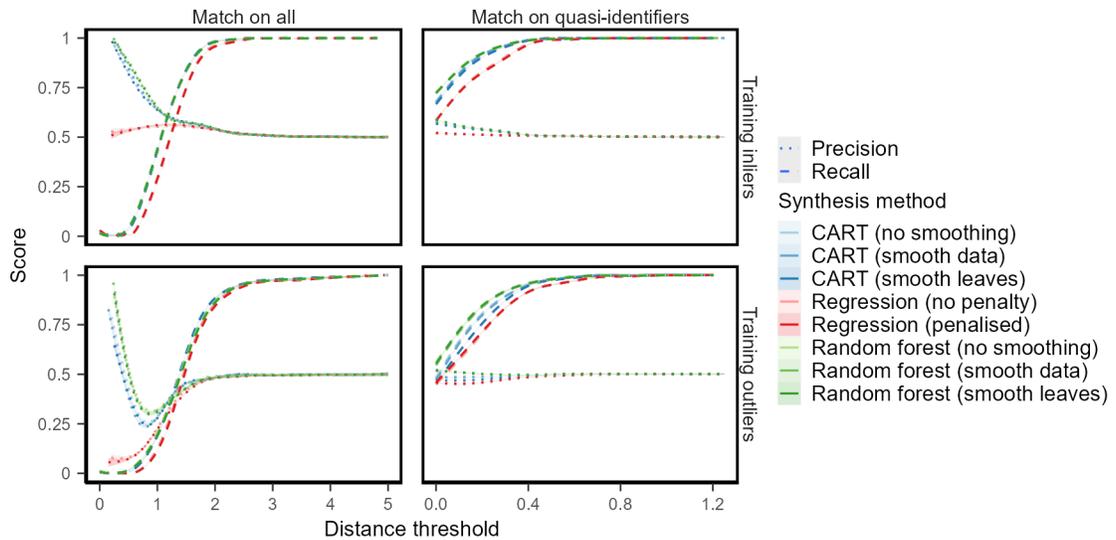


Figure 8.19: Precision-recall curves for changing intruder prior knowledge and training subsets.

only know quasi-identifiers or all variables.

The membership disclosure risk differs for inliers and outliers. Fitting the regression model to the area under the precision-recall curves of inliers and outliers indicates that the membership disclosure risk of inliers is significantly higher when matching on all variables, but the disclosure risk of outliers is significantly higher when matching on the quasi-identifiers (Figure 8.20 and Table 8.2). Looking at the separate precision-recall curves for the inliers and outliers (Figure 8.19), we can see that the difference between area under the precision-recall curve when matching on all variables is mostly explained by lower precision scores at lower distance thresholds. It appears that the synthetic data generation methods tend to generate observations that are closer to average, and so the synthetic observations are closer to the test observations than the training observations.

Looking at the closest training and synthetic pairs for each data synthesis method when matching on all variables (Table 8.3), the non-smoothed CART and random forest observations are identical to the closest training observation. Applying smoothing to the

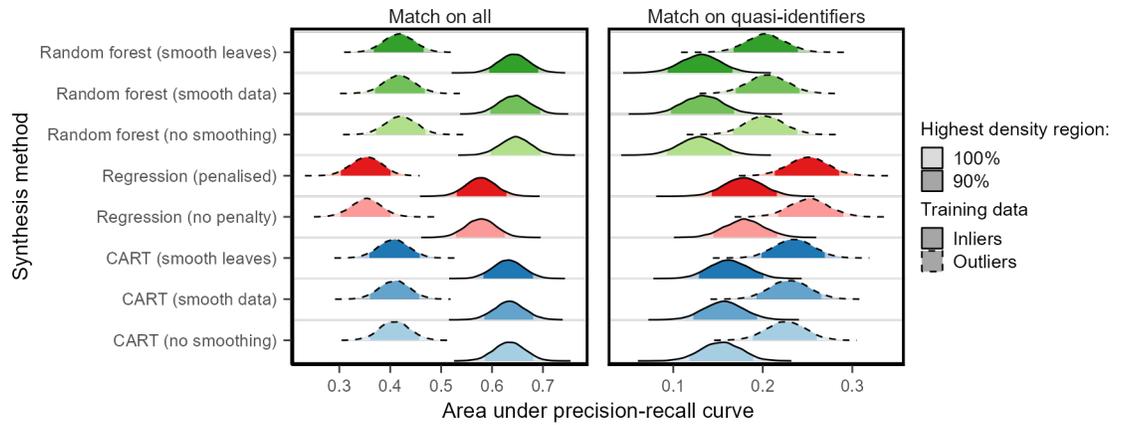


Figure 8.20: PPD of area under precision-recall curves for changing intruder prior knowledge and training subsets.

Table 8.2: Differences between area under precision-recall curve of inliers (β_0) and outliers (β_1), from the model in Equation (8.5) for both intruder prior knowledge scenarios.

Intruder knowledge	$\mathbb{E}(\beta_1 - \beta_0)$	90% HDPI $\beta_1 - \beta_0$
All variables	-0.225	(-0.228, -0.223)
Quasi-identifiers	0.073	(0.071, 0.074)

numeric variables introduces small differences between the synthetic and real observations. However, these differences are negligible and do not appear to affect the risk of membership disclosure. The situation is especially bad for the random forest synthesised datasets, where there were many more identical pairs of training and synthetic data than could be shown (Table A.1). In contrast, the regression synthesised observations are similar but not identical to the training observations. It is important to note that we would expect the closest of those pairs to be fairly similar because there were 4 million possible pairs of training and synthetic observations for each synthetic method shown. While the closest test and synthetic pairs (6.64 million possible pairs) are not identical, they are still very similar (Table 8.4).

Table 8.3: A subset[†] of the closest pairs of training (original) and synthetic Pima observations for each data synthesis method when matching on all variables.

Dataset	Distance	Preg.	Gluc.	BP	Skin.	BMI	DPF	Age	Diab.
Training	6.68E-03	1	107	68	19	26.5	0.165	24	0
Regression (pen.)	6.68E-03	1	108	67	18	27.3	0.131	23	0
Training	8.34E-03	1	111	62	13	24.0	0.138	23	0
Regression (no pen.)	8.34E-03	1	112	62	11	24.7	0.176	24	0
Training	1.10E-03	1	96	64	27	33.2	0.289	21	0
CART (smooth leaves)	1.10E-03	1	96	64	28	33.2	0.277	21	0
Training	2.12E-05	1	96	64	27	33.2	0.289	21	0
CART (smooth data)	2.12E-05	1	96	64	27	33.2	0.293	21	0
Training	0	1	99	58	10	25.4	0.551	21	0
CART (no smooth)	0	1	99	58	10	25.4	0.551	21	0
Training	0	9	152	78	34	34.2	0.893	33	1
RF (smooth leaves)	0	9	152	78	34	34.2	0.893	33	1
Training	1.32E-06	4	154	72	29	31.3	0.338	37	0
RF (smooth data)	1.32E-06	4	154	72	29	31.3	0.337	37	0
Training	0	5	139	80	35	31.6	0.361	25	1
RF (no smooth)	0	5	139	80	35	31.6	0.361	25	1

Table 8.4: A subset of the closest pairs of test and synthetic Pima observations for each data synthesis method when matching on all variables.

Dataset	Distance	Preg.	Gluc.	BP	Skin.	BMI	DPF	Age	Diab.
Testing	3.14E-03	1	89	66	23	28.1	0.167	21	0
Regression (pen.)	3.14E-03	1	88	66	24	27.6	0.167	22	0
Testing	4.46E-03	1	73	50	10	23.0	0.248	21	0
Regression (no pen.)	4.46E-03	1	73	49	9	22.0	0.253	21	0
Testing	7.00E-03	1	95	60	18	23.9	0.260	22	0
CART (smooth leaves)	7.00E-03	1	98	60	18	22.7	0.297	21	0
Testing	5.17E-03	7	102	74	40	37.2	0.204	45	0
CART (smooth data)	5.17E-03	7	100	74	39	36.6	0.242	46	0
Testing	4.65E-03	1	95	60	18	23.9	0.260	22	0
CART (no smooth)	4.65E-03	1	99	60	19	24.0	0.252	21	0
Testing	5.69E-03	1	97	66	15	23.2	0.487	22	0
RF (smooth leaves)	5.69E-03	1	96	68	14	22.5	0.493	22	0
Testing	7.08E-03	0	105	64	41	41.5	0.173	22	0
RF (smooth data)	7.08E-03	0	106	66	40	41.1	0.207	21	0
Testing	5.55E-03	2	107	74	30	33.6	0.404	23	0
RF (no smooth)	5.55E-03	2	107	75	29	34.2	0.364	24	0

[†] Table A.1 in Appendix A includes 14 pairs of random forest (no smooth), 2 pairs of random forest (smooth data), and 9 pairs of random forest (smooth leaves) that were equidistant to those in the Table 8.3.

Risk of attribute disclosure

Now we explore the results from our scenario modelling the attribute disclosure risk of a motivated intruder (see Section 8.2.3). First, we shall consider the two linear regression models (see Models 8.9 and 8.10). The chains for each model ran without issues and both trace plots and \hat{R} values ($\hat{R} < 1.01$) showed convergence. ELPD indicates that the full model is a significantly better fit to the data (Table 8.5). Therefore, despite the small difference we observed between observations for between different synthesis models, their coefficients do improve the fit. Most of the posterior samples from each model have been excluded for brevity but they showed that the model fit the data well. However, we can see that combinations of attribute prediction methods and attribute are not modelled well (Figure 8.21). Since the simplified data is very small (120 observations) and not identically distributed, we would not expect it to be able to model such fine aspects of the data. That said, we can clearly see that the lower variance of the regression prediction model and some other panels are not well modelled.

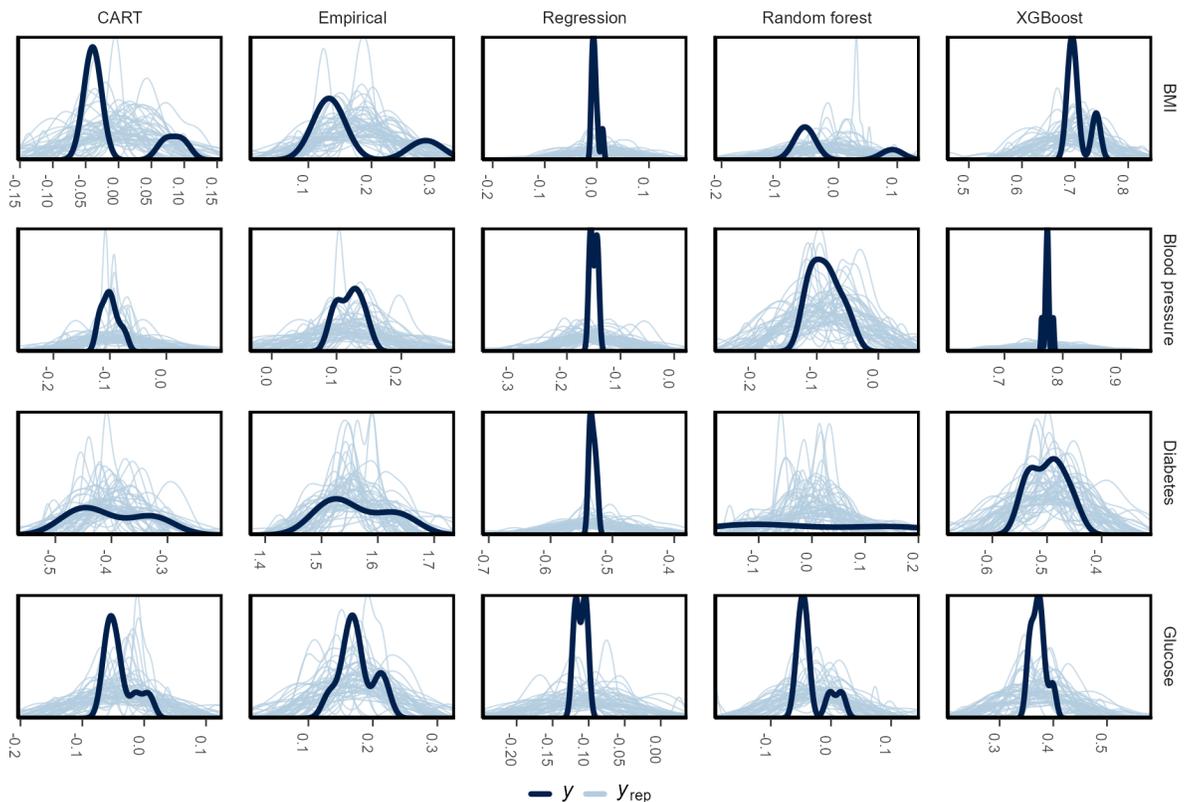


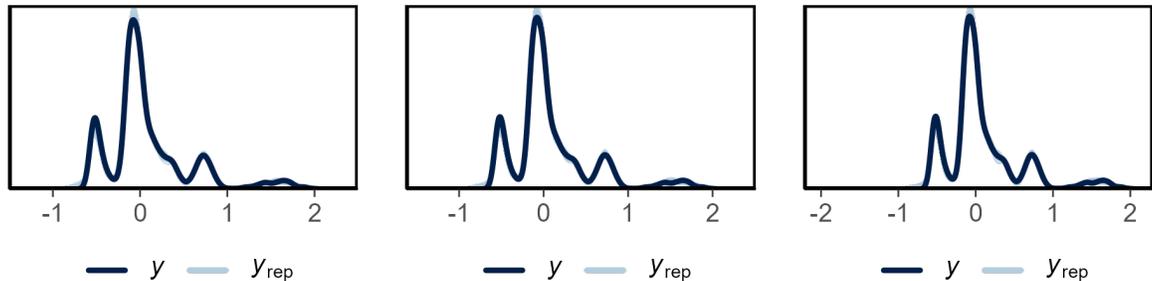
Figure 8.21: Posterior samples from linear regression (Model 8.9) grouped by attribute and attribute disclosure method.

Next, we consider the Gaussian, log normal, and gamma distributional regression models (see Models 8.11, 8.12 and 8.13 respectively). All trained well, with all \hat{R} values well below 1.01 and trace plots indicating good mixing of chains. Overall, it is difficult to

Table 8.5: ELPD difference for linear regression models.

	ELPD difference	SE difference
Full simple model (Model 8.9)	0.0	0.0
No synthesis method (Model 8.10)	-5.9	4.4

discern much difference between posterior distributions of the three models, since each fits the observed data well (Figure 8.22). Separating the posterior distribution plots by attribute and attribute prediction method really highlights the flexibility of the distributional model (Figure 8.23)⁶. Recall, that the linear regression model was unable to model differences in variance (Figure 8.21). In contrast, the distributional model can capture the difference in variances that we see for some groups. There are still some difficult values at the upper tail of scores for CART and diabetes attribute. It would be very difficult to find a regression model that could fit those without overfitting to the data.



(a) Gaussian (Model 8.11). (b) Log-normal (Model 8.12). (c) Gamma (Model 8.13).

Figure 8.22: Posterior distributions $\tilde{y}_{jrst}|y_{jrst}$ of distributional models.

The ELPD of the Gaussian distributional model was significantly larger than the gamma and log-normal (Table 8.6). As such, we will only consider simplifications for the Gaussian distributional model (Model 8.14). Of these simplifications, the models with either collapsed categories (Model 8.14c) or horseshoe shrinkage priors (Model 8.14d) both essentially fit the data as well as the full model (Table 8.7). The models without synthesis method (Model 8.14a and Model 8.14b) fit significantly worse. Recall that the model with the horseshoe shrinkage priors was difficult to train and contained divergent transition whereas the model with collapsed categories trained without issue. Therefore the model with collapsed categories (Model 8.14c) is our preferred distributional model for the attribute disclosure scores.

⁶We omit the corresponding figures for the log-normal and gamma distributional models, as they are virtually indistinguishable.

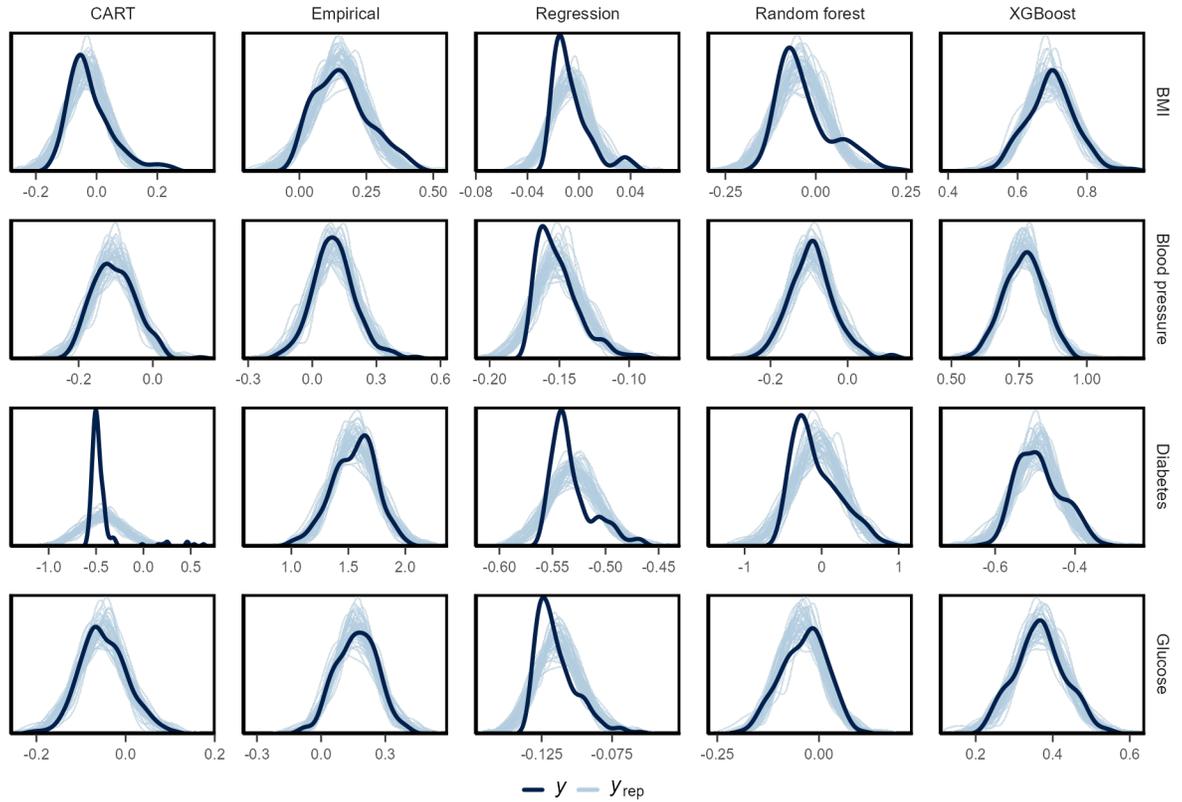


Figure 8.23: Posterior distribution $\tilde{y}_{jrst}|y_{jrst}$ from normal distributional model (see Model 8.11), grouped by target variable and disclosure method.

Table 8.6: ELPD difference for Gaussian, log normal, and gamma distributional models.

	ELPD difference	SE difference
Gaussian distributional (Model 8.11)	0.0	0.0
Log-normal distributional (Model 8.12)	-0.7	0.4
Gamma distributional (Model 8.13)	-44.1	4.9

Table 8.7: ELPD difference of full and simplified Gaussian distributional models.

Simplification	ELPD difference	SE difference
Collapsed categories for $\gamma_{2,s}, \gamma_{3,t}, \gamma_{4,(j,t)}$ (Model 8.14c)	0.0	0.0
Regularised horseshoe prior (Model 8.14d)	-0.1	6.9
Full model (Model 8.11)	-0.4	6.6
σ_{jst} not dependent on synthesis method (Model 8.14b)	-36.9	11.2
μ_{jst} not dependent on synthesis method (Model 8.14a)	-43.2	13.0

Now, we explore the Gaussian hierarchical distributional models (Model 8.15). They trained well with \hat{R} values ($\hat{R} < 1.04$) and trace plots indicating good mixing of chains. Posterior draws fit the data reasonably well for most combinations of the attribute prediction method and target variables, and for individual subjects (see Figure 8.24 and Figure 8.25, respectively). The scale of the posterior draws is generally good. There are some clear bimodalities in the data that have not modelled. For example, the combination of CART attribute prediction method and diabetes attribute was also problematic for the Gaussian distributional model (Figure 8.23). However, in general the model fits a challenging distribution quite well.

Consider the results for simplifications to the hierarchical distributional model (Table 8.8). The difference between the full model and model without LOF are clearly the best fitting, with little to separate them. Usually, we would prefer the more parsimonious model between two models that fit nearly equally well. However, one of our questions of interest is about whether outliers have a higher disclosure risk (Question 8.2.1e). Therefore, in this case we prefer hierarchical distributional model that includes LOF (Model 8.15).

Table 8.8: ELPD difference of full and simplified Gaussian hierarchical distributional models.

Simplification	ELPD difference	SE difference
Full model (Model 8.15)	0.0	0.0
Without local outlier factor (Model 8.16b)	-0.2	0.3
Collapsed categories for $\gamma_{2,s}, \gamma_{3,t}, \gamma_{4,(j,t)}$ (Model 8.16a)	-3656.3	91.4
Non-distributional (Model 8.16c)	-4348.8	115.3

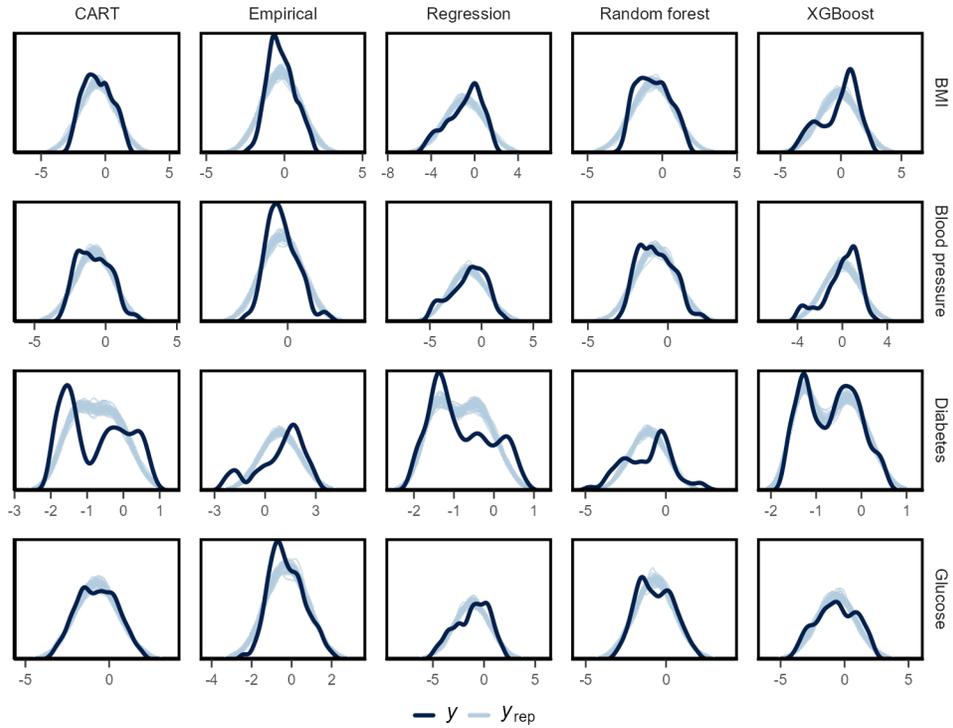


Figure 8.24: Posterior distribution of Gaussian hierarchical distributional (Model 8.11), grouped by target variable and disclosure method.

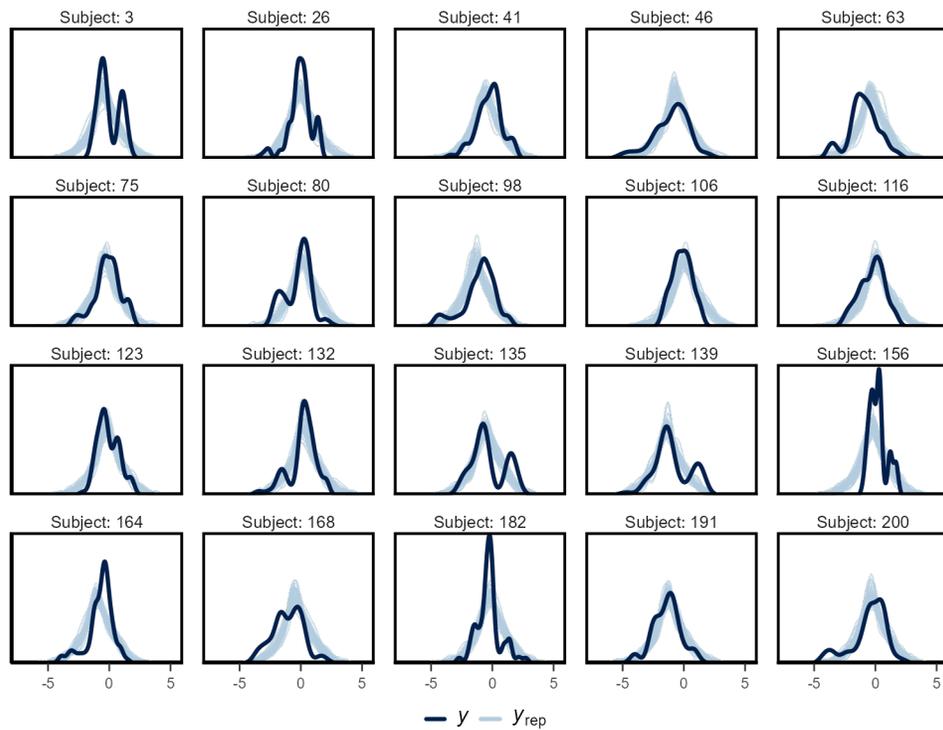


Figure 8.25: Posterior distribution of Gaussian hierarchical distributional (Model 8.11), grouped by subject.

Results: Evaluating the risk of disclosure

We use the best fitting non-hierarchical and hierarchical models from the previous section (Models 8.14c and 8.15) to answer the questions of interest posed (see Questions 8.2.1). We begin by exploring which attribute prediction method results in the highest risk of disclosure (Question 8.2.1a). That is, of the five methods that we explored, which represents the “best” intruder? We first consider the posterior predictive distribution (PPD) of the hierarchical model (Figure 8.26b). However, there is little evidence of any differences between attribute disclosure methods. We consider the PPD of the distributional regression model (Figure 8.26a). The CART, random forest and regression models (see Model 8.8a, Model 8.8b and Model 8.8c) were consistently good at predicting all variables. While, XGBoost and empirical matching were good at predicting the categorical and numeric variables, respectively, and terrible at predicting the other type. While there are clear differences between the disclosure risk scores for different prediction methods, it is difficult to define a “best” attribute prediction method.

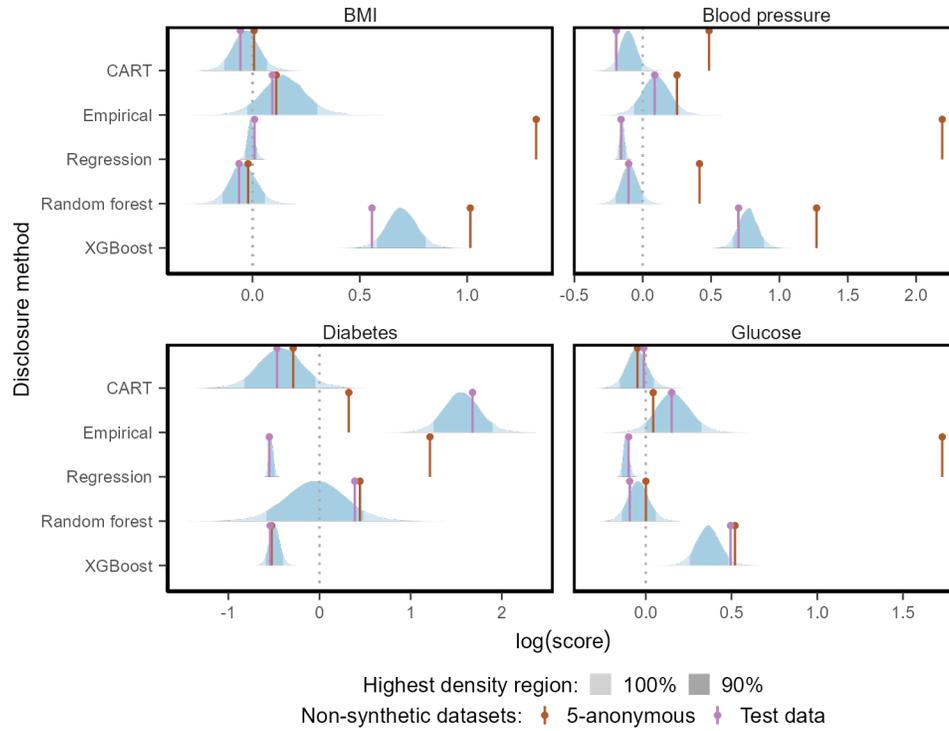
Now we move on to considering which synthesis models offer the best protection against disclosure (Question 8.2.1b). There is evidence that the disclosure risk of some synthesis methods is lower than others (Figure 8.27). The rankings of synthesis methods slightly change between the distributional regression and hierarchical distributional regression models. However, note that the scales of the coefficients for synthesis models are much smaller than the coefficients for attribute prediction methods. Consequently, despite there being evidence of a significant difference in disclosure risk between the synthetic datasets, the actual effect on disclosure risk is negligible. We can see that negligible difference when we look at the highest density posterior intervals (HDPIs) of the PPDs for each model (Figure 8.28). For both inference models (see Model 8.14c and Model 8.15), the HDPI of disclosure risk almost entirely depends on the attribute prediction method. As such, it is difficult to conclude that any synthesis model offers better disclosure risk than others.

The next question that we answer is whether smoothing or regularisation affect the risk of disclosure (Question 8.2.1c). There was very weak evidence from the distributional model that smoothing the leaves improved the disclosure risk of random forest, and stronger evidence that a weight penalty improved the disclosure risk of regression. However, as we discussed for Question 8.2.1b, the impact of any synthesis method is negligible.

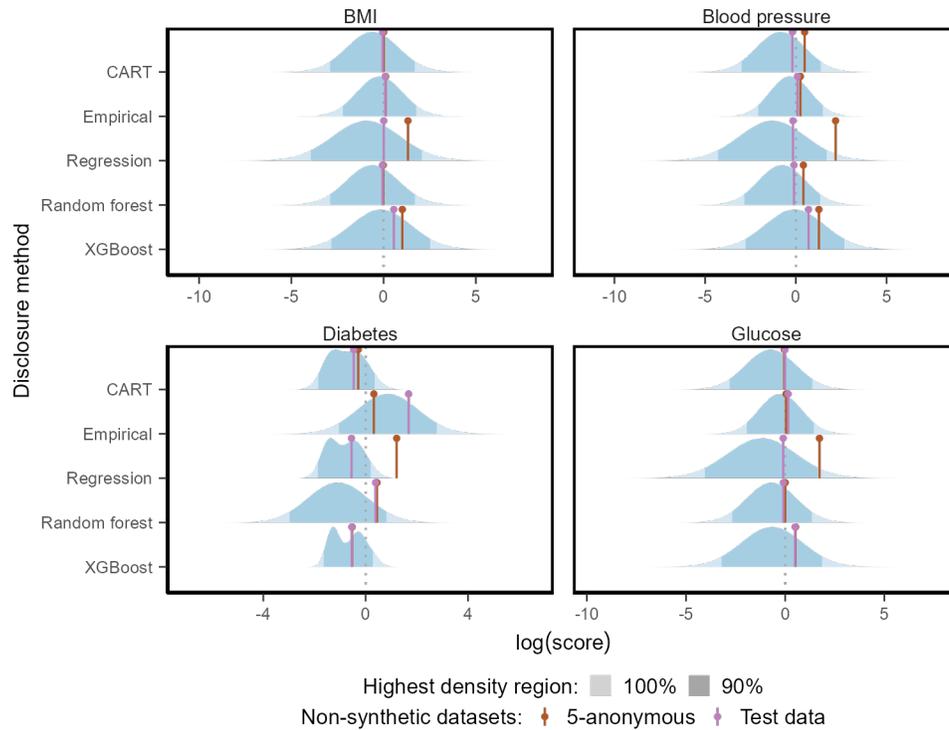
Let us now consider how the risk of disclosure of synthetic data compares to the test partition and 5-anonymous baseline (Question 8.2.1d). Recall, that the test baseline dataset contains no information about the training data (see Section 8.2.2). There is not any evidence that the synthetic datasets and test baseline have different disclosure risks (see Figure 8.26). This is a very positive result, as we have shown that the risk of disclosure of synthetic data is similar to a dataset that represents little disclosure risk.

The disclosure risk for the 5-anonymous baseline (see 3) is generally quite similar to the synthetic datasets. However, the disclosure risk of the 5-anonymous data is significantly lower if the intruder uses regression as a prediction method, or if the intruder is predicting blood pressure. Given the results in the utility assessment (see Section 8.3.1), it seems likely that these differences are due to the significantly lower quality of the k -anonymous data. Furthermore, a motivated intruder could recognise that regression is a particular poor attribute prediction method against some anonymisation procedures (see Winkler (2007)). In which case, they may implement a prediction method that is more robust against a variety of anonymisation procedures, such as CART or random forest. This would completely negate almost all the scenarios that k -anonymous data has a lower disclosure risk.

The final question to consider is whether the disclosure risk of outliers is higher (Question 8.2.1e). The attribute disclosure scores tended to be lower for observations that were more outlying, with a slightly positive relationship between local outlier factor and attribute disclosure score (Figure 8.30). On average, log local outlier factor increasing by 1 will correspond to a 1.15 increase in log attribute disclosure score (Figure 8.29). So in fact, our results show evidence that the disclosure risk of outliers is lower.

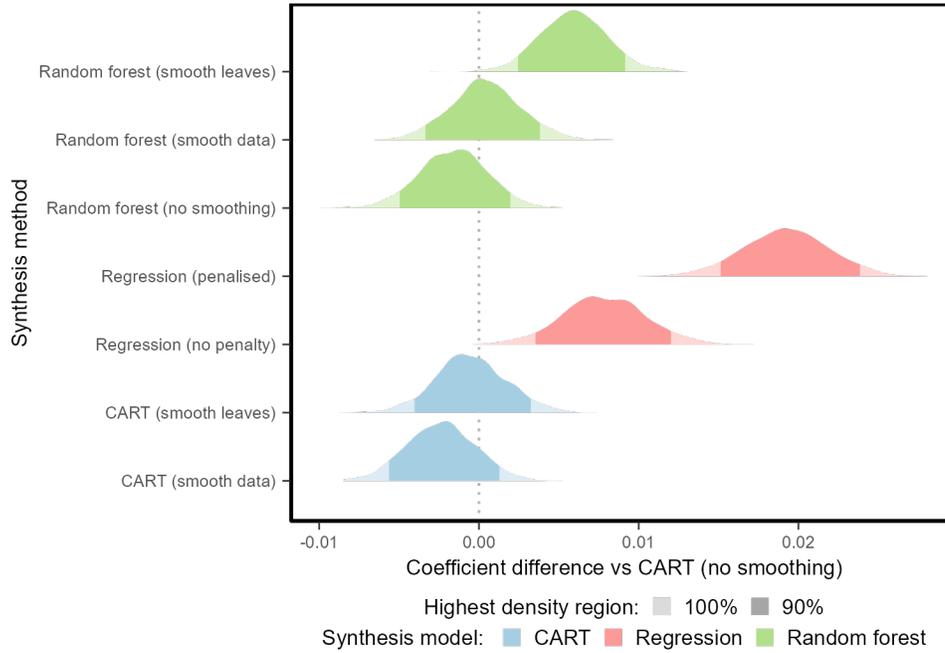


(a) Gaussian distributional regression (Model 8.14c).

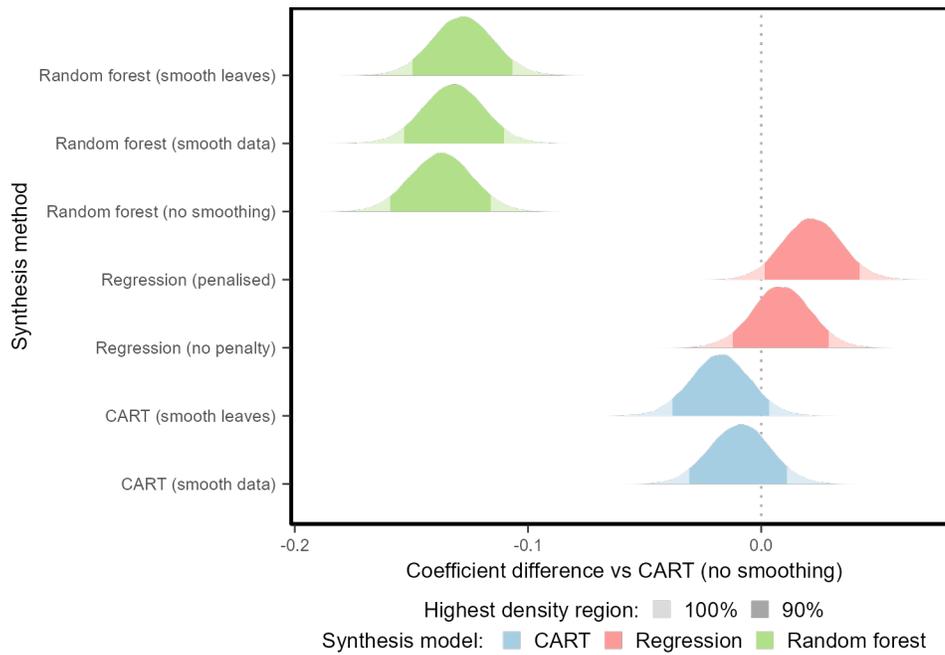


(b) Gaussian hierarchical distributional regression (Model 8.15).

Figure 8.26: Posterior distributions of attribute prediction method coefficient, for each target variable and disclosure method.

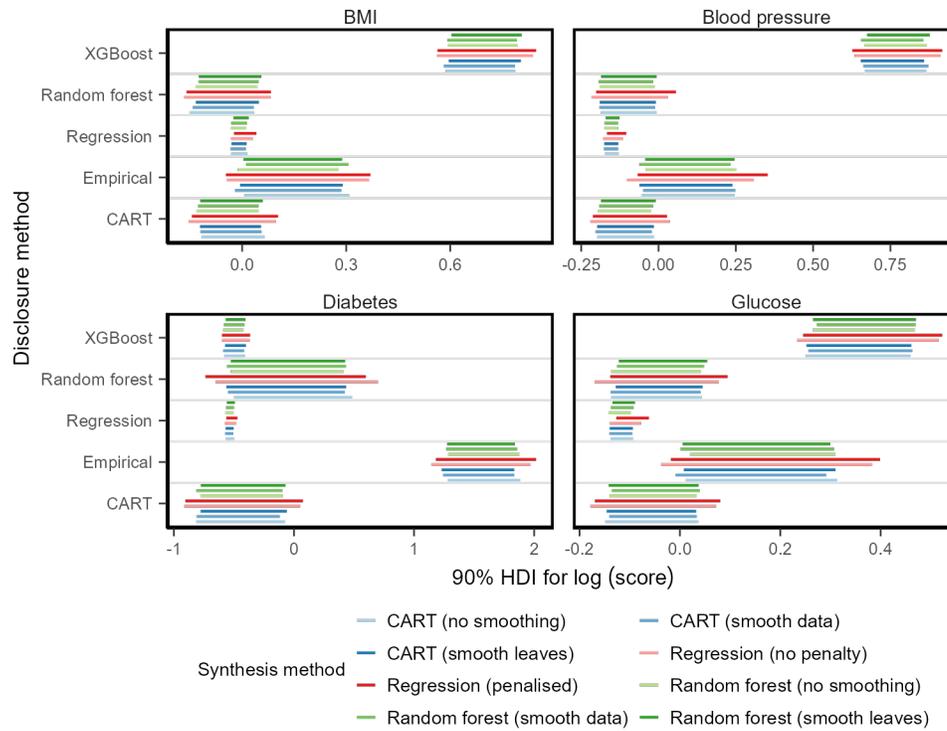


(a) Gaussian distributional regression.

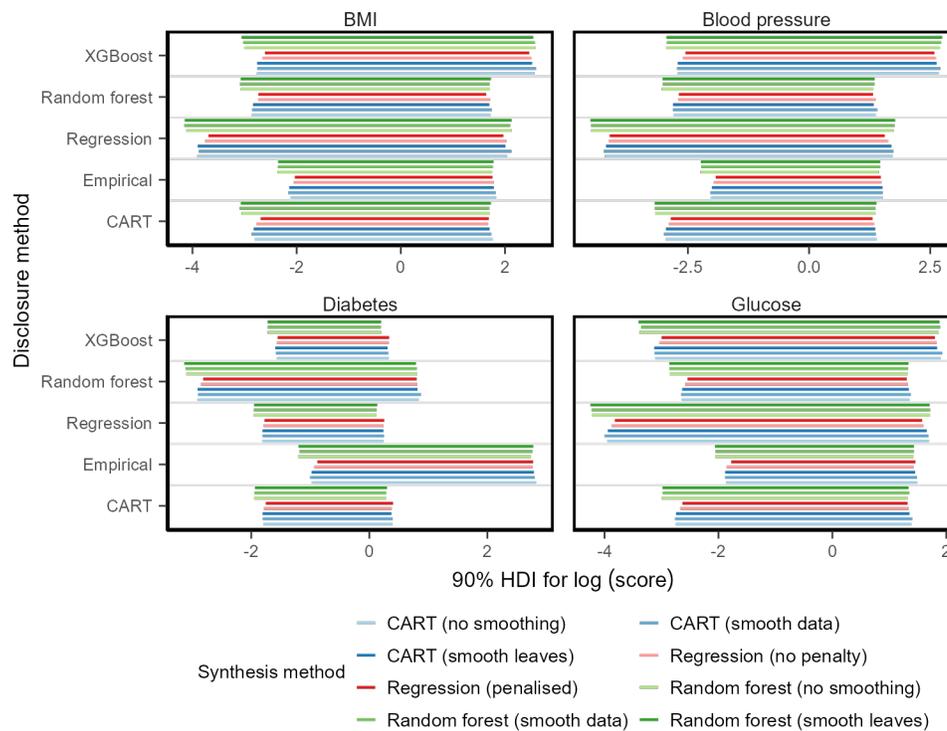


(b) Gaussian hierarchical distributional regression.

Figure 8.27: Posterior distributions of synthesis method coefficient, for Gaussian distributional regression (Model 8.14c) and Gaussian hierarchical distributional regression (Model 8.15).



(a) Gaussian distributional regression.



(b) Gaussian hierarchical distributional regression.

Figure 8.28: 90% HDPIs of attribute disclosure score from Gaussian distributional regression (Model 8.14c) and hierarchical distributional regression (Model 8.15).

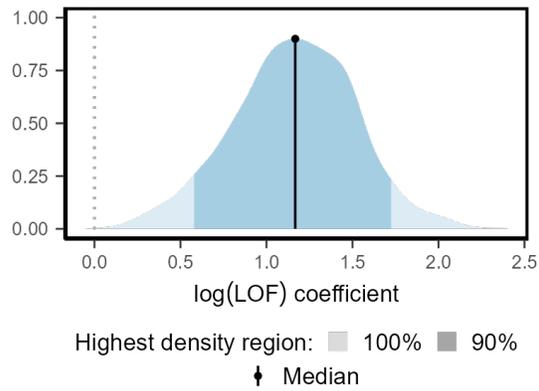


Figure 8.29: Posterior density for the coefficient of log local outlier factor (see Model 8.15).

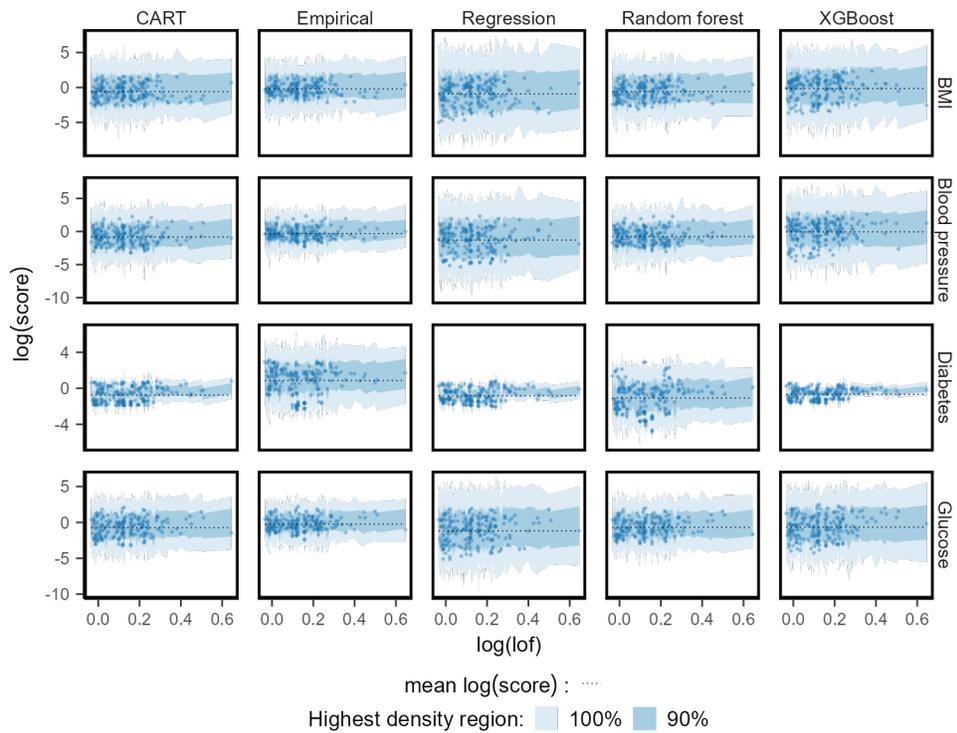


Figure 8.30: Posterior predictive distribution of log attribute disclosure score given log local outlier factor (see Model 8.15).

8.4 Conclusions

Based on the results of this chapter, of the synthesis methods that were compared, CART appears to be the best. CART synthesised datasets consistently performed well across all utility assessments. The distributions for some of the regression synthesised count variables were clearly affected by poor model specification, however, that lack of robustness against misspecification is a major disadvantage in comparison to CART. It is somewhat surprising that the random forest synthesised datasets performed worse in the utility assessments. We would generally expect an ensemble of decision tree models to outperform a single decision tree. The lack of parameter tuning for the random forest synthesiser may have been a contributing factor, but these results do mirror similar findings for incompletely synthesised data (Caiola & Reiter, 2010) and missing data imputation (Drechsler & Reiter, 2011).

Overall, most utility assessments do not identify many differences between synthetic datasets. This provides strong justification of our decision to evaluate utility using a variety of utility assessments. There are however, clear indications that k -anonymisation significantly lowered the quality of the data. When utility assessments do identify a difference between synthetic datasets, it is generally that regression synthesised data (Model 8.1) is worse than the other synthesis methods. As stated previously, many of the issues with the regression data are likely the result of poor model specification and could be resolved with better choices of synthesis model. The utility assessment which identifies the clearest differences between synthetic datasets is the inference task, which justifies our choice to implement a task based utility assessment.

For our disclosure risk assessment, we implemented a unified comparison of multiple attributes and prediction methods. Fitting models for this unified comparison required overcoming several challenges. It was difficult to identify major differences, especially in the hierarchical models. While we could identify differences in the non-hierarchical model, the results differed between attributes of different types. Whether this reflects the reality of the situation or is a result of the complications of the data is unknown and requires further experimentation. These findings are somewhat disappointing, especially the hierarchical results, as we would prefer to assess disclosure risk by modelling the risk for individuals.

Despite those difficulties there were some interesting results. We found that average disclosure risk for the training observations was mostly affected by the attribute prediction method and not by the synthetic data. This demonstrates the dangers of evaluating disclosure risk with a single set of fixed intruder assumptions, which was an issue that we discussed in the literature review in Section 6.3.2. Furthermore, we identified that the attribute disclosure risk would be lower for outliers. Our findings suggest that, for some intruder strategies, outliers are less vulnerable to attribute disclosures. This is

somewhat contrary to the widely held view. However, they support our discussion in Section 3.1.1 that an intruder using predictive modelling may find it more difficult to predict the attributes of outliers.

The membership disclosure assessment demonstrated that a simple membership inference attack can successfully predict training observations when data was synthesised from CART or random forest models, regardless of smoothing, while the risk for data that was synthesised with regression methods was significantly lower. Some CART and many random forest synthesised observations were identical or near identical to training observations. If membership disclosures are an issue, then these results are concerning. As we have previously discussed, the assumption that the intruder knows the values of all variables for a training observation only reflects a disclosure of sensitive information in cases where membership of the dataset is sensitive.

Chapter 9

Generation and assessment of large synthetic microdata

One major motivator for research into synthetic data is the potential of releasing large samples of microdata. Microdata is data that contain detailed information about synthetic individuals (Reiter, 2005a). In the previous chapter, we demonstrated methods for generating and assessing synthetic data with the Pima dataset. However, the Pima dataset is a small and relatively simple dataset compared to a census or healthcare database. In this chapter, we expand on the work of the previous chapter by demonstrating methods for generating and assessing synthetic data on the larger 130 Hospitals Diabetes dataset (also referred to as diabetes-130). We implement changes to our methods of assessing disclosure risk that affect both the larger scale of the new dataset, and the difficulties that we faced with our inferences of the disclosure risk results.

9.1 Exploratory analysis

The diabetes-130 dataset contains 101,766 inpatient encounters (also called visits) from 130 hospitals and integrated delivery networks during the years 1999–2008 that include any type of diabetes as a diagnosis code (Table A.2). The data was originally extracted for a study investigating the link between measuring HbA1c (blood sugar) and early readmission Strack et al. (2014). The dataset is available for download from UCI Machine Learning Repository (Dua & Graff, 2017).

An initial exploration of the dataset highlights several challenges for data synthesis. The first is that many of the patients in the dataset have repeated visits. Of the 101,766 patient encounters only 71,518 of these encounters are unique patients. That is, most patients make a single visit but a small number make significantly more (Figure 9.1). Repeated visits are not independent of each other. As such, we must remove the repeated visits or account for the dependence between repeated visits by a single person.

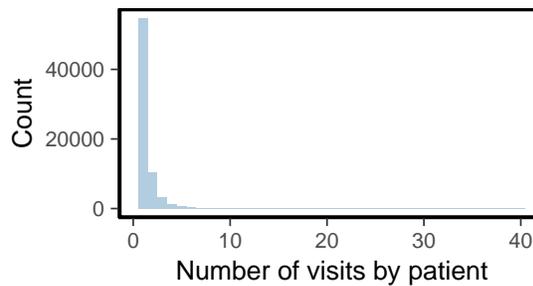


Figure 9.1: Total number of visits for each patient in 130 hospitals dataset.

The 48 variables that are observed per visit are a diverse mix of data types and distributions. The data is sparse, containing high cardinality variables such as ICD-9 diagnosis codes (Figures 9.4 & 9.5) and variables with imbalanced class distributions (Figure 9.3). There are dependencies between variables, such as the dosage of specific medications and the overall count of prescribed medications. Some dependencies span across multiple visits, such as age, weight and the number of inpatient visits in the year preceding the current visit (Figure 9.6). Generally, variables that contain patient demographic information remain constant across their multiple visits.

The count variables (Figure 9.2) are right skewed and several are zero-inflated. The number of diagnoses, number of procedures and time in hospital variables have clearly been truncated. The truncated time in hospital variable is due to the exclusion of patient visits longer than 14 days from the sample, but the truncation of the other variables is not mentioned in the original paper (Strack et al., 2014).

The distribution of the lab procedure variable is particularly interesting. The major mode of the distribution is what we would expect for a count variable with zero-inflation. However, at semiregular intervals there are spikes of random noise. These spikes are more prevalent when the number of lab procedures is under 50. Often, spikes throughout the distribution of a numeric variable can indicate rounding, however the random noise suggests that this is not the case. One possible explanation is if there are standard sets of lab procedures that doctors tend to use. This could cause certain numbers of procedures to be more common than others, especially at the lower end of the distribution. This explanation would align with random variation of the peaks, and the seemingly inverse relationship between the number of lab procedures and the magnitude of the peaks. Without a clear understanding of the underlying mechanism that causes this pattern, it will be very challenging to model.

There are also restrictions in the data, observations that would be impossible outside of errors in recording the data. These include non-negative values and structural zeros (Upton & Cook, 2014) as well as more complicated dependencies that affect multiple visits. For example, patient demographic information should not change across repeat

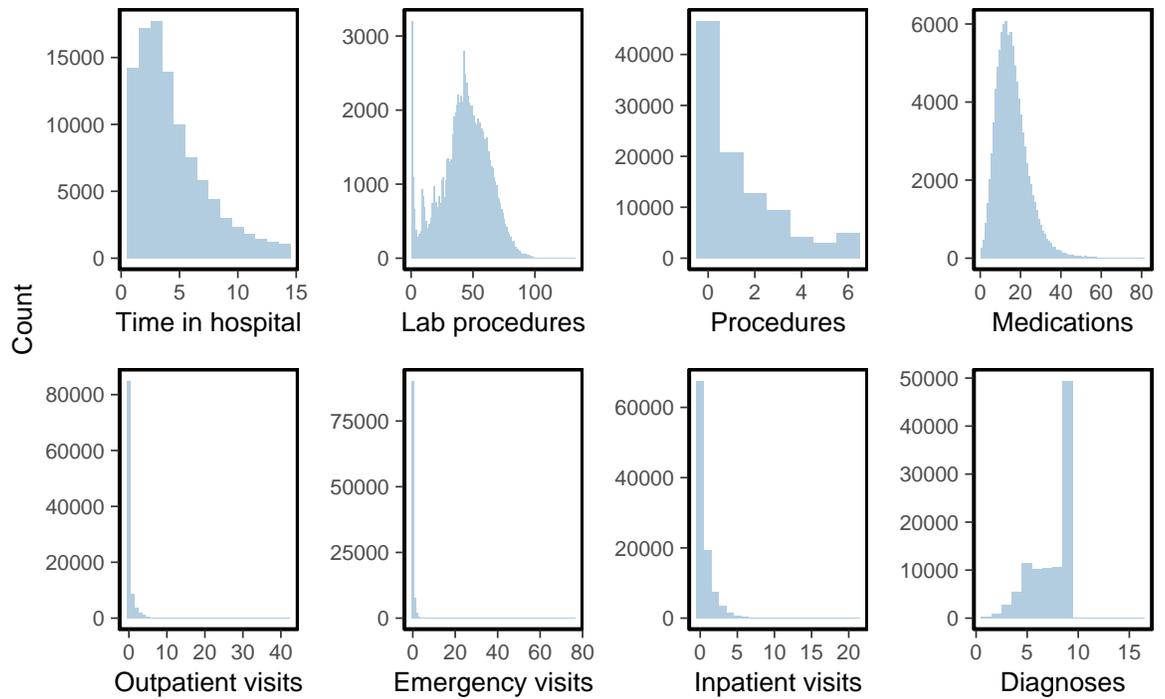


Figure 9.2: Distributions for count variables in the 130 hospitals data.

visits. With the obvious exception of age, which should never decrease. Another example is, if a patient were to die during a visit, then there should not be a later visit. Values that should be impossible are occasionally observed in the dataset, presumably this is due to data input mistakes. Due to all of these factors, the 130 Hospitals dataset is challenging to synthesise without making multiple simplifying assumptions.

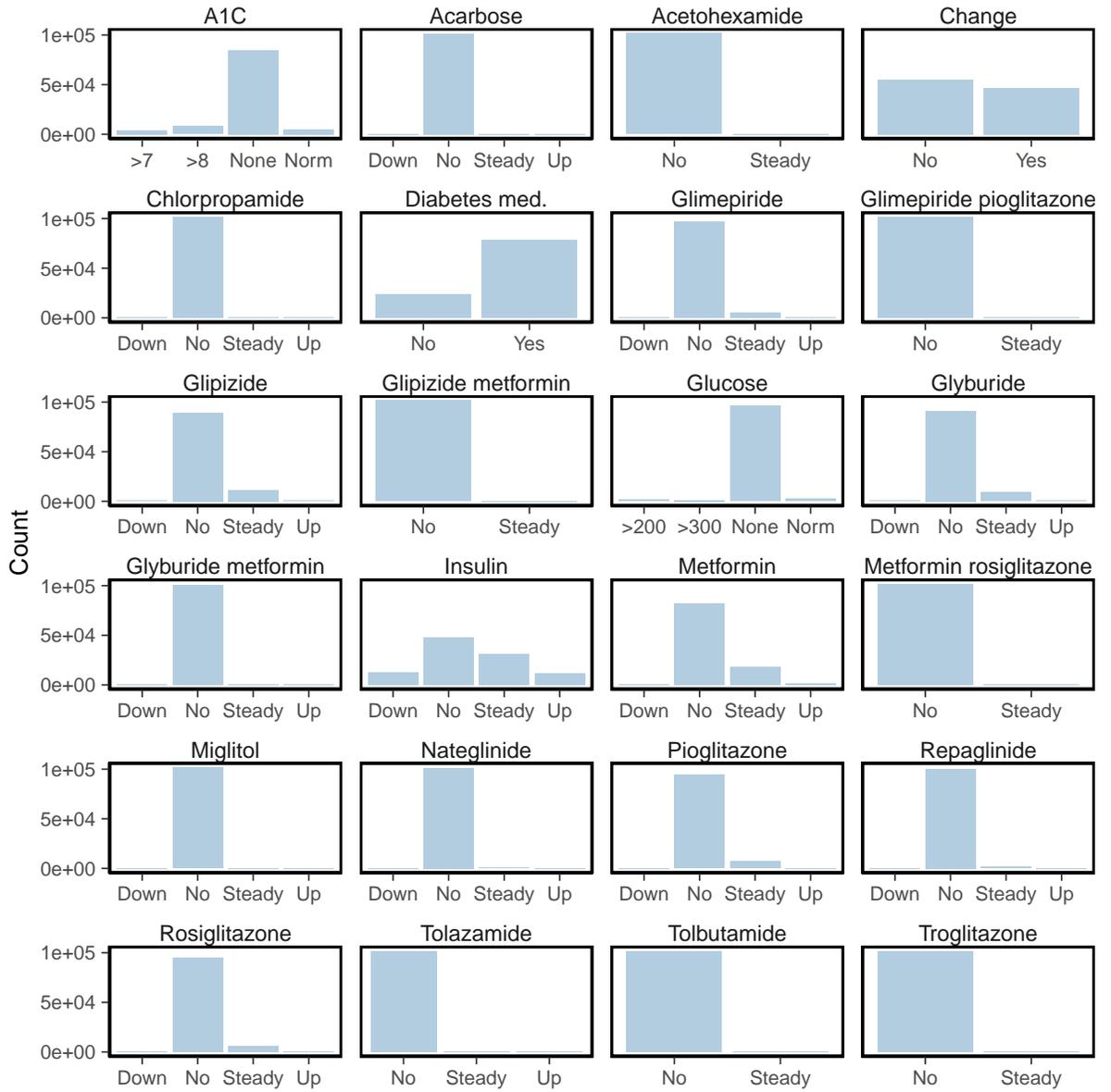
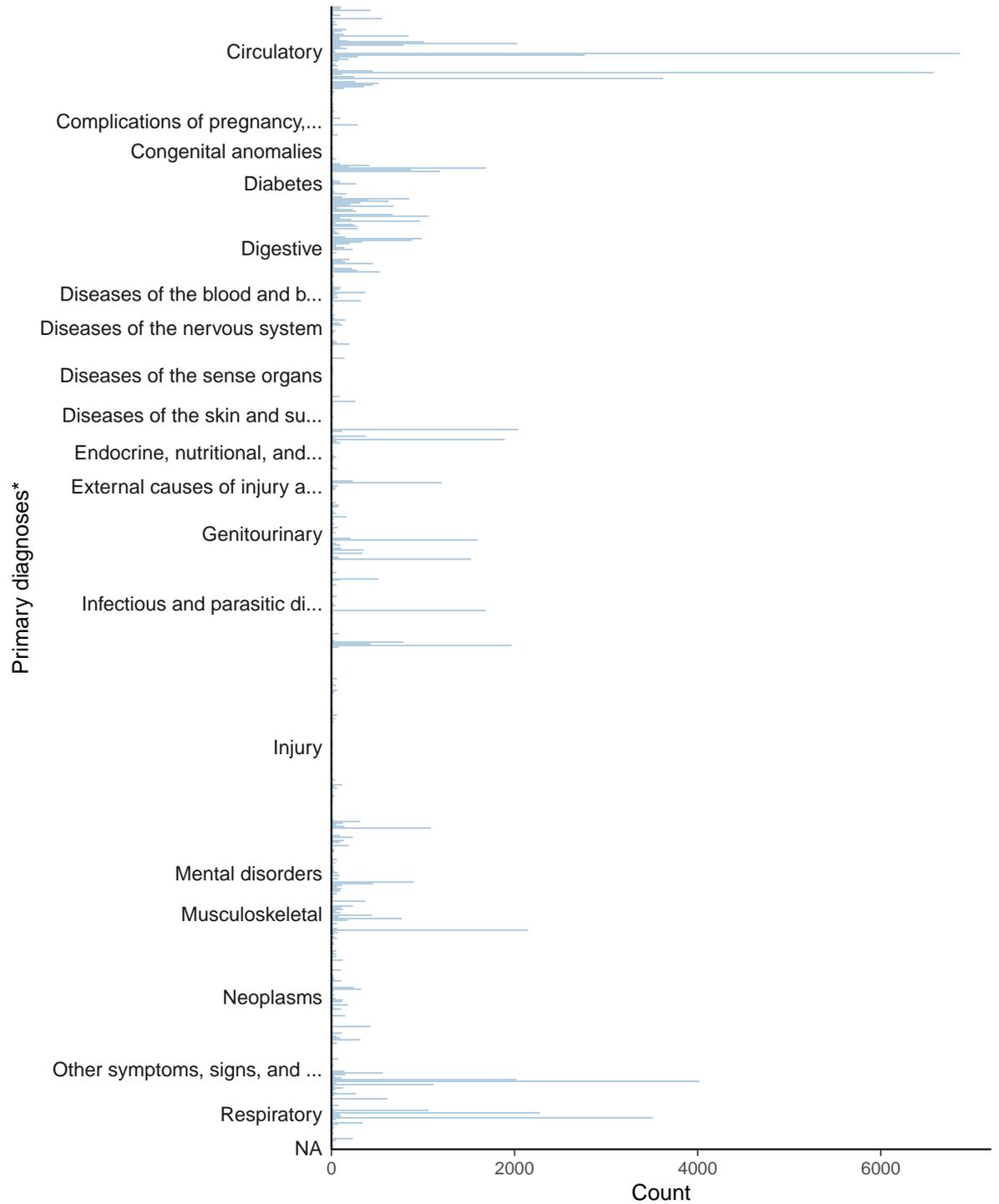


Figure 9.3: Distributions of medicine variables in the 130 hospitals data.



*One bar for each ICD-9 code, but only ICD-9 groupings are labelled to avoid visual clutter.

Figure 9.4: ICD-9 code for primary diagnosis of each visit in the 130 hospitals data.

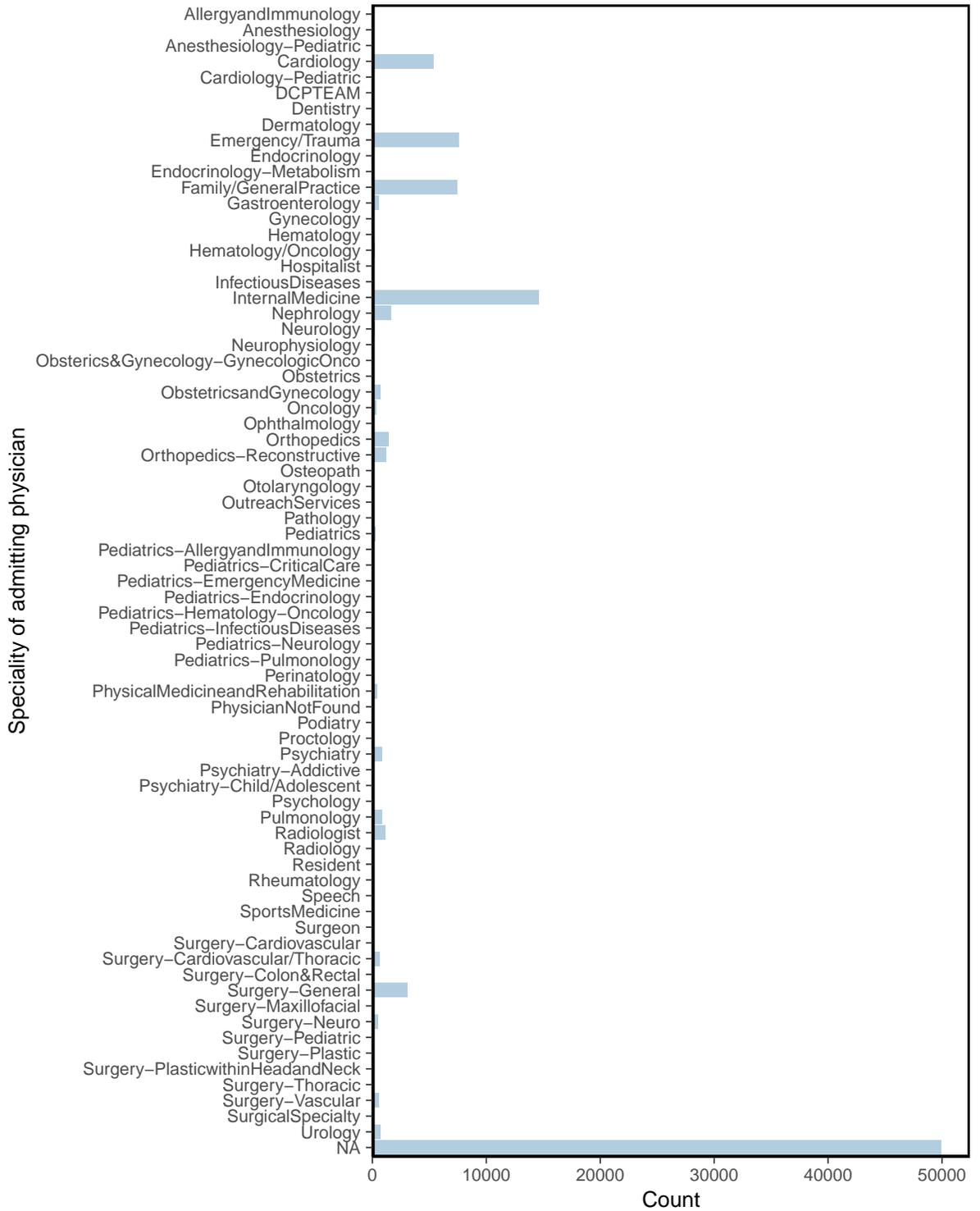


Figure 9.5: Speciality of admitting physician for each visit in the 130 hospitals data.

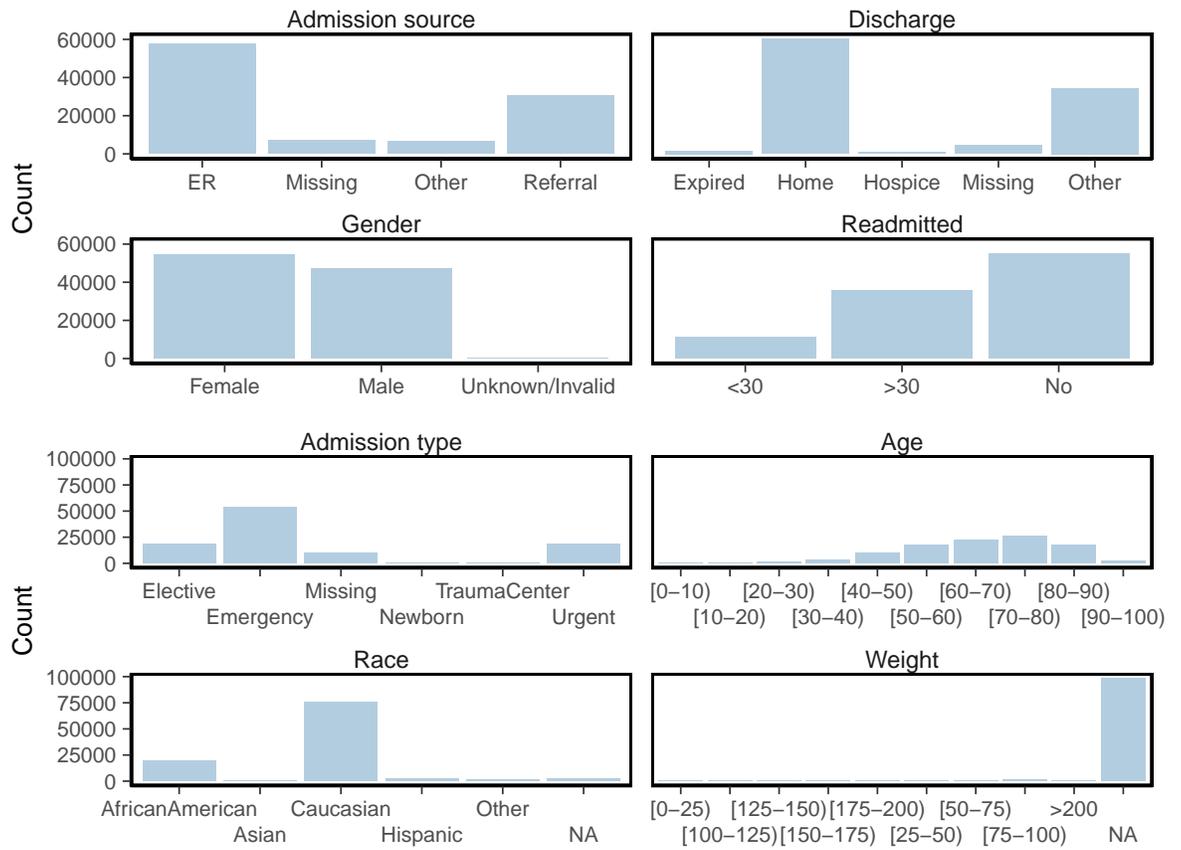


Figure 9.6: Distributions of categorical variables across all visits in the 130 hospitals data.

9.2 Methods

9.2.1 Synthesising Diabetes-130 data

In Chapter 8, we compared several variations of different methods of sequential synthesis. In this chapter, we will synthesise the 130 hospitals diabetes data with a smaller selection of models that reflect the results of the previous chapter and the additional challenges of the Diabetes-130 dataset.

Due to the large number of variables and observations, sequential synthesis of the Diabetes-130 data will be much slower than it was for the Pima data. We attempt to generate synthetic data with regression, CART and random forests, but it becomes clear that synthesising the large dataset with random forests is not feasible. The compute time is notably slower at each subsequent step and, less than halfway through the sequence of variables, progress slows to a crawl. Synthesising the high cardinality variables last helps to speed up computation, but not enough to finish the synthesis process. It is possible that a more efficient implementation of random forest sequential modelling would allow for synthesis of this dataset, however, due to time constraints we leave this for future research.

Smoothing of numeric variables had little effect on the utility or disclosure risk of the CART and random forest synthesised Pima data. There is no guarantee that this result will generalise to the diabetes dataset, which is very different to the Pima data. However, given the additional computational burden of synthesising the larger dataset, we forgo the comparison of smoothing and leave it as an avenue for future research.

In the previous chapter, we saw that the regression models performed poorly. Penalised regression methods were the worst performing on most utility assessments, but the non-penalised methods were still poor. Even simple checks of the univariate distributions showed that transforming and then synthesising count variables from linear regression models was not an effective synthesis method. The tails of the variables that were synthesised in this manner included impossible or very extreme values. Two approaches for synthesising numeric variables are explored. The first was to also synthesise these variables with CART and the second was to model these variables with count regression models, following the approach described in Section 4.4.2 (Kleinke & Reinecke, 2013).

An initial exploration of count regression models is carried out by fitting intercept only models to the data, and then using histograms to compare values that were generated from each model (Figure 9.7). Poisson, quasi-Poisson and negative binomial models are all considered. Additionally, hurdle and zero-inflated (or one-inflated) variants of Poisson and negative binomial models are also explored. Based on the results of this exploration, we decide to use quasi-Poisson models to synthesise time in hospital, medications, inpatient, outpatient and emergency visits, while number of lab procedures is synthesised from a

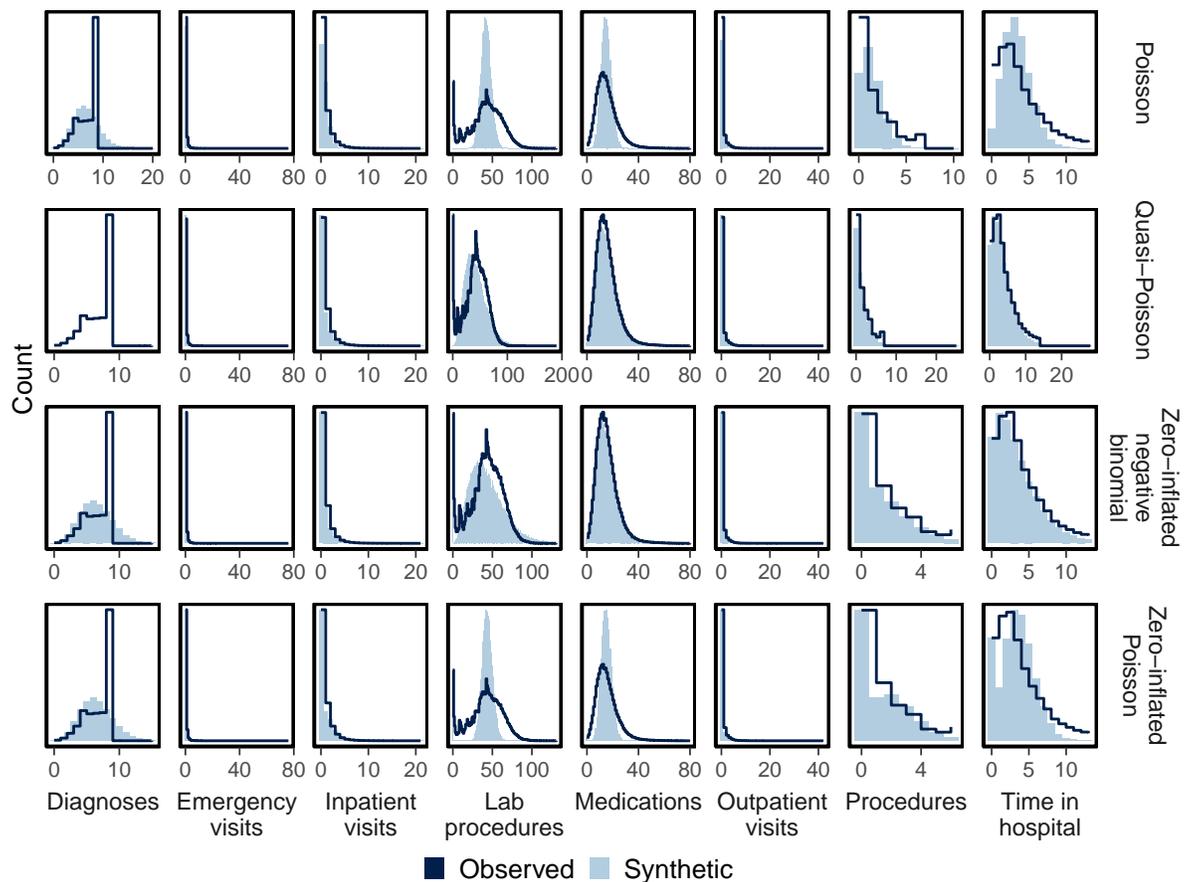


Figure 9.7: Results from the initial exploration of regression models for synthesising count variables.

zero-inflated negative binomial model. Number of diagnoses is clearly censored and none of the regression models seem appropriate, so we will use a CART synthesis model. The CART model will capture the truncated nature of the distribution, however, a regression model that is appropriate for right censored may have enabled the synthesis of number of diagnoses, as if the censoring had never been applied. We are not aware of any examples of synthesising right censored data with regression models. Although, such models have been utilised in the context of missing data imputation and their application to synthetic data is a potential future area of research (van Buuren & Groothuis-Oudshoorn, 2011, Section 3.7.3).

In total, 10 datasets were synthesised solely from CART models and 10 from the mixture of CART and generalised linear regression models. This was a significant reduction from the 100 sets of Pima data that were synthesised for each method in Chapter 8, however evaluating the results of a large number of replications would have been overly time consuming for such a large dataset. Previous results on the Pima data also showed that the variation in the assessment results between replications was relatively small in comparison to other sources of variance, so we do not expect the reduction to significantly

change the results.

Preprocessing

The pre-synthesis preprocessing loosely follows the methods of Strack et al. (2014). We make exceptions in cases where we feel that the authors of the original paper oversimplified the data. We remove all but the first visit for each patient and then reduce the dimensionality by grouping together similar levels of categories into single categories (Table A.3). The variables for the dosage of individual medicines are dominated by the “no” category, with the remaining three levels — decreased dose, steady dose and increased dose — very rarely or never observed. The extra levels would increase model complexity and their sparsity would be challenging to model. As such, we simplify all individual medicines except insulin to binary variables that indicate if the patient was prescribed that medicine. We also remove patients with a missing primary diagnosis ($n = 21$) and gender ($n = 3$). Then, we randomly divide the remaining visits ($n=71504$) into training (75%) and testing (25%) partitions, without stratification. We choose to not use stratified sampling because the data contains many categorical variables but no clear outcome variable.

Order of synthesis

As in Chapter 8, we synthesise the data using sequential synthesis, with each variable synthesised conditional on all previous variables. Note that we keep the order constant for all synthesis models. In the previous chapter, we determined synthesis order solely based on assumptions about the dependence between the variables. However, the diabetes-130 data contains variables with many missing values or very imbalanced classes. We expect this to be more difficult to model. While assumptions on dependence between variables are still an important factor for the order of synthesis, we want to synthesise very imbalanced and heavily missing variables towards the end to reduce their influence on other synthesised variables.

We first synthesise the variables that define a general profile of the patient and their health. Following that, we synthesise the variables that define the type of hospital visit:

1. race,
2. gender,
3. age,
4. primary diagnosis,
5. secondary diagnosis,
6. tertiary diagnosis,
7. number of inpatient visits,
8. number of outpatient visits,
9. number of emergency visits.

Next, we synthesise variables that relate to tests, procedures, and medications given during the visit and whether the patient was readmitted following that visit:

- | | |
|-------------------------------|-------------------------------|
| 10. admission type, | 18. number of medications, |
| 11. admission source, | 19. HbA1C result |
| 12. discharge information, | 20. max glucose serum result, |
| 13. physician speciality, | 21. diabetes medication, |
| 14. time in hospital, | 22. change of medication, |
| 15. number of procedures, | 23. insulin, |
| 16. number of lab procedures, | 24. readmitted. |
| 17. number of diagnoses, | |

Finally, we synthesise the variables with high degrees of missingness or class imbalances:

- | | |
|-------------------------|-------------------------------|
| 25. weight | 37. tolazamide, |
| 26. metformin, | 38. miglitol, |
| 27. glipizide | 39. tolbutamide, |
| 28. glyburide, | 40. glipizide metformin, |
| 29. pioglitazone, | 41. troglitazone, |
| 30. rosiglitazone, | 42. acetohexamide, |
| 31. glimepiride, | 43. metformin pioglitazone, |
| 32. repaglinide, | 44. metformin rosiglitazone, |
| 33. glyburide metformin | 45. citoglipton, |
| 34. nateglinide, | 46. examide, |
| 35. acarbose | 47. glimepiride pioglitazone. |
| 36. chlorpropamide, | |

9.2.2 Baseline datasets

We compare the utility and privacy of the synthesised datasets against the training and testing datasets and a k -anonymised training dataset ($k = 50$). For the k -anonymised training dataset, we treat age, race, gender, and weight as quasi-identifying variables, where, “missing” was treated as a category. Treating missingness as its own category is not generally recommended, as this can bias the data (van Buuren & Groothuis-Oudshoorn, 2011, Section 1.3.7). However, masking procedures such as k -anonymising data are already known to bias data (Winkler, 2007) and it is reasonable to assume that weight is quasi-identifying. The k -anonymised dataset contains 51,812 observations. This is a much higher percentage of the training observations (96.7%) than the k -anonymised dataset in the Pima example (see Section 8.2.2). Presumably, the low number of observations that were removed by k -anonymisation is due to the application of other de-identification procedures. While we are not aware of the details of de-identification procedures that were applied to the 130 hospitals dataset, at the absolute minimum, HIPAA compliant

de-identification (see Example 3.3) will have been mandatory to allow for the public release of the data.

9.2.3 Assessing the utility of synthetic Diabetes-130 data

For the initial exploration of the synthetic data, we check how well univariate and conditional distributions from the training data are preserved. We plot the proportions of categorical variables and distributions of numeric variables and compare with those of the training data.

We check conditional relationships between variables in each synthetic replication using dimension-wise prediction (see Definition 5.2). We predict each variable using the same three prediction models as the Pima example (Models 8.4a, 8.4b and 8.4c). Although differences in the size of the dataset and the types of variables necessitate some changes. Unlike the Pima example, the Diabetes data contains categorical and count variables, which we score with AUROC and Poisson log loss respectively.

We increase the number of trees of the random forest model to 500, because we expect that more trees will be required to stabilise the prediction error of the larger dataset. An investigation of how the random forest fits to the training partition shows that prediction error has stabilised at 250 trees. While it is not guaranteed that the same holds true for the synthetic data, their distributions should be simpler than the original data. Therefore, we believe it is unlikely that more trees would be required than the training data.

For the boosted trees, we implement subsampling. At each round of boosting, the tree is fit to a uniform random sample that contains 50% of the training data. This helps to prevent overfitting (Hastie et al., 2009, p. 365) and has the nice side effect of reducing the runtime of each boosting round. In the expectation that more boosting rounds would be required for the boosted trees algorithm to stabilise, the maximum number of boosting rounds and the number of rounds without improvement before early stopping activated were increased to 500 and 25 respectively. Although, in practice we found that early stopping occurred within 50 rounds of boosting, for most variables. The step size parameter for each boosted tree is optimised using 10-fold cross validation, and the remaining hyperparameters are fixed at the default values (XGBoost developers, 2023). It would be preferable to optimise the maximum depth, however, this drastically increases the run time.

We also compute pairwise correlation differences, as given in Equation (5.2), for numeric variables. In this exploration, we are particularly interested in how well the distributions of categorical variables with imbalanced classes and the tails of numeric variables are preserved, as these tend to be the more challenging variables to model.

We check for four specific examples of structural zeros for each synthetic record:

1. Diabetes medications is false but any specific diabetes medication (e.g. insulin,

metformin, etc.) is true.

2. The number of diabetes medications prescribed is less than the number of specific diabetes medications that are true.
3. The number of lab procedures is less than the number of lab results (HbA1C and max glucose serum).
4. The number of diagnoses is less than the number of non-missing diagnoses (primary, secondary and tertiary).

We confirm that all four are structural zeros by checking the training and test subsets and finding zero observations that match these definitions.

We will calculate the pMSE ratio as a general utility measurement for both datasets. We follow the same procedure that was described in Section 8.2.2.

Inference task

We assess task performance for each synthetic replication by repeating the inference carried out on the 130 hospitals dataset by Strack et al. (2014). They explore the impact of HbA1c measurement on early readmission rates for patients with diabetes by systematically fitting logistic regression models of increasing complexity.

Despite flaws in the methodology of Strack et al. (2014), which we will discuss throughout this section, we choose to replicate the majority of their methodology. As in the previous chapter (Section 8.2.2), the purpose of the comparison of inferences is to gauge whether the inference results are similar for the synthetic datasets. So, while it is not ideal if the inferences for the real data are incorrect, it is less problematic than if valid inferences of the real data were our primary goal.

We preprocess the data following the original inference as faithfully as possible. This is in contrast to the pre-synthesis preprocessing, which ignored choices that were felt to be over-simplifying. This additional preprocessing consists of

- removing patients that were discharged to hospices or expired during their visit,
- recoding the readmission variable to a binary factor for whether a patient was readmitted within 30 days,
- recoding the missing category into the ‘other’ category, for admission source and discharge disposition,
- recoding the Asian and Hispanic categories of race into ‘other’, and
- the creation of a new variable that combines HbA1c and change of medicine into a single variable (Table 9.1).

Collapsing categories removes information from the data and can bias data, as can non-imputation of missing data.

In the original paper a plot of the relationship between age and probability of readmission shows that age could be grouped into three distinct intervals for which the probability of readmission stayed relatively similar. This plot was used to justify collapsing age from ten categories to three. We replicate this plot for each synthetic dataset to see if the same conclusion can be drawn.

Table 9.1: How HbA1c and change variable was defined (Strack et al., 2014).

HbA1c measurement	Diabetes medication changed?	HbA1c change
> 8	Yes	High, changed
> 8	No	High, not changed
> 7	-	Normal
Normal	-	Normal
None	-	Not measured

We calculate various sample statistics for each synthetic dataset:

- The proportion of patients where HbA1c was measured.
- The proportion of patients where HbA1c was less than 8%.
- The proportion of patients where HbA1c was not measured and they had a medication change during their hospitalisation.
- The proportion of patients where HbA1c was measured and they had a medication change during their hospitalisation.
- The proportion of patients where HbA1c was greater than 8% and they had a medication change during their hospitalisation.

For each synthetic dataset, we calculate the 95% confidence intervals of these proportions using the normal approximation, and we estimate sample variance using the simple variance estimator (Equation (4.5a)). Then, we calculate the confidence interval overlap for the synthetic and training data (Equation (5.7)).

Rather than replicate the regression methods of the original paper, we fit a simpler, Bayesian regression model that contains only the main effects. Strack et al. (2014) implements a step-wise procedure to regression modelling, where a many permutations of covariates are considered and the most “statistically significant” models are chosen. Inference carried out in this manner is well known to lead to an increase in false positives from inference, see, e.g. (Gelman & Loken, 2013). Additionally, comparison of inference

results across multiple datasets would be challenging because the covariates for the inference models differ. We were able to faithfully replicate their step-wise procedure and the large difference between the models fit to the training and test partitions was evidence that the models chosen by their procedure were overfit to the real data. None of the interaction terms that were chosen in the inference model for the training partition were included in the model for the testing partition, and vice versa. As such, we opt for a simpler choice of inference model that does not include any interaction terms.

Model 9.1 (Diabetes-130 inference model). We model readmission within 30 days

$$y_i \sim \text{Bernoulli}(p_i),$$

with a simple linear regression that contains all the main effects

$$\text{logit}(p_i) = \alpha + \sum_{j=1}^9 \mathbf{X}_{ij}^T \boldsymbol{\beta}_j, \quad (9.1)$$

where $j = 1, \dots, 9$ correspond to the variables discharge, race, source, speciality, time in hospital, age, gender, primary diagnosis and A1C change. We assign weakly informative priors to each parameter

$$\alpha \sim \mathcal{N}(0, 2.5), \quad \boldsymbol{\beta} \stackrel{\text{iid}}{\sim} \mathcal{N}(0, 2.5) \quad \text{and} \quad \sigma \sim \text{Exponential}(5). \quad (9.2)$$

When fitting this model, we treat age as a continuous variable. This deviates from the original paper, but we feel that this is the more correct treatment of the age variable because treating age as categorical does not reflect its natural ordering.

Recall, that the posterior distribution of a parameter can be estimated by combining the posterior distributions for multiple synthetic replications, Section 4.2.2. We run four chains, each containing 2000 samples, for each baseline dataset and synthetic replication. For each synthesis method, we combine the chains for all replications and calculate 90% credible intervals. Then we compute the percent overlap with the 90% credible interval of the training partition. The percent overlap of each coefficient is compared for all synthesis method and baseline datasets.

We briefly explore the idea of fitting an additional inference model with regularised horseshoe priors (Equation (2.10)) on the coefficients, to more closely mirror the variable selection aspect of the stepwise procedure in the original research. We are able to train these horseshoe prior models on the baseline partitions, however, upon trying, one finds it extremely challenging to train those models on the synthetic replications. Despite the relatively simple model design, the posterior geometry that is induced by the horseshoe prior is very complex and so the chains sample very slowly and contain many divergent

transitions. We attempt to address this by setting the acceptance probability to be extremely small. This change reduces the number of divergent transitions but paired plots of the posterior draws indicate that the model is not exploring the entire posterior. It is possible that keeping the low acceptance probability and increasing the number of samples would have helped. However, there was no guarantee of this working and it would have been time consuming, so ultimately we decided to settle for the inference model with weakly informative Gaussian priors (Model 9.1).

9.2.4 Assessing the disclosure risk of synthetic Diabetes-130 data

We assess attribute disclosure risk following a similar approach to that applied in Chapter 8. However, we make adjustments to reflect the challenges faced during that assessment and also the new challenges introduced by the larger scale of the 130 hospitals dataset. We evaluate attribute disclosure risk by comparing the errors of attribute predictions made on the training partition of the data from models trained on the synthetic replications of data or testing and k -anonymous baseline datasets.

In the previous chapter, we considered the attribute disclosure risk of four attributes, which were predicted by an intruder that knew two quasi-identifying variables. While we were able to draw inferences about the disclosure risks for the four attributes that we considered, it was not clear if the results could be extrapolated to other attributes in the data. Since the 130 hospitals data contains a much larger number of attributes, we can address this issue. We assume that the intruder will attempt to predict the values of the attributes primary diagnosis, secondary diagnosis, tertiary diagnosis, medical speciality, admission type, maximum glucose serum, HbA1C measurement, whether the patient was prescribed any diabetes medication, and the individual prescription information for several diabetes medications. Note, we do not expect that an attacker will be interested in predicting that an attribute is missing. Therefore, if a subject has a missing value for an attribute, we will exclude it from the disclosure risk assessment of that attribute.

Recall that, in the previous chapter, a large source of the difficulty of modelling the attribute disclosure scores was due to differences in the prediction error distributions of each attribute. In particular, the difficulty was due to the difference between the numerical and the categorical attributes. While attempting to unify the results for different types of attributes is an interesting problem, the larger scale of the 130 hospitals data is computationally challenging in itself. In addition, we do not believe that the numeric variables in the data reflect the types of variables that an intruder would be interested in. In fact, several of the numeric variables are summaries derived from other categorical variables in the data. Due to all of these factors, we only consider the attribute disclosure risk of categorical variables.

Recall that the method of attribute prediction was a strong predictor of the disclosure

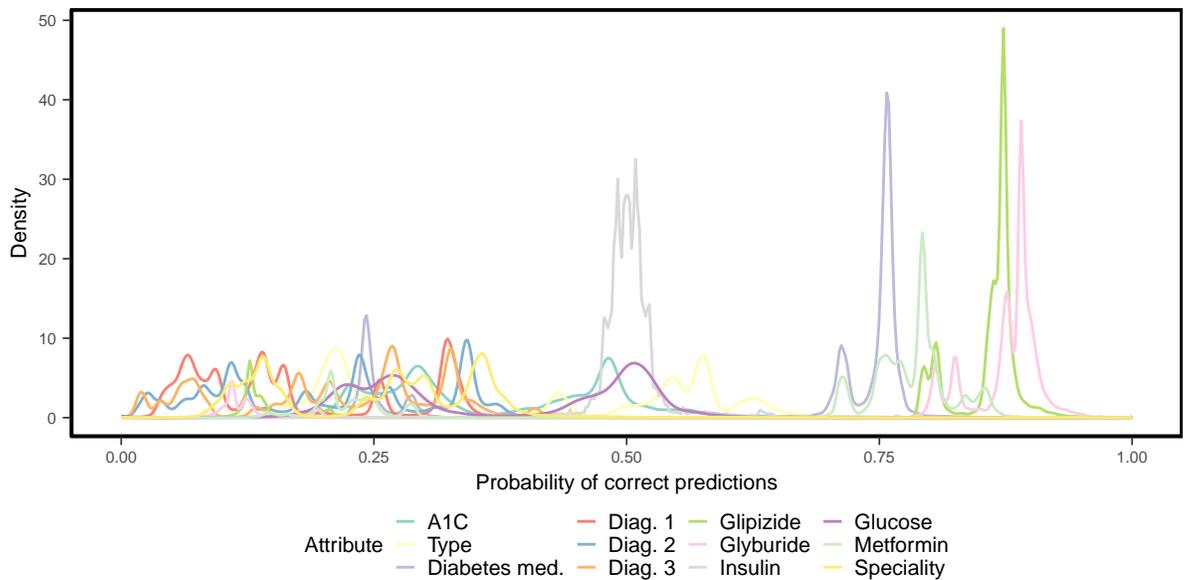


Figure 9.8: Proportion of correct attribute predictions for different attributes in the 130 hospitals training partition.

risk scores for the Pima data, Section 8.3.2. In fact, the method of attribute prediction was a stronger predictor than the method of synthesis. Consequently, we will continue to assess attribute disclosure risk for a range of prediction methods. We will assume that the intruder utilises the following attribute prediction models: CART, random forest, XGBoost, and the empirical matching procedure (see, Definition 7.1).

The dataset of results consisted of 50,084,848 observations

$$p_{ijrst} = P(f_j(t_i | \mathbf{Q}_{sr}, \mathbf{t}_{sr}) = t_i),$$

where p_{ijrst} is the probability that the j^{th} attribute prediction model will predict the correct class for t_i , the t^{th} attribute of the i^{th} training observation, when trained on \mathbf{Q}_{sr} and \mathbf{t}_{sr} , the quasi-identifier and target variable columns from the r^{th} replication of a synthetic dataset that was generated with the s^{th} synthesis method.

An initial exploration of the data highlights large variations in the probabilities of correct predictions for different attributes and training observations (Figures 9.8 and 9.9). Paired differences indicate that the probability of correct predictions are smaller for XGBoost than for the other attribute prediction methods. However, they show little evidence of differences between synthetic and baseline datasets (Figure 9.10).

Initially, we thought to consider modelling p_{ijrst} as beta distributed, since this seems like a natural approach for modelling probabilities. However, some observed values are exactly zero or one, which is outside the support of the beta distribution. We could model zeros and ones with a zero-one inflated beta distribution. However, this would model those observations with a separate process. Recall from earlier (Section 8.2.3), that we consider

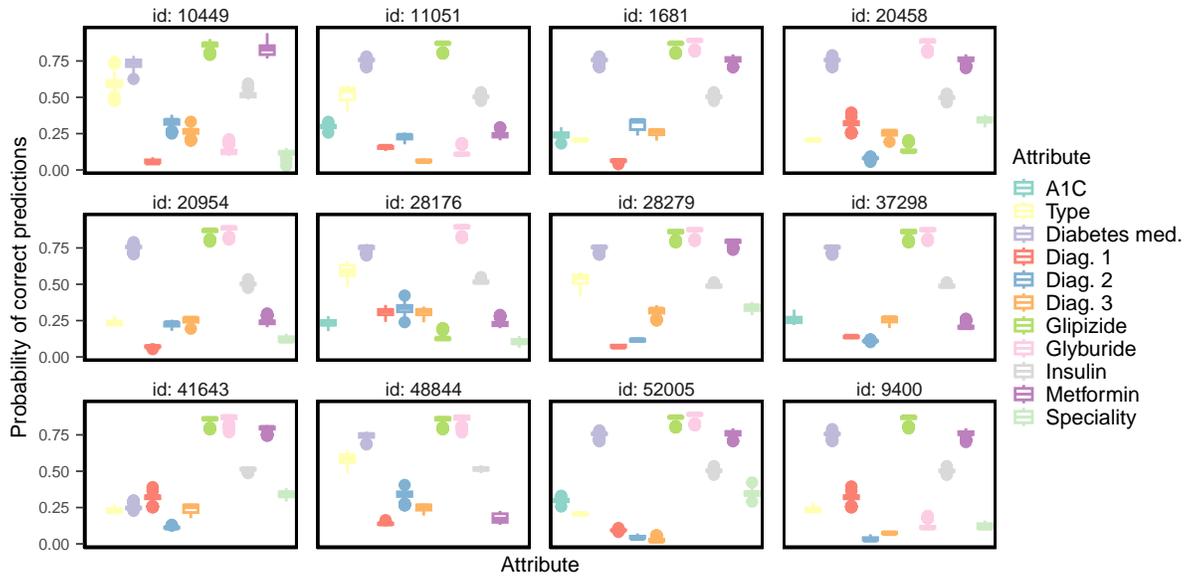


Figure 9.9: Probability of correct attribute predictions for small sample of subjects in the 130 hospitals training partition.

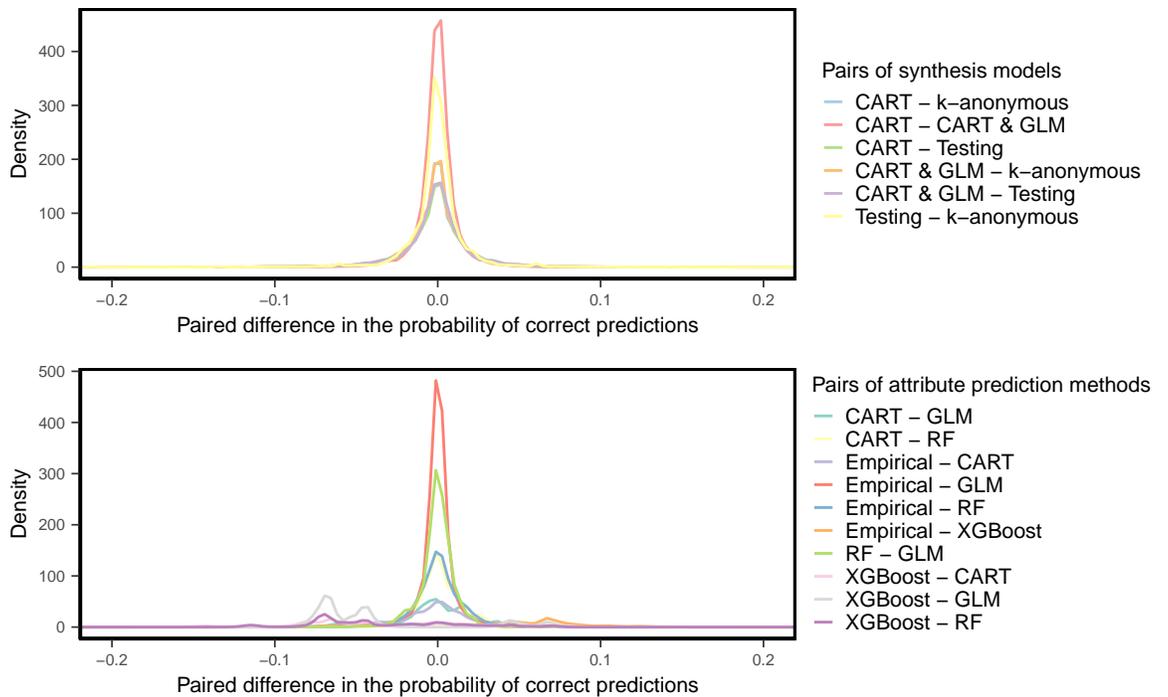


Figure 9.10: Paired differences in the probability of correct attribute predictions for all pairs of attribute prediction methods, and all pairs of synthesis methods and baseline datasets.

certain predictions to be due to a lack of precision, rather than a separate process. Given that we have a strong belief that zeros and ones are the result of the same process as the other observations, a zero-one inflated beta model is inappropriate.

Instead, we reformulate our results as Bernoulli trials for whether an attribute prediction is correct. For each observation, we draw a binary random variable with probability of success p_{ijrst} ,

$$y_{ijrst} \sim \text{Bernoulli}(p_{ijrst}).$$

Notice, for each $ijst$, we can view the n_r replications as a series of Bernoulli trials. Therefore, we further simplify by modelling the total number of correct predictions as a Binomial variable.

$$y_{ijst} \sim \text{Binomial}(p_{ijst}, n_r), \quad (9.3)$$

where n_r is 10 for synthetic datasets and 1 for baseline datasets.

Modelling the attribute disclosure scores

Our data has a repeated measurements design where, for each training observation i , predictions are made for all combinations of attribute, attribute prediction model, and synthesis model. Even simple regression models, without any hierarchical structure, can take an entire day to run in Stan. Fitting models with hierarchical structure could take weeks. Reformulating our results as Bernoulli reduces our data to 9,106,336 observations. This is a huge improvement from our initial results. Unfortunately, the data is still large enough that model fitting will be time consuming. Additionally, the amount of data is large enough that we expect the likelihood will dominate the prior, hence, removing much of the benefit of a Bayesian model. Given these factors, we fit our models using maximum likelihood methods instead of MCMC.

Our generalised linear mixed model accounts for the repeat measurements by including a random intercept for each training subject. We could reasonably treat attribute, attribute disclosure model, and synthesis model as either fixed or random effects, depending on whether we are interested in unobserved levels of these covariates. Initially attribute was also modelled as a fixed effect, since our experiments already included most attributes in our data, while synthesis method and attribute disclosure methods were treated as random effects because there were many potential methods of both synthesis and attribute prediction that were not tested. However, we found the models would not converge when either synthesis method or attribute prediction model were treated as random effects. Presumably, this is because the differences between levels of those variables are so small that the random effect variance estimates as near zero. Also, models do not converge unless attribute is treated as a random effect. As such, we fit a generalised linear model of the

form

$$\begin{aligned} y_{ijst} &\sim \text{Binomial}(p_{ijst}, n_s), \\ \mathbb{E}[\text{logit}(p_{ijst})] &= \alpha + a_i + b_j + \beta_{1s} + \beta_{2t}, \end{aligned} \tag{9.4}$$

where

$$\begin{aligned} a_i &\sim \mathcal{N}(0, \sigma_{\text{subj}}^2), \\ b_j &\sim \mathcal{N}(0, \sigma_{\text{attr}}^2), \end{aligned}$$

and

$$\beta_{11} = \beta_{21} = 0.$$

The large number of unique subjects (53,628) was another source of convergence issues that arose when attempting to fit the model to the entire dataset. We attempt to address this by randomly sampling 30,000 subjects for model training. Also, we can later use the 23,628 observations that we removed as a validation set. However, despite the reduction in the amount of training data, there are still convergence warnings during model fitting. In fact, these warnings persist unless the model is fit with 10,000 or fewer training subjects. The documentation for the `lme4` package, which we are using to fit mixed models, recommends fitting a model with multiple optimisers and checking if they each converge to the same model (Bates et al., 2015). We fit the model to 30,000 subjects, using eight different optimisers. In all cases, the optimisers converge to models that have identical log-likelihoods and very similar fitted values. While we cannot be certain, it is unlikely that this would happen if the models were not converging. Therefore, we assume that the warnings are false positives.

We validate the trained model by repeatedly drawing from the fitted values of the model in Equation (9.4): $\hat{\alpha}$, \hat{b}_j , $\hat{\beta}_{1s}$, $\hat{\beta}_{2t}$, and $\hat{\sigma}_{\text{subj}}$. We use these fitted values to predict counts for the validation data, which we compare with their observed values. Then, we assess the effects of attribute disclosure and data synthesis models on probability of a correct attribute prediction. We use the 95% confidence intervals for the model coefficients $\hat{\beta}_{1s}$ and $\hat{\beta}_{2t}$, to assess differences in the effects of each attribute disclosure model and data synthesis model. Also, we calculate 95% prediction interval for the probability of a correct attribute prediction. However, we are not aware of any method to include the uncertainty of random effects in prediction intervals. As such, the 95% prediction intervals are only conditioned on the uncertainty of the fixed effects.

9.3 Results

9.3.1 Results of utility assessments of synthetic Diabetes-130 data

The univariate distributions of all categorical variables are reasonably close to the training data (Figures 9.11, 9.12, and 9.13). In addition, dimension-wise prediction shows that the conditional distributions of most categorical variables are also quite good (Figure 9.14). However, there are some categorical variables for which the synthetic data losses are extremely poor, such as acarbose and chlorpropamide. Both are binary variables with severe class imbalances where the rarer classes are observed for less than 0.3% of the training and test datasets. Predicting variables with severe class imbalances is a challenging problem in and of itself. However, the training data loss demonstrates that some variables in the training data are strong predictors for acarbose and chlorpropamide. Whereas, the synthetic data loss indicates that those strong relationships have not been captured by either data synthesis model.

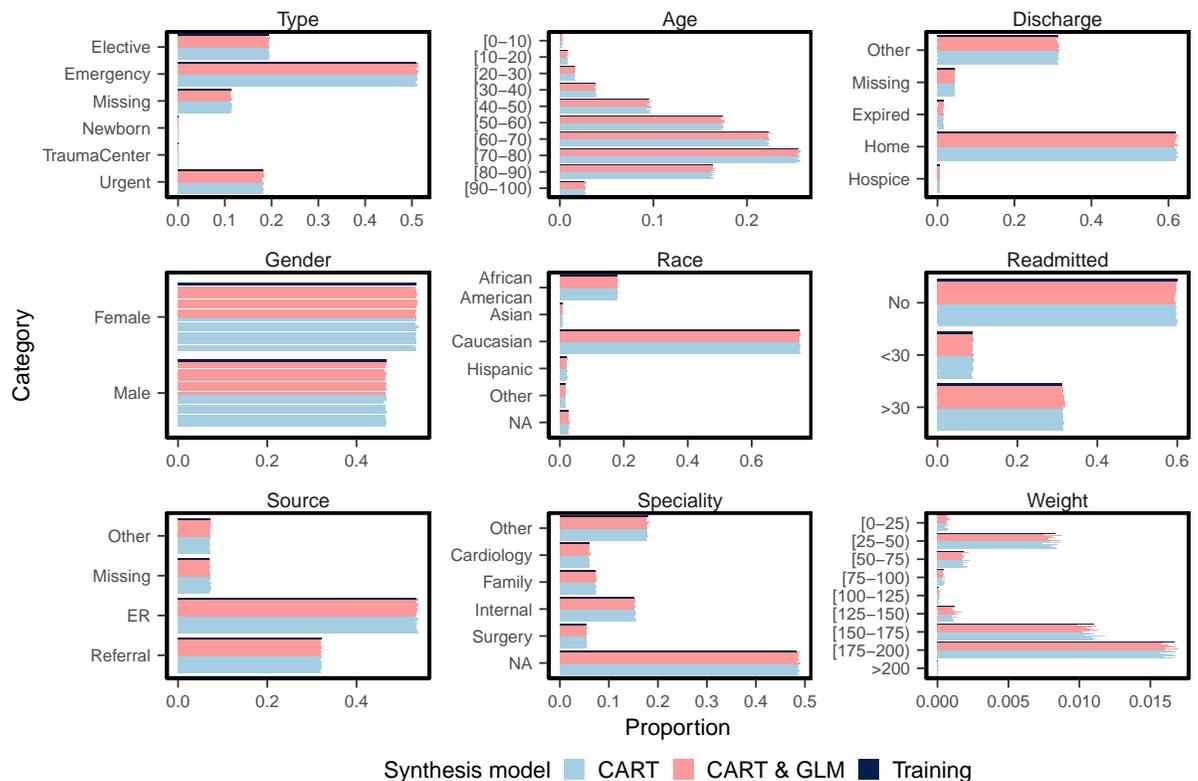


Figure 9.11: Comparing the proportions of levels of categorical variables in synthetic and baseline datasets.

The univariate distributions of most numeric variables appear to be quite reasonable (Figure 9.15). Notice that the lower tails do not contain any impossible values, which was a major problem when synthesising the Pima data with regression (Section 8.3.1). In addition, the upper tails of the regression synthesised variables are notably longer than

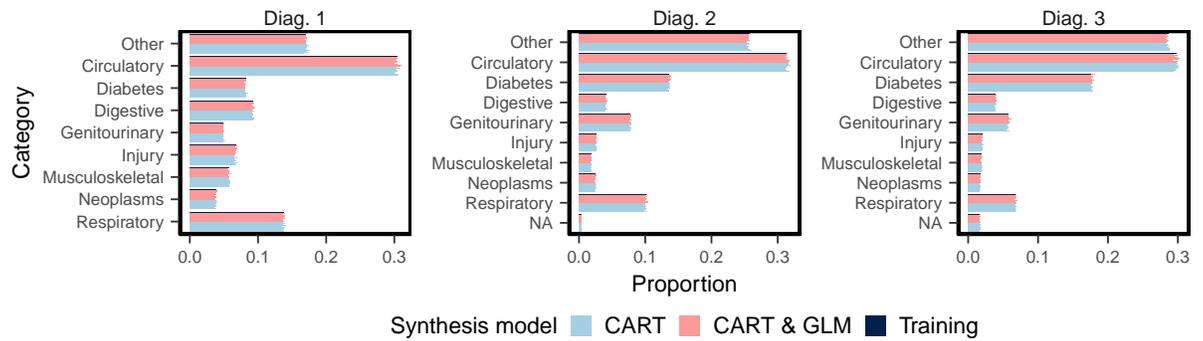


Figure 9.12: Comparing the proportions of levels of diagnosis code variables in synthetic and baseline datasets.

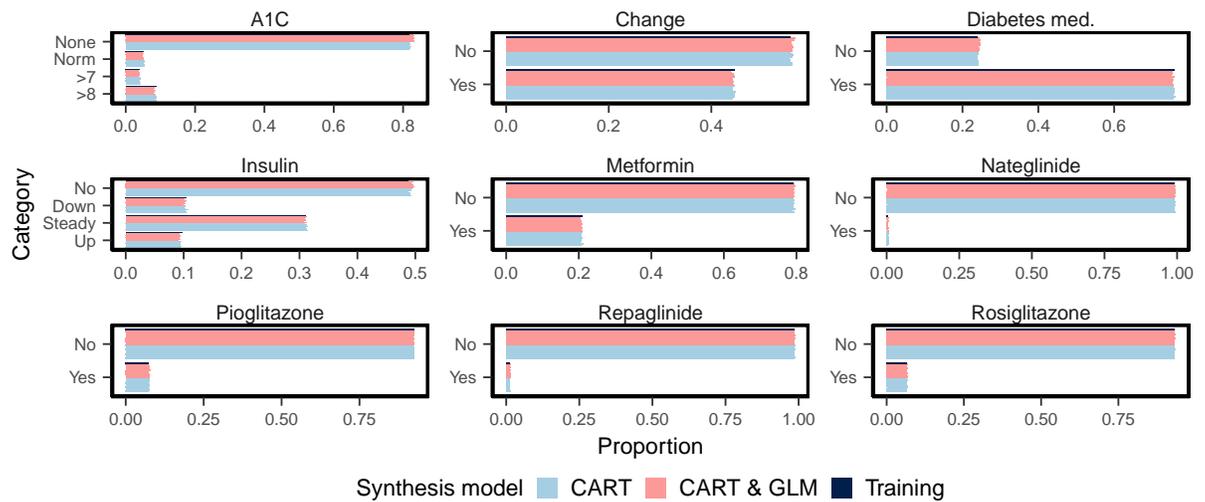


Figure 9.13: Comparing the proportions of levels of medicine variables in synthetic and baseline datasets.

those synthesised with CART, which is expected given that the output of the regression model is unbounded, whereas the CART model output is not. The regression data contains three numerical variables with impossible or very unlikely values. These are, 200 medications prescribed, 100 emergency visits to the hospital within one year, and hospital stays that are over 14 days. We lack the subject specific knowledge to say if the first two examples are impossible, or just very unlikely. However, visit lengths of over 14 days were specifically excluded from the 130 hospitals data, so those values should not be observed in the data. These longer visit lengths are due to a failure to account for the truncation of the variable when defining the synthesis model. Consequently, impossible values were generated and the visit length is more right skewed than the training data.

The zero-inflated negative binomial distribution has modelled the upper tail of the number of lab procedures reasonably well but it does not fit the smaller non-zero values in the data or the multiples of 10, which are inflated in comparison to what one would expect for the commonly used distributions of count variables. It is possible that a better

understanding of the mechanisms that leads to lab procedures being arranged or input into the systems would allow for the specification of a regression model that better fit the data. However, without that information, the regression is a poor choice of synthesis model, and the flexible nature of non-parametric methods such as CART is more suited to variables with such challenging distributions.

Dimension-wise prediction scores for all numeric variables look very good and do not indicate any issues with the conditional distributions. However, pairwise correlation differences for the synthetic datasets are larger than either of the baseline datasets (Figures 9.16, 9.14 & 9.17). There is weak evidence that the differences are smaller for the dataset in which those variables were solely synthesised with CART, but not enough of a difference to say with any certainty.

Replications of both types of synthetic datasets contained examples of all four types of structural zeros. This is not particularly surprising, given that we did not consider the strict relationships that govern those variables when building the data synthesis models.

The pMSE ratios for both synthetic datasets are greater than one for both discriminator models and larger than one for the replications where the count variables were generated with regression (Figure 9.19). This indicates that both the logistic regression and CART discriminators are able to differentiate between synthetic and training samples of the 130 hospitals data. In addition, it indicates that they are able to more reliably differentiate the difference for the dataset with regression synthesised count variables.

Comparison of results for inference task on diabetes-130 data

Table 9.2: 95% confidence interval overlaps for various sample statistics in 130 hospitals data.

	CART	CART & GLM	Testing	k -anon.
HbA1c measured	0.544	0.000	0.789	0.942
HbA1c measured & meds changed	0.229	0.000	0.714	0.939
HbA1c not measured & meds changed	0.977	0.584	0.350	0.736
HbA1c \leq 8%	0.735	0.946	0.790	0.730
HbA1c $>$ 8% & meds changed	0.277	0.459	0.788	0.809
Readmitted & HbA1c measured	0.606	0.977	0.791	0.949

Point estimates for all sample statistics are reasonably close to the training data for both the synthetic and baseline datasets (Figure 9.20). However, the 95% confidence interval overlaps for these statistics tell a mixed story (Table 9.2). In contrast to the k -anonymous data, both synthetic datasets and the test data have much less overlap with the real data. However, given that the training and test data are independent and identically distributed samples of the real population (see Section 8.2.2), the overlap of the

Table 9.3: Ratio of estimates (divided by training estimate) for various sample statistics in 130 hospitals data.

	CART	CART & GLM	Testing	k -anon.
HbA1c measured	0.983	0.998	0.921	1.003
HbA1c measured & meds changed	0.960	1.003	0.906	0.977
HbA1c not measured & meds changed	0.999	1.006	1.010	1.022
HbA1c $\leq 8\%$	1.011	1.011	1.002	1.010
HbA1c $> 8\%$ & meds changed	0.968	1.008	0.976	0.997
HbA1c measured & readmitted	1.056	1.007	0.999	0.987

k -anonymous data appears to be an unrealistically high. If we instead consider the testing data percentage overlap to be a more reasonable target, then the percentage overlap of both synthetic datasets is reasonable. Comparing just the synthetic datasets, there is an equal amount of statistics that each has the greater overlap. However, the average confidence interval overlap for the entirely CART synthesised dataset is greater than the dataset that was synthesised with both CART and regression.

Recall, that confidence interval overlap are a poor utility measure if the variance is small in comparison to the point estimate (see Section 5.4.1). Instead, we compare the statistics for the synthetic and training data using ratio of point estimates. These ratios are all close to 1, showing that the statistics are well preserved across all synthetic and baseline datasets (Table 9.3)

Plotting age against probability of readmission shows a stable linear trend between age and re-admissions for the synthetic datasets (Figure 9.21). In their analysis, Strack et al. (2014) identified three distinct intervals for age. Although, it is not clear to us that these intervals exist for either the training or testing partitions. There is weak evidence of a steeper increase in the probability of readmission from 0 to 30. However, it would be perfectly reasonable to conclude that there is a consistent linear trend of the probability of readmission increasing with age. That said, the probability of readmission of younger patients does appear to be higher in the synthetic data than either the training or test data.

Moving onto the comparisons of inference models, we plot the 90% credible intervals for the posteriors of the Bayesian inference model (9.1) that was fit to the synthetic and baseline datasets (Figure 9.22). In general, the posterior interval overlaps for the synthetic models are smaller than either the testing partition or k -anonymous models (Figure 9.23). The significant positive relationships for age, discharge, and visit length are captured in the synthesis model, although the strength of the effect of the latter two effects was underestimated for both synthetic datasets. We were not able to run models with shrinkage priors on the synthetic replications but we were able to fit them to the training and testing partitions, and the majority of coefficients were shrunk to zero (Figure 9.24). The only

coefficients for the testing partition that were not shrunk towards zero were age, discharge, and visit length. Our synthetic data somewhat captures all of the coefficients with the strongest relationships to readmission and struggles with those weaker relationships.

9.3.2 Results of disclosure risk assessment of synthetic Diabetes-130 data

Consider the plots that compare model predictions and observed values for the validation data (Figures 9.25 and 9.26). These show that the model fits the validation data well, except for a very small amount of under-dispersion in the tails of some attributes.

From the coefficients, there is no evidence of a significant difference in the probability of correct attribute predictions the synthesis models (Table 9.4). Let us consider the difference between the CART synthesis model and the k -anonymous or test baseline datasets. In this case, there is evidence of a small but significant increase in the probability of a correct prediction if the intruder has access to either of the two baseline datasets. The coefficients for attribute prediction method show that XGBoost is by far the worst prediction method. Furthermore, random forest, generalised linear models, and the empirical method are all slightly better attribute prediction methods than CART.

We would like to know if the attribute prediction method results apply to all attributes, or if there are some variables that CART or XGBoost are better at predicting. However, attempts to fit a model with a random slope for each combination of attribute and attribute prediction method were unsuccessful. This occurred due to the relatively small variance of random slope effect in comparison to the variance of the other random effects. As such, we are unfortunately unable to answer questions about the relationships between attribute and attribute prediction method.

The 95% prediction intervals show that there is little difference between the probability of a correct prediction for any method of attribute prediction and method any of the synthetic or baseline datasets (Figure 9.27).

Table 9.4: Estimates of fixed and random effects from fitted model on the log-odds scale (Equation (9.4)).

Term	Estimate	Std. error	95% conf. interval
Intercept	0.811	2.272E-01	(0.468, 1.404)
Synthesis model = CART	reference		
Synthesis model = k -anon	1.011	2.435E-03	(1.006, 1.015)
Synthesis model = CART & GLM	1.001	1.028E-03	(0.999, 1.003)
Synthesis model = test	1.006	2.425E-03	(1.002, 1.011)
Prediction method = CART	reference		
Prediction method = empirical	1.020	1.483E-03	(1.017, 1.023)
Prediction method = RF	1.007	1.464E-03	(1.004, 1.010)
Prediction method = GLM	1.017	2.031E-03	(1.013, 1.021)
Prediction method = XGBoost	0.843	1.228E-03	(0.840, 0.845)
SD(ID)	0.284		
SD(attribute)	0.970		

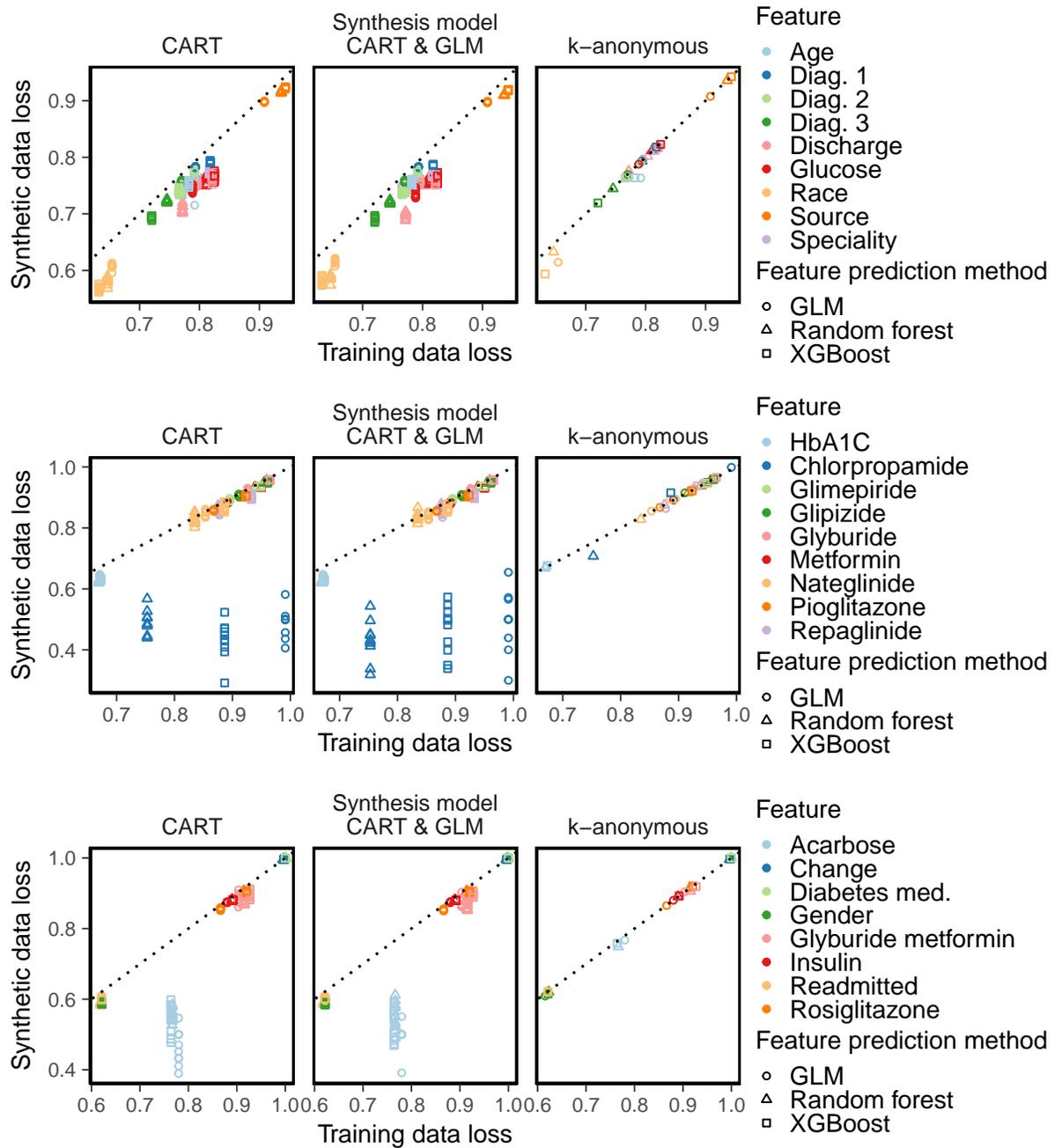


Figure 9.14: Feature prediction scores for categorical variables from the 130 hospitals dataset.

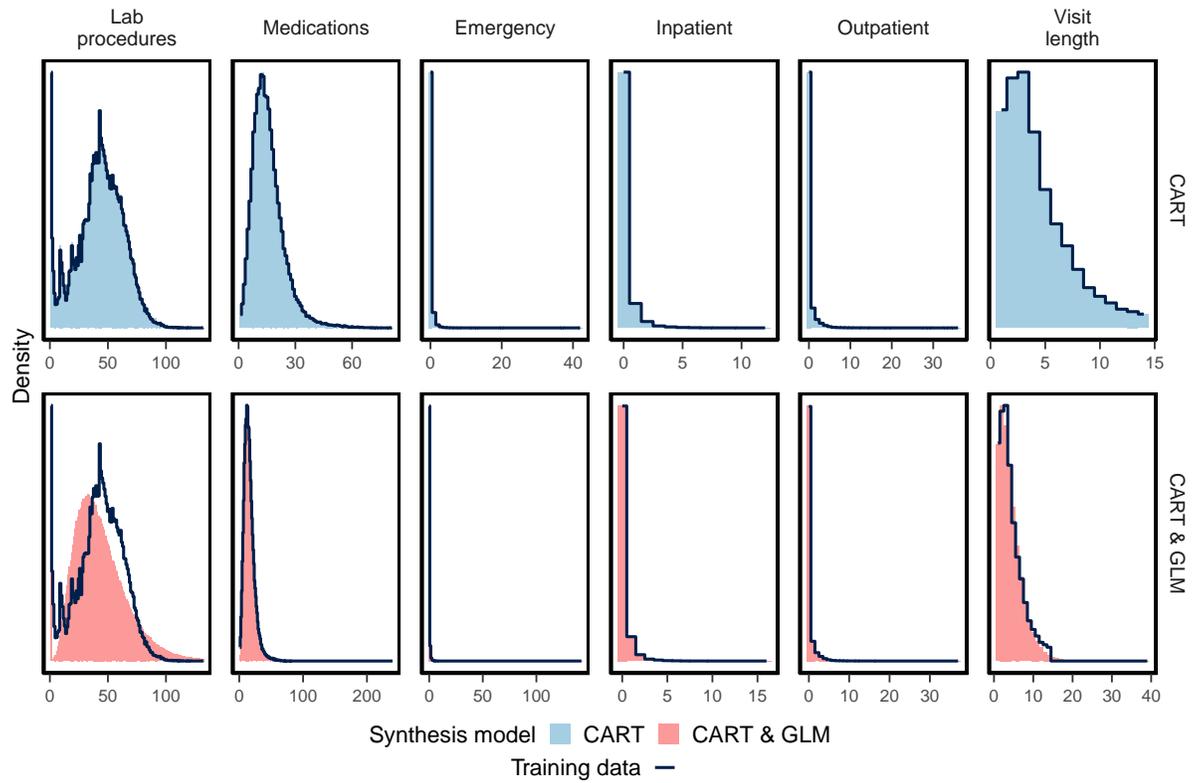


Figure 9.15: Comparing the distributions of numeric variables in synthetic and baseline datasets.

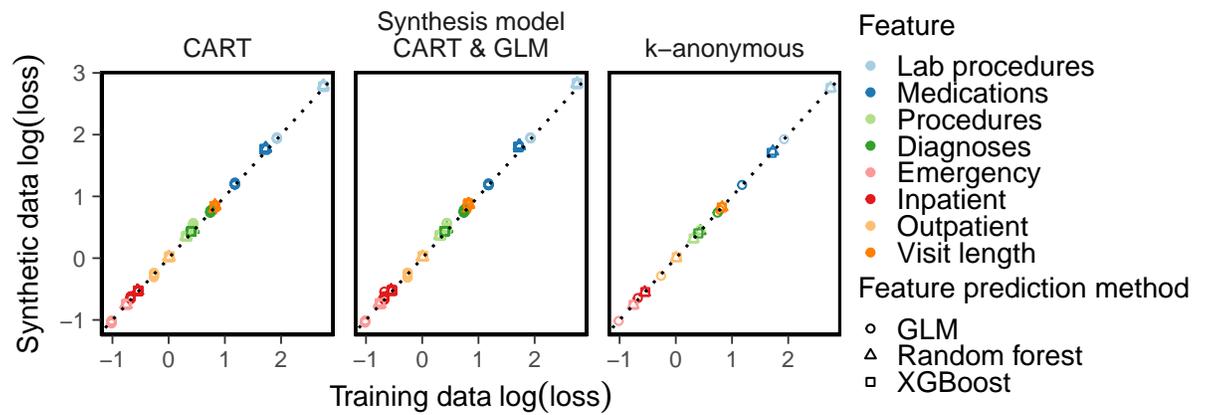


Figure 9.16: Feature prediction scores for numeric variables from the 130 hospitals dataset.

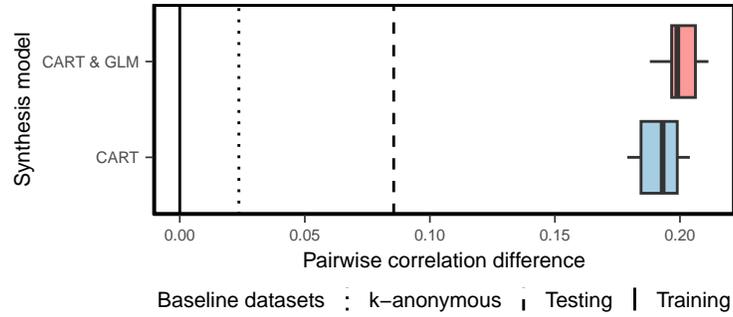


Figure 9.17: Pairwise correlation differences for synthesis replications.

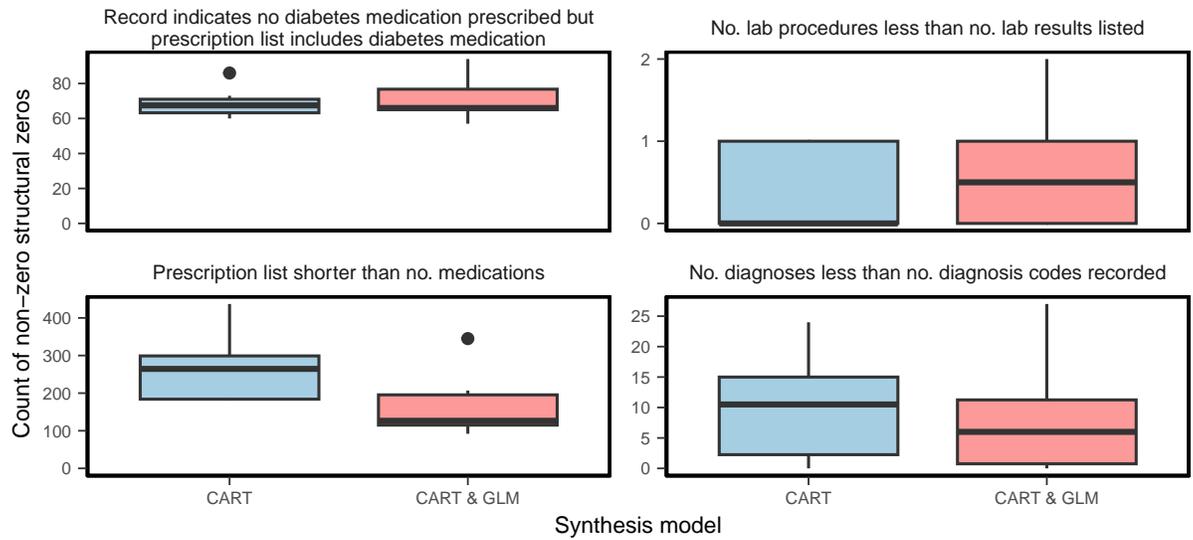


Figure 9.18: Counts of structural zero observations for synthetic replications.

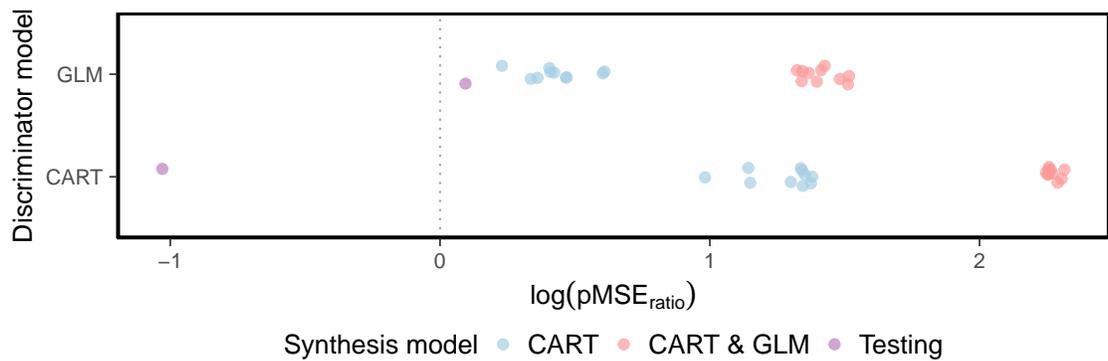


Figure 9.19: pMSE score ratios for synthetic replications of 130 hospitals data.

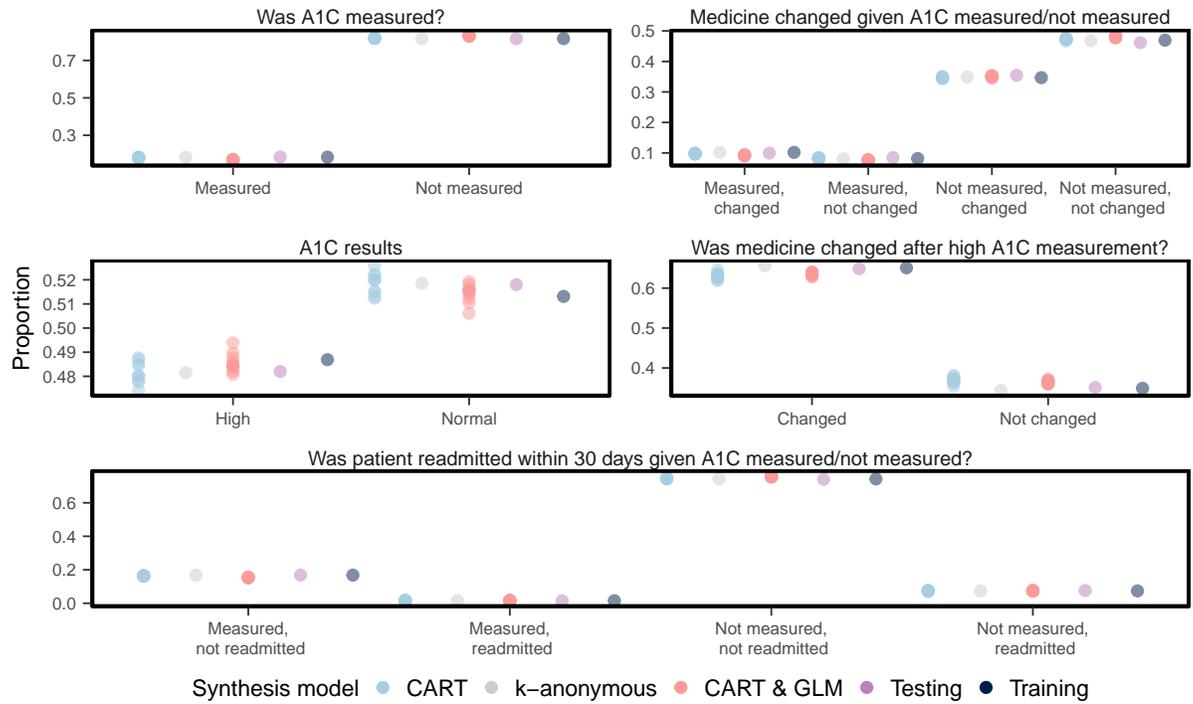


Figure 9.20: Sample statistics for synthetic and baseline 130 hospitals datasets.

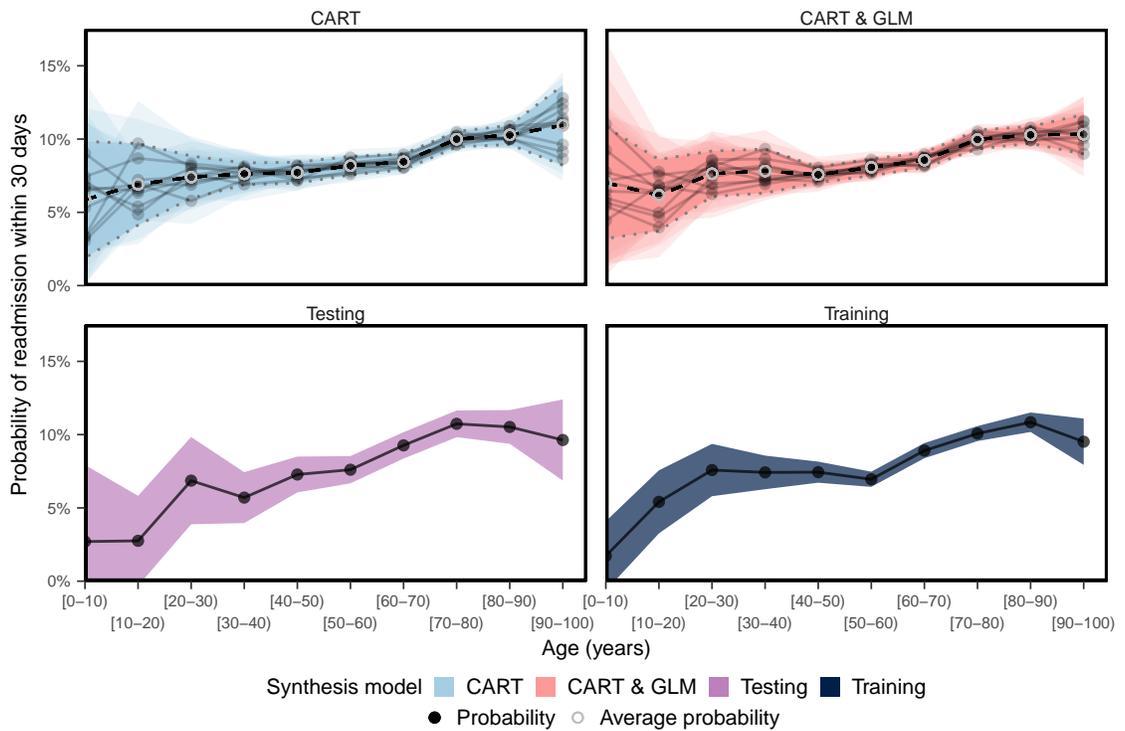


Figure 9.21: Percentage of re-admissions for different age brackets in synthetic and baseline 130 hospitals datasets.

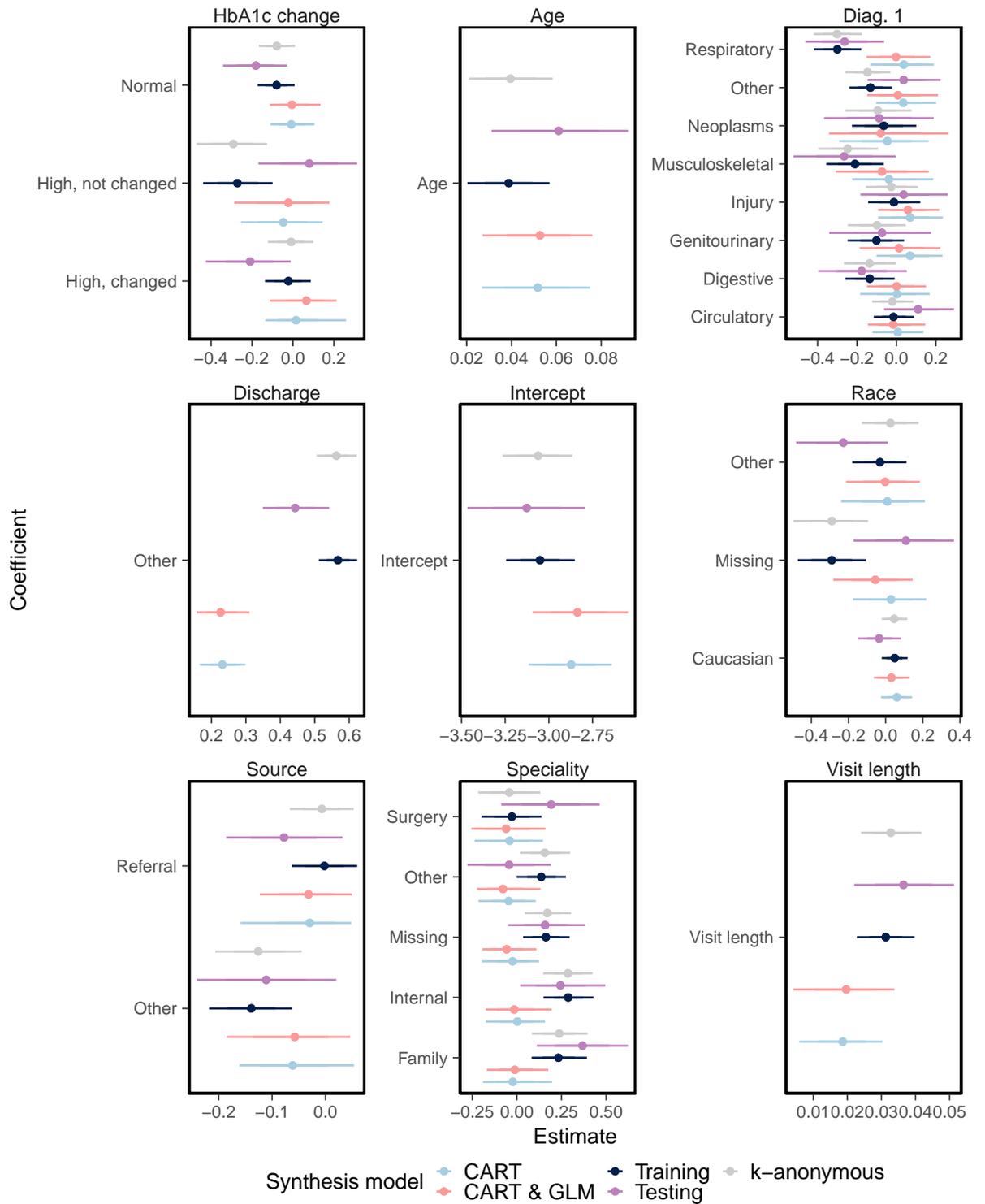


Figure 9.22: 90% credible intervals for coefficients of Bayesian regression model with Gaussian priors (Equation (9.1)).

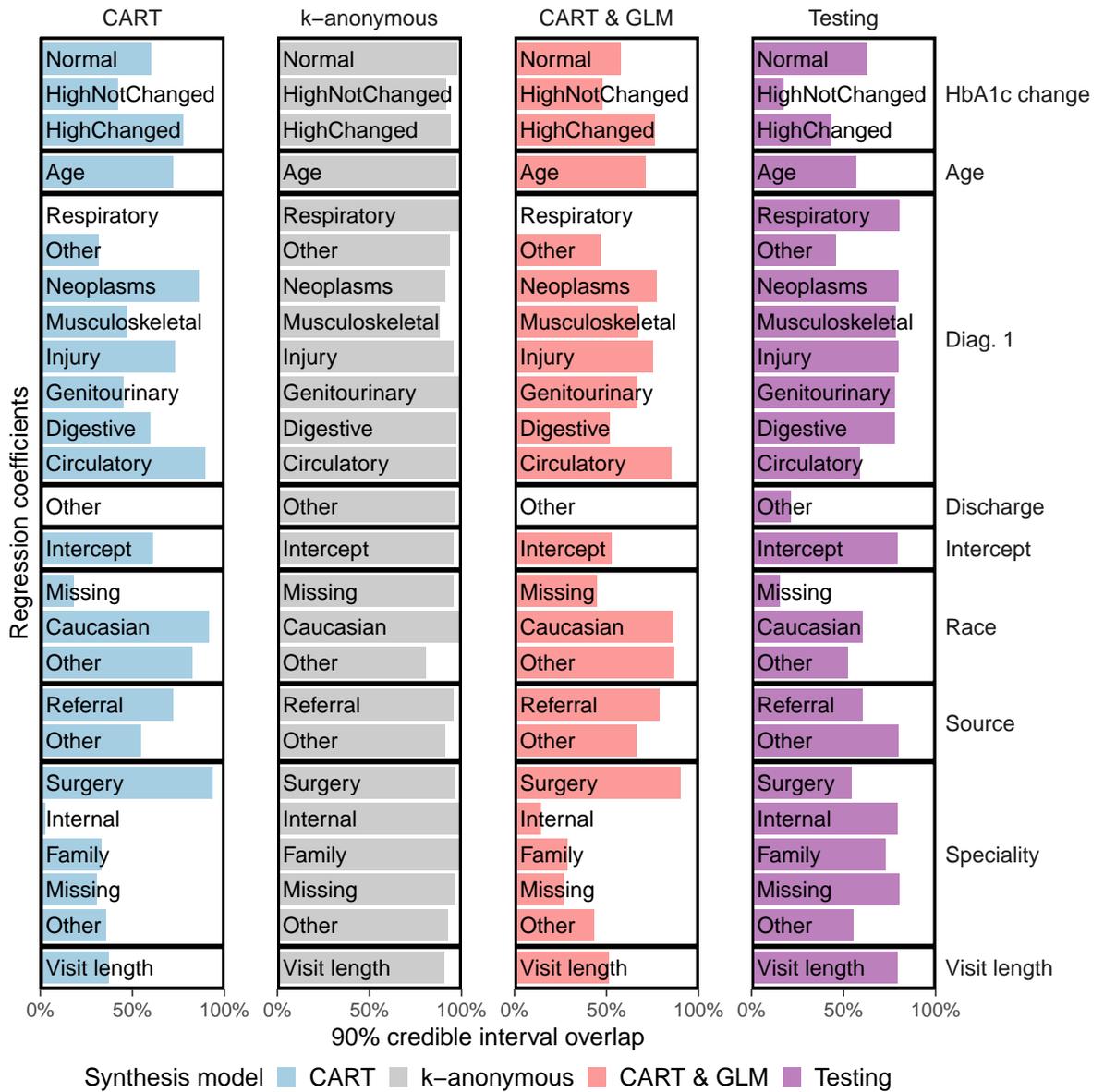


Figure 9.23: 90% credible interval overlap for coefficients of Bayesian regression model (Equation (9.1)).

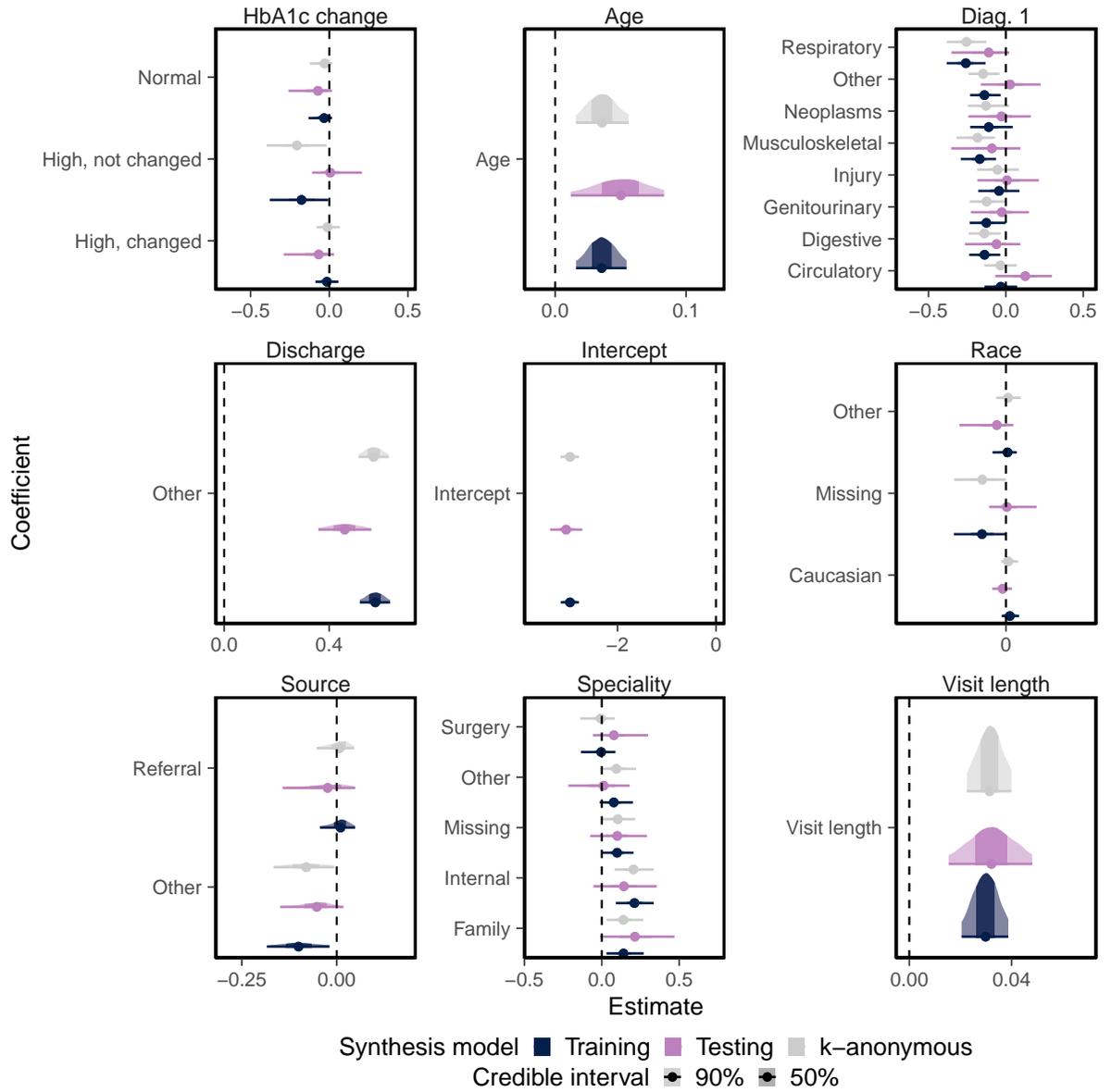


Figure 9.24: 90% credible intervals for coefficients of Bayesian regression model with horseshoe priors.

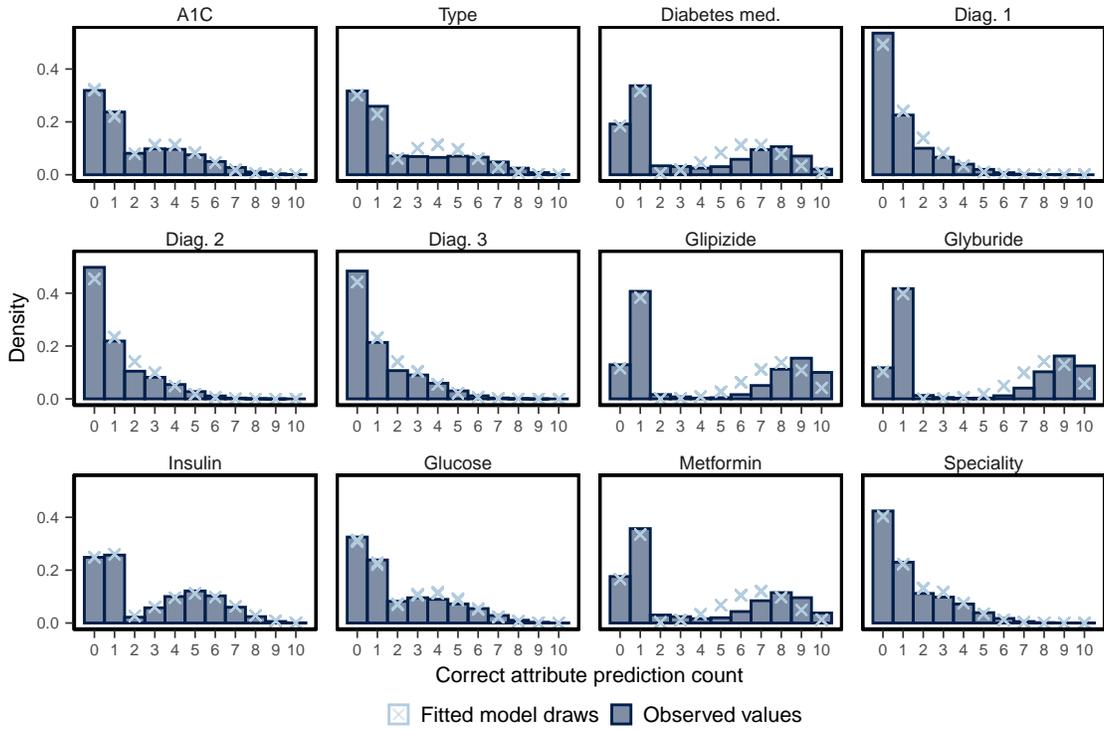


Figure 9.25: Observed and predicted values (50 replications from fitted model (9.4)) for each attribute in validation data.

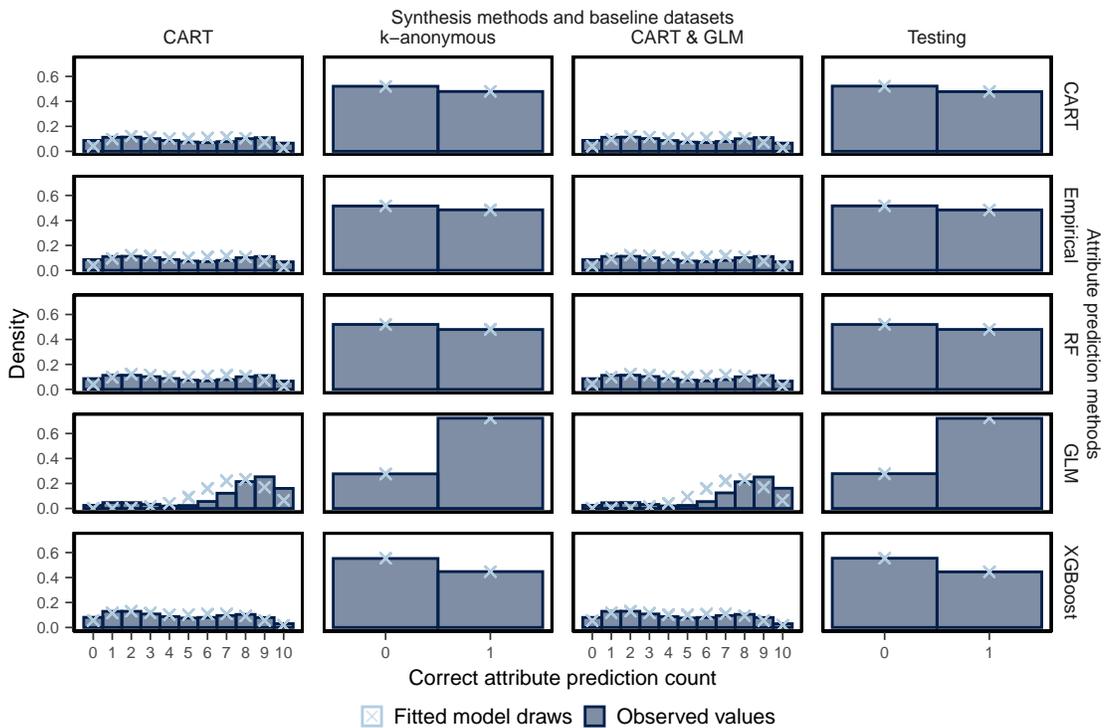


Figure 9.26: Observed and predicted values (50 replications from fitted model (9.4)) for each attribute prediction model and synthesis model in validation data.

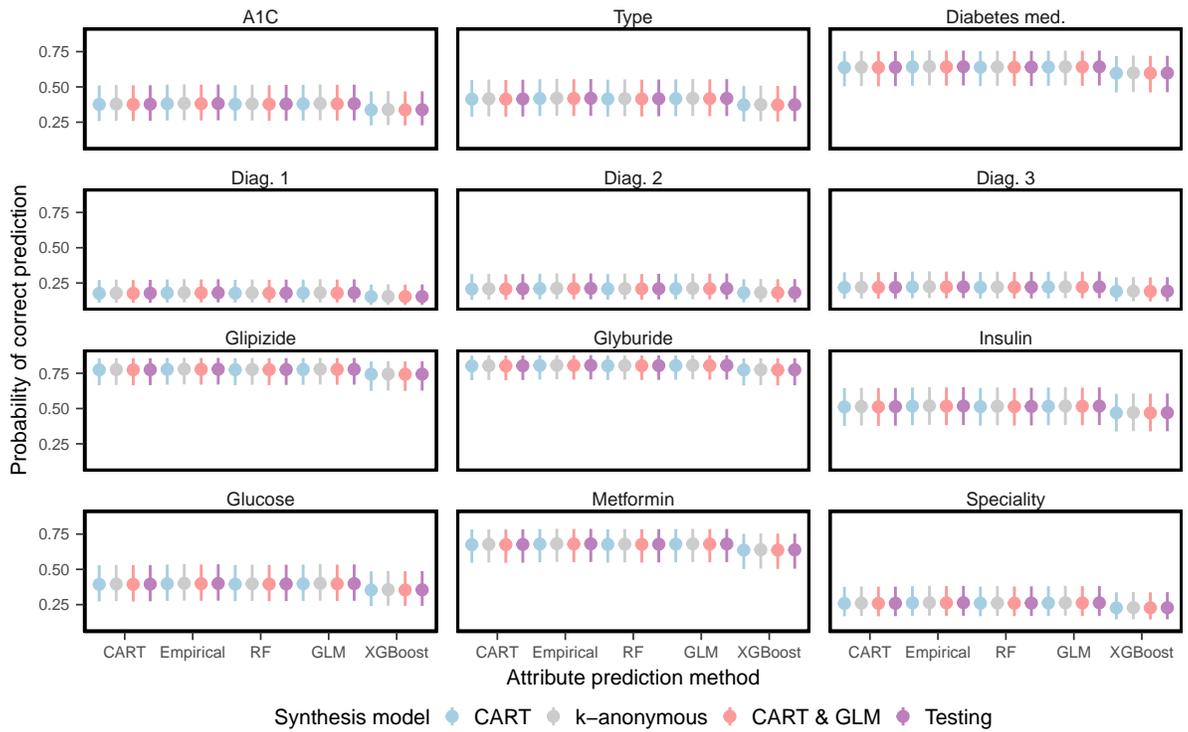


Figure 9.27: 95% prediction intervals from the fitted model (Equation (9.4)), conditioned on the uncertainty of fixed effects only.

9.4 Conclusions

In this chapter, we expanded on the work of Chapter 8 by applying sequential synthesis methods to generate the larger 130 Hospitals dataset and then assessed the data by implementing the framework described in Chapter 7. Exploratory plots of the synthetic datasets demonstrate that the sequential synthesis was able to capture the univariate distributions of most variables and that conditional relationships were also reasonably well preserved. However, the propensity scores indicated that both discriminator models were able to differentiate between the training and synthetic datasets.

The synthetic datasets had worse performance on the inference task than the baseline datasets although, if we only consider the coefficients that were significant for the real data, the inference results were far more reasonable. An analyst who only had access to the synthetic data would be able to identify almost all significant coefficients that an analyst with access to the test data would. It would have been interesting if we had been able to fit the model with shrinkage priors, as the overlap may have been improved by the shrinking of coefficients with weaker effects. That said, the difficulty of fitting the models with shrinkage priors to some synthetic replications should be seen as a negative of the synthetic data.

In addition, we modified the attribute disclosure assessment from the previous chapter to model the probability of a correct prediction instead of the prediction loss. This change was necessary since we previously found it difficult to model the prediction losses. However, we prefer approach after the change, as it allows us to directly assess the probability that an intruder correctly predicts an attribute. Unfortunately, the changed approach is only suited to categorical variables, as it models disclosures as binary events. While an application of the method to numeric variables would be possible, this would involve the specification of a hard threshold, where a prediction within the threshold would be considered to be correct.

Our attribute disclosure results show that the probability of disclosure was very slightly lower for the synthetic data than either of the baseline datasets. However, the difference in probability between the synthesis and baseline datasets was smaller than the difference between attribute prediction method. In fact, the effect of any of the datasets on the probability of disclosure was negligible, in comparison to the effect of the attribute or subject. As with the previous chapter, this highlights the importance of carrying out an attribute disclosure assessment that considers multiple methods of attribute prediction.

In stark contrast to the previous chapter, the k -anonymous baseline performed well for every utility assessment. A large reason for this difference is that 2% of observations were removed with k -anonymisation in this chapter, in comparison to removing 75% of observations in the Pima chapter. The reduction in the quality of the data will be drastically different between the two k -anonymous datasets. The choice of appropriate comparisons

for utility and disclosure risk warrants further discussion. Such discussion is included in the next chapter, Chapter 10.

Overall, the results of this chapter show that larger datasets generated with sequential synthesis methods look somewhat realistic. However, the synthetic data performed more poorly on some of the more extensive checks than in the Pima scenario of the previous chapter. This reflects the additional complications of synthesising a dataset that contains many more variables. The larger size of the data also contributed to issues with both generation and assessment of the data, as we found it prohibitively slow to synthesise variables with random forest models, to calculate distances between observations for membership disclosure assessments and identifying outliers, and to run hierarchical Bayesian inference models in our inference of the disclosure assessment results.

Chapter 10

Discussion

10.1 Discussion of Literature Review

In this thesis, we carried out an extensive review of a variety of utility and disclosure risk assessment methods. This review does not exist elsewhere in the literature. Others have noted issues and a lack of consensus for methods of assessing synthetic data. Our critiques and discussions of the limitations of utility assessments is informative for future approaches for evaluating the utility of synthetic data. We discussed the assessment of disclosure risk for completely synthetic data and argued that attribute disclosure is the most reasonable choice. Our views on this are echoed by Reiter (2023). Despite the arguments for assessing attribute disclosure risk, much of the synthetic data literature prefers to assess membership disclosure risk. We suspect that the ease of assessing the results of membership disclosure in comparison to attribute disclosure plays a large part in this preference. Furthermore, we discussed some issues with the current methods for attribute disclosure assessment.

10.2 Our Framework

In Chapter 7, we described our methodology for the assessment of synthetic data. This builds on the lessons learned in our literature review. The broad range of utility assessments implemented in this thesis echoes the examples of others in the literature. Given the issues discussed in our literature review, we believe that considering a variety of assessments is the best approach to assessing utility.

Our practical framework for the assessment of disclosure risk builds on the work of Elliot (2014) and Taub et al. (2018). We addressed limitations of their work (mentioned in our literature review) by extending their work to a wider set of intruder assumptions. Our framing of disclosure assessment in general terms allows for others to integrate their own beliefs and assumptions about a disclosure scenario into their assessment.

Furthermore, we described and implemented a novel approach of modelling the disclosure risk scores and then evaluating the models. This addresses the difficulty of assessing attribute disclosure in the scenario of there being many sensitive attributes. Additionally, modelling the scores allows for the simultaneous evaluation of multiple sets of prior assumptions and data synthesis methods. This addresses other limitations that previously existed with many approaches to disclosure assessment. We demonstrated our framework for two comprehensive examples. In Chapter 8, eight variants of three synthetic data generation methods were compared on a small dataset. In Chapter 9, we demonstrate the framework on a much larger and more complex dataset that contains over 50,000 observations and 40 variables.

10.3 Our Utility Assessment

The results of utility assessments in Chapter 8 highlighted that many utility assessments were unable to identify differences between the real and synthetic data even when other assessments showed clear differences in the quality of the data. The results of Chapter 9 were similar. However, the additional complexity of the data led to poorer quality synthesis. As such, more utility assessments were able to identify differences. Despite many assessments failing to identify differences, the combination of assessment results highlighted clear patterns in both datasets. This supports our conclusions in the literature review and further justifies the advantages of evaluating synthetic data with a diverse and varied set of utility assessments.

We implemented several utility assessment that were based on the results of hypothesis tests. Assessments of this type have been considered before but our results showed that such comparisons are problematic as they depend heavily on the choice of hypothesis. In the first hypothesis test example in Chapter 8, the best synthetic dataset depended on whether weakly informative or regularising priors were chosen. In the second hypothesis test example in Chapter 9, we compared Bayesian regression models fit to real and synthetic datasets. The synthetic data performance was poor in comparison to the training data but a comparison of the training and test data indicated the extreme over-fitting of the procedure. Researchers should be aware of the potential to unintentionally bias their results by implementing such subjective assessments.

10.4 Our Disclosure Risk Assessment

Our disclosure assessment of the Pima data in Chapter 8 considered the attribute disclosure risk of many synthetic datasets for a large set of intruder prior assumptions. Consequently, there were a large number of comparisons and a difficult distribution, which

required us to fit a very complex distributional model.

In the 130 hospitals example in Chapter 9, we considered the disclosure risk of many more sensitive attributes. The increased number of sensitive attributes and the observations necessitated a different modelling approach. By reframing intruder predictions as a set of Bernoulli trials we were able to drastically simplify the problem and fit a far simpler model to the data than in the previous chapter. The potential of this approach to attribute disclosure risk assessment is very exciting as it allows for subject level disclosure risk to be simultaneously modelled for large numbers of comparisons.

The inference of the disclosure results in both Chapter 8 and Chapter 9 highlighted that the impact of synthetic data generation model on disclosure risk was small in comparison to both attribute prediction methods and attributes. This is noteworthy because it indicates the potential to draw misleading conclusions from disclosure assessments if they only consider a single set of assumptions about the intruder. Given this finding, we must be wary of the results of disclosure assessments that only consider a single set of assumptions about the intruder. Furthermore, the results justify our choice to consider several intruder prediction methods and show why our approach to modelling the results simultaneously is necessary.

In Chapters 8 and 9, we compared synthetic data to two baseline datasets. In Chapter 8, the disclosure risk of the synthetic datasets are comparable to the test baseline, while the disclosure risk was lower for the 130 hospitals dataset in Chapter 9. This difference likely reflects the lower quality of the synthetic 130 hospitals data. Disclosure risk of the synthetic data in comparison to the k -anonymous baseline changed from Chapter 8 to Chapter 9. This reflects the clear difference of the quality of the datasets for the respective examples.

10.5 Limitations and future work

Through our examples we have demonstrated the potential of our methods for the assessment of attribute disclosure risk. However, these demonstrations are limited to a small set of synthesis models and datasets. As such, the exploration of other examples would highlight unforeseen issues that need to be addressed.

In hindsight, the choice to implement k -anonymous data as a baseline was uninformative. While there is a genuine use case for the comparison of the disclosure risk of synthetic data with a non-synthetic statistical disclosure control (SDC), the sample average baseline that Taub et al. (2018) implements reflects a more reasonable disclosure scenario. Furthermore, it would simplify the analysis.

The difficulties faced in Chapter 8 demonstrated the problems of a simultaneous comparison of disclosure risk for numeric and categorical variables. However, implementing

numeric variables into the Bernoulli model of Chapter 9 would be possible and interesting to explore further. Hu and Savitsky (2021) implement one such approach, where numeric variables are correctly predicted if the prediction is within a predefined threshold. The requirement to specify this threshold is not ideal, but the payoff of simultaneous risk assessment of numeric and categorical variables would justify this.

Appendix A

Additional tables

Table A.1: The closest pairs of training (original) and synthetic Pima observations for each data synthesis method when matching on all variables.

*Training set outliers

Dataset	Distance	Preg.	Gluc.	BP	Skin.	BMI	DPF	Age	Diab.
Training	6.68E-03	1	107	68	19	26.5	0.165	24	0
Regression (pen.)	6.68E-03	1	108	67	18	27.3	0.131	23	0
Training	8.34E-03	1	111	62	13	24.0	0.138	23	0
Regression (no pen.)	8.34E-03	1	112	62	11	24.7	0.176	24	0
Training	1.10E-03	1	96	64	27	33.2	0.289	21	0
CART (smooth leav.)	1.10E-03	1	96	64	28	33.2	0.277	21	0
Training	2.12E-05	1	96	64	27	33.2	0.289	21	0
CART (smooth data)	2.12E-05	1	96	64	27	33.2	0.293	21	0
Training	0	1	99	58	10	25.4	0.551	21	0
CART (no smooth)	0	1	99	58	10	25.4	0.551	21	0
Training	0	9	152	78	34	34.2	0.893	33	1
RF (smooth leav.)	0	9	152	78	34	34.2	0.893	33	1
Training	0	1	99	58	10	25.4	0.551	21	0
RF (smooth leav.)	0	1	99	58	10	25.4	0.551	21	0
Training	0	0	137	68	14	24.8	0.143	21	0
RF (smooth leav.)	0	0	137	68	14	24.8	0.143	21	0
Training	0	6	80	66	30	26.2	0.313	41	0
RF (smooth leav.)	0	6	80	66	30	26.2	0.313	41	0
Training	0	2	122	70	27	36.8	0.340	27	0
RF (smooth leav.)	0	2	122	70	27	36.8	0.340	27	0
Training	0	1	120	80	48	38.9	1.162	41	0
RF (smooth leav.)	0	1	120	80	48	38.9	1.162	41	0

Training	0	1	111	62	13	24.0	0.138	23	0
RF (smooth leav.)	0	1	111	62	13	24.0	0.138	23	0
Training	0	2	122	76	27	35.9	0.483	26	0
RF (smooth leav.)	0	2	122	76	27	35.9	0.483	26	0
Training	0	7	114	76	17	23.8	0.466	31	0
RF (smooth leav.)	0	7	114	76	17	23.8	0.466	31	0
Training	1.32E-06	3	171	72	33	33.3	0.199	24	1
RF (smooth data)	1.32E-06	3	171	72	33	33.3	0.200	24	1
Training	1.32E-06	4	154	72	29	31.3	0.338	37	0
RF (smooth data)	1.32E-06	4	154	72	29	31.3	0.337	37	0
Training	0	4	99	72	17	25.6	0.294	28	0
RF (no smooth)	0	4	99	72	17	25.6	0.294	28	0
Training	0	9	154	78	30	30.9	0.164	45	0
RF (no smooth)	0	9	154	78	30	30.9	0.164	45	0
Training*	0	1	144	82	46	46.1	0.335	46	1
RF (no smooth)	0	1	144	82	46	46.1	0.335	46	1
Training	0	5	77	82	41	35.8	0.156	35	0
RF (no smooth)	0	5	77	82	41	35.8	0.156	35	0
Training	0	12	140	82	43	39.2	0.528	58	1
RF (no smooth)	0	12	140	82	43	39.2	0.528	58	1
Training	0	5	97	76	27	35.6	0.378	52	1
RF (no smooth)	0	5	97	76	27	35.6	0.378	52	1
Training	0	5	139	80	35	31.6	0.361	25	1
RF (no smooth)	0	5	139	80	35	31.6	0.361	25	1
Training*	0	3	191	68	15	30.9	0.299	34	0
RF (no smooth)	0	3	191	68	15	30.9	0.299	34	0
Training	0	9	112	82	32	34.2	0.260	36	1
RF (no smooth)	0	9	112	82	32	34.2	0.260	36	1
Training	0	3	99	80	11	19.3	0.284	30	0
RF (no smooth)	0	3	99	80	11	19.3	0.284	30	0
Training	0	5	158	84	41	39.4	0.395	29	1
RF (no smooth)	0	5	158	84	41	39.4	0.395	29	1
Training*	0	1	71	78	50	33.2	0.422	21	0
RF (no smooth)	0	1	71	78	50	33.2	0.422	21	0
Training*	0	8	167	106	46	37.6	0.165	43	1
RF (no smooth)	0	8	167	106	46	37.6	0.165	43	1
Training	0	7	114	76	17	23.8	0.466	31	0

RF (no smooth)	0	7	114	76	17	23.8	0.466	31	0
----------------	---	---	-----	----	----	------	-------	----	---

Table A.2: List of variables from Diabetes-130 dataset and their descriptions (Strack et al., 2014).

Variable name	Type	Description and values	% missing
Encounter ID	Numeric	Unique identifier of an encounter	0%
Patient number	Numeric	Unique identifier of a patient	0%
Race	Nominal	Values: Caucasian, Asian, African American, Hispanic, and other	2%
Gender	Nominal	Values: male, female, and unknown/invalid	0%
Age	Nominal	Grouped in 10-year intervals:[0, 10],[10, 20),..., [90, 100)	0%
Weight	Numeric	Weight in pounds	97%
Admission type	Nominal	Integer identifier corresponding to 9 distinct values, for example, emergency, urgent, elective, newborn, and not available	0%
Discharge disposition	Nominal	Integer identifier corresponding to 29 distinct values, for example, discharged to home, expired, and not available	0%
Admission source	Nominal	Integer identifier corresponding to 21 distinct values, for example, physician referral, emergency room, and transfer from a hospital	0%
Time in hospital	Numeric	Integer number of days between admission and discharge	0%
Payer code	Nominal	Integer identifier corresponding to 23 distinct values, for example, Blue Cross/Blue Shield, Medicare, and self-pay	53%
Medical Speciality	Nominal	Integer identifier of a speciality of the admitting physician, corresponding to 84 distinct values, for example, cardiology, internal medicine, family/general practice, and surgeon	0%
Number of lab procedures	Numeric	Number of lab tests performed during the encounter	0%
Number of procedures	Numeric	Number of procedures (other than lab tests) performed during the encounter	0%

Number of medications	Numeric	Number of distinct generic names administered during the encounter	0%
Number of outpatient visits	Numeric	Number of outpatient visits of the patient in the year preceding the encounter	0%
Number of emergency visits	Numeric	Number of emergency visits of the patient in the year preceding the encounter	0%
Number of inpatient visits	Numeric	Number of inpatient visits of the patient in the year preceding the encounter	0%
Diagnosis 1	Nominal	The primary diagnosis (coded as first three digits of ICD9); 848 distinct values	0%
Diagnosis 2	Nominal	Secondary diagnosis (coded as first three digits of ICD9); 923 distinct values	0%
Diagnosis 3	Nominal	Additional secondary diagnosis (coded as first three digits of ICD9); 954 distinct values	1%
Number of diagnoses	Numeric	Number of diagnoses entered to the system	0%
Glucose serum test result	Nominal	Indicates the range of the result or if the test was not taken. Values: > 200, > 300, normal, and none" if not measured	0%
A1c test result	Nominal	Indicates the range of the result or if the test was not taken. Values: > 8 if the result was greater than 8%, > 7 if the result was greater than 7% but less than 8%, normal if the result was less than 7%, and none if not measured	0%
Change of medications	Nominal	Indicates if there was a change in diabetic medications (either dosage or generic name). Values: change and no change	0%
Diabetes medications	Nominal	Indicates if there was any diabetic medication prescribed. Values: yes and no	0%

24 ables medica- tions	vari- for	Nominal	For the generic names: metformin, repaglinide, nateglinide, chlorpropamide, glimepiride, acetohexamide, glipizide, glyburide, tolbutamide, pioglitazone, rosiglitazone, acarbose, miglitol, troglitazone, tolazamide, examide, sitagliptin, insulin, glyburide-metformin, glipizide-metformin, glimepiride-pioglitazone, metformin-rosiglitazone, and metformin-pioglitazone, the feature indicates whether the drug was prescribed or there was a change in the dosage. Values: up if the dosage was increased during the encounter, down if the dosage was decreased, steady if the dosage did not change, and no if the drug was not prescribed	0%
Readmitted		Nominal	Days to inpatient readmission. Values: < 30' if the patient was readmitted in less than 30 days, > 30' if the patient was readmitted in more than 30 days, and No for no record of readmission.	0%

Table A.3: List of groupings made to levels of categorical variables in Diabetes-130 dataset.

Grouped categories	Subcategories	%
Primary diagnoses		
Circulatory	390–459, 785	30.56
Respiratory	460–519, 786	13.56
Digestive	520–579, 787	9.27
Diabetes	250.xx	8.21
Injury	800–999	6.71
Musculoskeletal	710–739	5.81
Genitourinary	580–629, 788	4.92
Neoplasms	140–239	3.63
Other	—	17.34
Discharge Disposition		
Discharged to home	Discharged to home	63.33
Missing	Not Mapped, NULL	4.65
Other	—	32.03
Admission Source		
Physician/Clinical referral	Physician Referral, Clinic Referral	32.37
Emergency Room	Emergency Room	53.25
Missing	Not Mapped, NULL	7.11

Grouped categories	Subcategories	%
Other	—	7.27
<hr/> Medical Speciality <hr/>		
Surgery	All surgical categories	5.36
Internal Medicine	Internal Medicine	15.20
Cardiology	Cardiology, Cardiology-Pediatric	6.02
Family or General Practice	Family/General Practice	7.11
Missing	Missing	48.08
Other	—	18.22
<hr/> Readmitted <hr/>		
Within 30 days	< 30	8.98
Not within 30 days	> 30, Not readmitted	91.02
<hr/> Race <hr/>		
African American	African American	18.04
Caucasian	Caucasian	74.74
Other	Hispanic, Asian, Other	4.48
Missing	Missing	2.74
<hr/> Gender <hr/>		
Female	—	53.21
Male	—	46.79
<hr/> Age <hr/>		
Under 30	[0–10), [10–20), [20–30)	2.58
30 to 60	[30–40), [40–50), [50–60)	31.25
Over 60	[60–70), [70–80), [80–90), [90–100)	66.17
<hr/> HbA1c <hr/>		
High & changed	HbA1c > 8 & medicine changed	5.80
High & not changed	HbA1c > 8 & medicine not changed	3.12
Normal	HbA1c > 7, Normal	9.44
Not Measured	HbA1c Not Measured	81.65

Bibliography

- Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., & Zhang, L. (2016, October 14). *Deep learning with differential privacy* (2). arXiv: [1607.00133](https://arxiv.org/abs/1607.00133) [stat.ML]. <https://doi.org/10.48550/arXiv.1607.00133>
- Abowd, J. M., Stinson, M., & Benedetto, G. (2006). *Final report to the Social Security Administration on the SIPP/SSA/IRS public use file project* (tech. rep.). U.S. Census Bureau. Retrieved August 4, 2023, from <https://www.census.gov/library/working-papers/2006/adrm/2006-11-AbowdStinsonBenedetto-SSAfinal.html>
- Alfons, A., Kraft, S., Templ, M., & Filzmoser, P. (2011). Simulation of close-to-reality population data for household surveys with application to EU-SILC. *Statistical Methods and Applications*, *20*(3), 383–407. <https://doi.org/10.1007/s10260-011-0163-2>
- Arjovsky, M., & Bottou, L. (2017, January 17). *Towards principled methods for training generative adversarial networks* (1). arXiv: [1701.04862](https://arxiv.org/abs/1701.04862) [stat.ML].
- Arjovsky, M., Chintala, S., & Bottou, L. (2017, December 6). *Wasserstein GAN* (3). arXiv: [1701.07875](https://arxiv.org/abs/1701.07875) [stat.ML].
- Backes, M., Berrang, P., Humbert, M., & Manoharan, P. (2016). Membership privacy in microrna-based studies. *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, 319–330. <https://doi.org/10.1145/2976749.2978355>
- Barrientos, A. F., Williams, A. R., Snoke, J., & Bowen, C. M. (2023). *A feasibility study of differentially private summary statistics and regression analyses with evaluations on administrative and survey data* (3). arXiv: [2110.12055](https://arxiv.org/abs/2110.12055) [stat.AP].
- Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. <https://doi.org/10.18637/jss.v067.i01>
- Beaulieu-Jones, B. K., Wu, Z., Williams, C., Lee, R., Bhavnani, S. P., Byrd, J. B., & Greene, C. S. (2019). Privacy-preserving generative deep neural networks support clinical data sharing. *Circulation: Cardiovascular Quality and Outcomes*, *12*(7), e005122. <https://doi.org/10.1161/CIRCOUTCOMES.118.005122>

- Benedetto, G., Stanley, J. C., & Totty, E. (2017). *The creation and use of the SIPP Synthetic Beta v7.0* (tech. rep.). U.S. Census Bureau. Retrieved August 4, 2023, from <https://www.census.gov/library/working-papers/2018/adrm/SIPP-Synthetic-Beta.html>
- Benedetto, G., Stinson, M. H., & Abowd, J. M. (2013). *The creation and use of the SIPP Synthetic Beta* (tech. rep.). U.S. Census Bureau. Retrieved August 4, 2023, from https://www.census.gov/content/dam/Census/programs-surveys/sipp/methodology/SSBdescribe_nontechnical.pdf
- Bennett, J., & Lanning, S. (2007). The netflix prize. *Proceedings of KDD Cup and Workshop 2007*. Retrieved December 30, 2023, from <https://www.cs.uic.edu/~liub/KDD-cup-2007/proceedings/The-Netflix-Prize-Bennett.pdf>
- Bernaards, C. A., Belin, T. R., & Schafer, J. L. (2006). Robustness of a multivariate normal approximation for imputation of incomplete binary data. *Statistics in Medicine*, 26(6), 1368–1382. <https://doi.org/10.1002/sim.2619>
- Betancourt, M. (2021). *Sparsity blues*. Retrieved January 5, 2024, from https://betanalpha.github.io/assets/case_studies/modeling_sparsity.html
- Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.
- Bowen, C. M., & Snoke, J. (2020, October 12). *Comparative study of differentially private synthetic data algorithms from the NIST PSCR differential privacy synthetic data challenge* (3). arXiv: 1911.12704 [stat.AP]. <https://doi.org/10.48550/arXiv.1911.12704>
- Brand, J. P. L. (1999). *Development, implementation and evaluation of multiple imputation strategies for the statistical analysis of incomplete data sets* (Doctoral dissertation). Erasmus University Rotterdam.
- Breiman, L. (1984). *Classification and regression trees*. Routledge. <https://doi.org/10.1201/9781315139470>
- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32. <https://doi.org/10.1023/A:1010933404324>
- Breunig, M. M., Kriegel, H.-P., Ng, R. T., & Sander, J. (2000). LOF: Identifying density-based local outliers. *Proceedings of the 2000 ACM SIGMOD*, 93–104.
- Bürkner, P.-C. (2017). brms: An R package for bayesian multilevel models using stan. *Journal of Statistical Software*, 80(1), 1–28. <https://doi.org/10.18637/jss.v080.i01>
- Caiola, G., & Reiter, J. P. (2010). Random forests for generating partially synthetic, categorical data. *Transactions on Data Privacy*, 3(1), 27–42.
- California Consumer Privacy Act of 2018, AB 375 (2018). https://leginfo.ca.gov/faces/billTextClient.xhtml?bill_id=201720180AB375
- California Privacy Rights Act of 2020, Proposition 24 (2020). <https://vig.cdn.sos.ca.gov/2020/general/pdf/topl-prop24.pdf>

- Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M. A., Guo, J., Li, P., & Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, 76(1).
- Carvalho, C. M., Polson, N. G., & Scott, J. G. (2009). Handling sparsity via the horseshoe. In D. van Dyk & M. Welling (Eds.), *Proceedings of the twelfth international conference on artificial intelligence and statistics* (pp. 73–80). PMLR.
- Centers for Disease Control and Prevention. (2019). *Diabetes tests*. Retrieved June 6, 2023, from <https://www.cdc.gov/diabetes/basics/getting-tested.html>
- Che, Z., Purushotham, S., Li, G., Jiang, B., & Liu, Y. (2018, July). Hierarchical deep generative models for multi-rate multivariate time series. In J. Dy & A. Krause (Eds.), *Proceedings of the 35th international conference on machine learning* (pp. 784–793). PMLR. <https://proceedings.mlr.press/v80/che18a.html>
- Chen, T., & Guestrin, C. (2016). *XGBoost: A scalable tree boosting system* (3). arXiv: 1603.02754 [cs.LG]. <https://doi.org/10.48550/arXiv.1603.02754>
- Choi, E., Biswal, S., Malin, B., Duke, J., Stewart, W. F., & Sun, J. (2017). *Generating multi-label discrete patient records using generative adversarial networks*. arXiv: 1703.06490 [cs].
- Choquette-Choo, C. A., Tramèr, F., Carlini, N., & Papernot, N. (2020, December 5). *Label-only membership inference attacks*. arXiv: [cs/2007.14321](https://arxiv.org/abs/2007.14321).
- Collier, M., Nazabal, A., & Williams, C. K. I. (2021, March 21). *VAEs in the presence of missing data* (3). arXiv: 2006.05301 [cs.LG]. <https://doi.org/10.48550/arxiv.2006.05301>
- Cutforth, M., Watson, H., Brown, C., Wang, C., Thomson, S., Fell, D., Dilys, V., Scrimgeour, M., Schrempf, P., Lesh, J., Muir, K., Weir, A., & O’Neil, A. Q. a. (2023). Acute stroke CDS: Automatic retrieval of thrombolysis contraindications from unstructured clinical letters. *Frontiers in Digital Health*, 5, 1186516. <https://doi.org/10.3389/fdgth.2023.1186516>
- Dalenius, T. (1986). Finding a needle in a haystack or identifying anonymous census records. *Journal of Official Statistics*, 2(3), 329–336.
- Data Protection Act 2018, United Kingdom of Great Britain, Northern Ireland (2018). <https://www.legislation.gov.uk/ukpga/2018/12/contents>
- Debnath, A., Gupta, N., Waghmare, G., Wadhwa, H., Asthana, S., & Arora, A. (2021). Adversarial generation of temporal data: A critique on fidelity of synthetic data. *Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, 306–321.
- Dennett, A. (2017). *Synthetic LS spines*. Retrieved August 4, 2023, from <https://github.com/adamdennett/Synthetic-LS-spines>

- Dennett, A., Norman, P., Shelton, N., & Stuchbury, R. (2016). A synthetic longitudinal study dataset for England and Wales. *Data in Brief*, 9, 85–89. <https://doi.org/10.1016/j.dib.2016.08.036>
- Diabetes UK. (2017a). *Family relationships and type 1 diabetes*. Retrieved February 18, 2023, from <https://www.diabetes.org.uk/guide-to-diabetes/your-child-and-diabetes/family-relationships>
- Diabetes UK. (2017b). *Type 1 diabetes*. Retrieved February 18, 2023, from <https://www.diabetes.org.uk/diabetes-the-basics/types-of-diabetes/type-1>
- Diabetes UK. (2017c). *Types of diabetes*. Retrieved December 31, 2023, from <https://www.diabetes.org.uk/diabetes-the-basics/types-of-diabetes>
- Diabetes UK. (2017d). *What is gestational diabetes?* Retrieved December 31, 2023, from <https://www.diabetes.org.uk/diabetes-the-basics/gestational-diabetes/causes>
- Diabetes UK. (2019a). *Type 2 diabetes*. Retrieved February 18, 2023, from <https://www.diabetes.org.uk/diabetes-the-basics/types-of-diabetes/type-2>
- Diabetes UK. (2019b). *Type 2: Diabetes risk factors*. Retrieved December 31, 2023, from <https://www.diabetes.org.uk/diabetes-the-basics/types-of-diabetes/type-2/diabetes-risk-factors>
- Diabetes UK. (2023). *What causes type 2 diabetes?* Retrieved December 31, 2023, from <https://www.diabetes.org.uk/diabetes-the-basics/types-of-diabetes/type-2/causes>
- Dong, H.-W., Hsiao, W.-Y., Yang, L.-C., & Yang, Y.-H. (2017, September 19). *MuseGAN: Multi-track sequential generative adversarial networks for symbolic music generation and accompaniment* (2). arXiv: 1709.06298 [eess.AS]. <https://doi.org/10.48550/arxiv.1709.06298>
- Drechsler, J. (2009). Synthetic datasets for the German IAB establishment panel. *Working paper for the Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality*. Retrieved August 4, 2023, from <https://unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.46/2009/wp.10.e.pdf>
- Drechsler, J. (2010). Using support vector machines for generating synthetic datasets: UNESCO chair in data privacy international conference, PSD 2010 (J. Domingo-Ferrer & E. Magkos, Eds.).
- Drechsler, J. (2011a). *Synthetic datasets for statistical disclosure control* (P. Bickel, P. J. Diggle, S. Fienberg, U. Gather, I. Olkin, & S. Zeger, Eds.). Springer New York. <https://doi.org/10.1007/978-1-4614-0326-5>
- Drechsler, J. (2011b). Improved variance estimation for fully synthetic datasets. *In: Proceedings of the Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality*. https://unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.46/2011/18_Drechsler.pdf

- Drechsler, J. (2018). Some clarifications regarding fully synthetic data: UNESCO chair in data privacy international conference, PSD 2018 (J. Domingo-Ferrer & F. Montes, Eds.), 109–121.
- Drechsler, J., Bender, S., & Rässler, S. (2008). Comparing fully and partially synthetic datasets for statistical disclosure control in the German IAB establishment panel. *Transactions on Data Privacy*, 1(3), 105–130. <http://www.tdp.cat/issues/tdp.a006a08.pdf>
- Drechsler, J., Dundler, A., Bender, S., Rässler, S., & Zwick, T. (2008). A new approach for disclosure control in the IAB establishment panel — multiple imputation for a better data access. *AStA Advances in Statistical Analysis*, 92(4), 439–458. <https://doi.org/10.1007/s10182-008-0090-1>
- Drechsler, J., & Reiter, J. P. (2010). Sampling with synthesis: A new approach for releasing public use census microdata. *Journal of the American Statistical Association*, 105(492), 1347–1357. <https://doi.org/10.1198/jasa.2010.ap09480>
- Drechsler, J., & Reiter, J. P. (2011). An empirical evaluation of easily implemented, non-parametric methods for generating synthetic datasets. *Computational Statistics and Data Analysis*, 55(12), 3232–3243. <https://doi.org/10.1016/j.csda.2011.06.006>
- Dua, D., & Graff, C. (2017). UCI machine learning repository. <http://archive.ics.uci.edu/ml>
- Duchi, J. C. (2014). Derivations for linear algebra and optimization. Retrieved August 7, 2023, from https://stanford.edu/~jduchi/projects/general_notes.pdf
- Duncan, G., & Lambert, D. (1989). The risk of disclosure for microdata. *Journal of Business & Economic Statistics*, 7(2), 207–217. <https://doi.org/10.1080/07350015.1989.10509729>
- Dunson, D. B., & Xing, C. (2009). Nonparametric Bayes modeling of multivariate categorical data. *Journal of the American Statistical Association*, 104(487), 1042–1051. <https://doi.org/10.1198/jasa.2009.tm08439>
- Dwork, C. (2006). Differential privacy. In M. Bugliesi, B. Preneel, V. Sassone, & I. Wegener (Eds.), *Automata, languages and programming* (pp. 1–12). Springer Berlin Heidelberg. https://doi.org/10.1007/11787006_1
- Dwork, C., Kenthapadi, K., McSherry, F., Mironov, I., & Naor, M. (2006). Our data, ourselves: Privacy via distributed noise generation. In S. Vaudenay (Ed.), *Advances in cryptology - EUROCRYPT [2006]* (pp. 486–503). Springer Berlin Heidelberg.
- Dwork, C., & Roth, A. (2014). The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3-4), 211–407. <https://doi.org/10.1561/04000000042>

- Dwork, C., Smith, A., Steinke, T., Ullman, J., & Vadhan, S. (2015). Robust traceability from trace amounts. *2015 IEEE 56th Annual Symposium on Foundations of Computer Science*, 650–669. <https://doi.org/10.1109/FOCS.2015.46>
- El Emam, K., Mosquera, L., & Zheng, C. (2021). Optimizing the synthesis of clinical trial data using sequential trees. *Journal of the American Medical Informatics Association*, 28, 3–13.
- Elliot, M. (2005). Statistical disclosure control. In K. Kempf-Leonard (Ed.), *Encyclopedia of social measurement* (pp. 663–670). Elsevier. <https://doi.org/10.1016/B0-12-369398-5/00378-9>
- Elliot, M. (2014, October). *Final report on the disclosure risk associated with the synthetic data produced by the SYLLS team* (tech. rep.). University of Manchester.
- Elliot, M., Manning, A., & Ford, R. (2002). A computational algorithm for handling the special uniques problem. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10, 493–509. <https://doi.org/10.1142/S0218488502001600>
- Eyth, E., Basit, H., & Swift, C. J. (2023). Glucose tolerance test. *StatPearls*. Retrieved December 31, 2023, from <https://www.ncbi.nlm.nih.gov/books/NBK532915/>
- Fieller, E. C. (1954). Some problems in interval estimation. *Journal of the Royal Statistical Society. Series B (Methodological)*, 16(2), 175–185.
- Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1), 1–22.
- Frigerio, L., de Oliveira, A., Gomez, L., & Duverger, P. (2019, March 6). *Differentially private generative adversarial networks for time series, continuous, and discrete open data* (2). arXiv: 1901.02477 [cs.CR]. <https://doi.org/10.48550/arXiv.1901.02477>
- Gabry, J., & Goodrich, B. (2020). *Estimating regularized linear models with rstanarm*. Retrieved November 10, 2022, from <http://mc-stan.org/rstanarm/articles/lm.html>
- Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, 1(3), 515–534. <https://doi.org/10.1214/06-BA117A>
- Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014). *Bayesian data analysis* (3rd ed.). Taylor & Francis.
- Gelman, A., & Loken, E. (2013). *The garden of forking paths: Why multiple comparisons can be a problem, even when there is no “fishing expedition” or “p-hacking” and the research hypothesis was posited ahead of time*. http://www.stat.columbia.edu/~gelman/research/unpublished/p_hacking.pdf
- Golub, G. H., & Van Loan, C. F. (1996). *Matrix computations* (3rd ed.). The John Hopkins University Press.

- Goncalves, A., Ray, P., Soper, B., Stevens, J., Coyle, L., & Sales, A. P. (2020). Generation and evaluation of synthetic patient data. *BMC Medical Research Methodology*, 20(1), 108–. <https://doi.org/10.1186/s12874-020-00977-1>
- Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014, June 10). *Generative Adversarial Networks*. arXiv: [stat/1406.2661](https://arxiv.org/abs/1406.2661).
- Green, P. J. (1984). Iteratively reweighted least squares for maximum likelihood estimation, and some robust and resistant alternatives. *Journal of the Royal Statistical Society. Series B (Methodological)*, 46(2), 149–170. [https://doi.org/https://doi.org/10.1111/j.2517-6161.1984.tb01288.x](https://doi.org/10.1111/j.2517-6161.1984.tb01288.x)
- Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., & Courville, A. (2017). Improved training of wasserstein gans. *Advances in Neural Information Processing Systems, 2017-December*.
- Guo, J., Lu, S., Cai, H., Zhang, W., Yu, Y., & Wang, J. (2017, December 8). *Long text generation via adversarial training with leaked information* (2). arXiv: [1709.08624](https://arxiv.org/abs/1709.08624) [cs.CL]. <https://doi.org/10.48550/arxiv.1709.08624>
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer New York.
- Hayes, J., Melis, L., Danezis, G., & De Cristofaro, E. (2018, August 21). *LOGAN: Membership inference attacks against generative models* (4). arXiv: [1705.07663](https://arxiv.org/abs/1705.07663) [cs.CR]. <https://doi.org/10.48550/arXiv.1705.07663>
- Health Insurance Portability and Accountability Act of 1996, Public Law 104–191, 104th Congress (1996). Retrieved July 4, 2023, from <https://www.govinfo.gov/content/pkg/PLAW-104publ191/pdf/PLAW-104publ191.pdf>
- Hoffman, M. D., & Gelman, A. (2011). *The No-U-Turn sampler: Adaptively setting path lengths in hamiltonian monte carlo* (1). arXiv: [1111.4246](https://arxiv.org/abs/1111.4246) [stat.CO].
- Hollis, J. L., Crozier, S. R., Inskip, H. M., Cooper, C., Godfrey, K. M., Harvey, N. C., Collins, C. E., & Robinson, S. M. (2017). Modifiable risk factors of maternal postpartum weight retention: An analysis of their combined impact and potential opportunities for prevention. *International Journal of Obesity*, 41(7), 1091–1098. <https://doi.org/10.1038/ijo.2017.78>
- Homer, N., Szelinger, S., Redman, M., Duggan, D., Tembe, W., Muehling, J., Pearson, J. V., Stephan, D. A., Nelson, S. F., & Craig, D. W. (2008). Resolving individuals contributing trace amounts of dna to highly complex mixtures using high-density snp genotyping microarrays. *PLOS Genetics*, 4(8), 1–9. <https://doi.org/10.1371/journal.pgen.1000167>

- Hornby, R., & Hu, J. (2021). *Bayesian estimation of attribute disclosure risks in synthetic data with the AttributeRiskCalculation R package* (1). arXiv: [2103.09805](https://arxiv.org/abs/2103.09805) [stat.ME].
- Hu, J., Akande, O., & Wang, Q. (2021). *Multiple imputation and synthetic data generation with the R package NPBayesImputeCat* (3). arXiv: [2007.06101](https://arxiv.org/abs/2007.06101) [stat.CO].
- Hu, J., & Hoshino, N. (2018). The quasi-multinomial synthesizer for categorical data: UNESCO chair in data privacy international conference, PSD 2018 (J. Domingo-Ferrer & F. Montes, Eds.), 75–91.
- Hu, J., Reiter, J. P., & Wang, Q. (2014). Disclosure risk evaluation for fully synthetic categorical data: UNESCO chair in data privacy international conference, PSD 2014 (J. Domingo-Ferrer, Ed.), 185–199.
- Hu, J., Reiter, J. P., & Wang, Q. (2016, October 28). *Dirichlet process mixture models for modeling and generating synthetic versions of nested categorical data* (6). arXiv: [1412.2282](https://arxiv.org/abs/1412.2282) [stat.ME]. <https://doi.org/10.48550/arxiv.1412.2282>
- Hu, J., & Savitsky, T. D. (2021, February 2). *Bayesian data synthesis and disclosure risk quantification: An application to the consumer expenditure surveys* (2). arXiv: [1809.10074](https://arxiv.org/abs/1809.10074) [stat.AP]. <https://doi.org/10.48550/ARXIV.1809.10074>
- Hundepool, A., de Wolf, P.-P., Bakker, J., Reedijk, A., Franconi, L., Poletini, S., Capobianchi, A., & Domingo-Ferrer, J. (2014). *μ -Argus User's Manual*. Version 5.1. Retrieved July 4, 2023, from <https://research.cbs.nl/casc/Software/MUmanual5.1.3.pdf>
- Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Schulte Nordholt, E., Spicer, K., & de Wolf, P.-P. (2012). *Statistical disclosure control* (1st ed.). John Wiley & Sons, Ltd.
- Hunt, N. (2010). *Netflix prize update*. Retrieved March 14, 2010, from <https://web.archive.org/web/20100402143433/http://blog.netflix.com/2010/03/this-is-neil-hunt-chief-product-officer.html>
- Ichim, D. (2009). Disclosure control of business microdata: A density-based approach. *International Statistical Review*, 77(2), 196–211. <https://doi.org/10.1111/j.1751-5823.2009.00079.x>
- Imdb. the internet movie database*. (2023). Retrieved December 30, 2023, from <http://www.imdb.com/>
- Information Commissioner's Office. (2012). *Anonymisation: Managing data protection risk code of practice* (tech. rep.). <https://ico.org.uk/media/for-organisations/documents/1061/anonymisation-code.pdf>
- Information Commissioner's Office. (2018). *Guide to the general data protection regulation (gdpr)*. Information Commissioner's Office. <https://ico.org.uk/for-organisations/guide-to-the-general-data-protection-regulation-gdpr/>

- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning: With applications in R* (2nd ed.). Springer New York.
- Jordon, J., Yoon, J., & van der Schaar, M. (2019). PATE-GAN: Generating synthetic data with differential privacy guarantees. *International Conference on Learning Representations*.
- Karr, A. F., Kohonen, C. N., Oganian, A., Reiter, J. P., & Sanil, A. P. (2006). A framework for evaluating the utility of data altered to protect confidentiality. *The American Statistician*, 60(3), 224–232.
- Kingma, D. P., & Welling, M. (2014, May 1). *Auto-encoding variational Bayes* (10). arXiv: [1312.6114](https://arxiv.org/abs/1312.6114) [stat.ML].
- Kinney, S. K., Reiter, J. P., & Miranda, J. (2014). *Improving the synthetic longitudinal business database* (tech. rep. CES-14-12). U.S. Census Bureau, Center for Economic Studies.
- Kinney, S. K., Reiter, J. P., Reznick, A. P., Miranda, J., Jarmin, R. S., & Abowd, J. M. (2011). Towards unrestricted public use business microdata: The synthetic longitudinal business database. *International Statistical Review*, 79(3), 362–384. <https://doi.org/10.1111/j.1751-5823.2011.00153.x>
- Kleinke, K., & Reinecke, J. (2013). Multiple imputation of incomplete zero-inflated count data. *Statistica Neerlandica*, 67(3), 311–336. <https://doi.org/10.1111/stan.12009>
- Lewandowski, D., Kurowicka, D., & Joe, H. (2009). Generating random correlation matrices based on vines and extended onion method. *Journal of Multivariate Analysis*, 100(9), 1989–2001. <https://doi.org/10.1016/j.jmva.2009.04.008>
- Li, N., Li, T., & Venkatasubramanian, S. (2007). t -closeness: Privacy beyond k -anonymity and ℓ -diversity. *IEEE 23rd International Conference on Data Engineering*, 106–115.
- Li, N., Qardaji, W., Su, D., Wu, Y., & Yang, W. (2013). Membership privacy: A unifying framework for privacy definitions. *Proceedings of the 2013 ACM SIGSAC Conference on Computer & Communications Security*, 889–900. <https://doi.org/10.1145/2508859.2516686>
- Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R News*, 2(3), 18–22. <https://CRAN.R-project.org/doc/Rnews/>
- Lin, Z., Jain, A., Wang, C., Fanti, G., & Sekar, V. (2019, September 30). *Using GANs for sharing networked time series data* (5). arXiv: [1909.13403](https://arxiv.org/abs/1909.13403) [cs.LG]. <https://doi.org/10.48550/arXiv.1909.13403>
- Little, R. J. A. (1993). Statistical analysis of masked data. *Journal of Official Statistics*, 9(2), 407–426. <http://www.jos.nu/articles/abstract.asp?article=92407>

- Little, R. J. A., Liu, F., & Raghunathan, T. E. (2004). Statistical disclosure techniques based on multiple imputation. In *Applied Bayesian modeling and causal inference from incomplete-data perspectives* (pp. 141–152). John Wiley & Sons, Ltd.
- Longhurst, J., & Vickers, P. (2007–November 7). Microdata risk assessment in an NSI context. Retrieved December 27, 2023, from https://nces.ed.gov/FCSM/pdf/2007FCSM_Longhurst-IX-B.pdf
- Machanavajjhala, A., Kifer, D., Gehrke, J., & Venkatasubramanian, M. (2007). ℓ -diversity: Privacy beyond k -anonymity. *ACM Transactions on Knowledge Discovery from Data*, 1(1).
- Manrique-Vallier, D., & Hu, J. (2018). Bayesian non-parametric generation of fully synthetic multivariate categorical data in the presence of structural zeros. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 181(3), 635–647. <https://doi.org/10.1111/rssa.12352>
- Mateo-Sanz, J. M., Martínez-Ballesté, A., & Domingo-Ferrer, J. (2004–June 11). Fast generation of accurate synthetic microdata: CASC project international workshop, PSD 2004 (J. Domingo-Ferrer & V. Torra, Eds.).
- Matthews, G. J., & Harel, O. (2011). Data confidentiality: A review of methods for statistical disclosure limitation and methods for assessing privacy. *Statistics Surveys*, 5, 1–29. <https://doi.org/10.1214/11-SS074>
- Matthews, G. J., Harel, O., & Aseltine, R. H. (2009). Examining the robustness of fully synthetic data techniques for data with binary variables. *Journal of Statistical Computation and Simulation*, 80(6), 609–624. <https://doi.org/10.1080/00949650902744438>
- Mironov, I. (2017, August 25). *Rényi differential privacy* (3). arXiv: 1702.07476 [cs.CR]. <https://doi.org/10.48550/arXiv.1702.07476>
- Mithal, M. (2010, March 12). *Closing letter to Reed Freeman, Esq., counsel for Netflix, Inc.* Retrieved December 30, 2023, from <https://www.ftc.gov/legal-library/browse/cases-proceedings/closing-letters/netflix-inc>
- Narayanan, A., & Shmatikov, V. (2007). *How to break anonymity of the Netflix prize dataset* (2). arXiv: 0610105 [cs.CR].
- National Health Service. (2022). *Gestational diabetes: Overview*. Retrieved December 31, 2023, from <https://www.nhs.uk/conditions/gestational-diabetes/>
- National Health Service. (2023). *Symptoms*. Retrieved December 31, 2023, from <https://www.nhs.uk/conditions/type-2-diabetes/symptoms/>
- Nazábal, A., Olmos, P. M., Ghahramani, Z., & Valera, I. (2020). Handling incomplete heterogeneous data using VAEs. *Pattern Recognition*, 107, 107537. <https://doi.org/10.1016/j.patcog.2020.107501>
- NHS Digital. (2019). *Statistics on obesity, physical activity and diet, England, Part 3: Adult overweight and obesity* [2019]. Retrieved February 18, 2023, from <https://digital.nhs.uk/>

nhs.uk/data-and-information/publications/statistical/statistics-on-obesity-physical-activity-and-diet/statistics-on-obesity-physical-activity-and-diet-england-2019/part-3-adult-obesity

Nowok, B., Raab, G., & Dibben, C. (2017). Providing bespoke synthetic data for the UK longitudinal studies and other sensitive data with the `synthpop` package for R. *Statistical Journal of the IAOS*, 33(3), 785–796. <https://doi.org/10.3233/SJI-150153>

Nowok, B., Raab, G. M., & Dibben, C. (2016). `synthpop`: Bespoke creation of synthetic data in R. *Journal of Statistical Software*, 74(11). <https://doi.org/10.18637/jss.v074.i11>

Office for Civil Rights. (2012, November 26). *Guidance regarding methods for de-identification of protected health information in accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule* (tech. rep.). U.S Department of Health & Human Services (HHS).

Office for National Statistics. (2022). *Maximising the quality of census 2021 population estimates*. Retrieved December 27, 2023, from <https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/methodologies/maximisingthequalityofcensus>

Office for Statistics Regulation. (2023). *2022 census in scotland*. Retrieved December 27, 2023, from <https://osr.statisticsauthority.gov.uk/publication/2022-census-in-scotland/pages/2/>

Office of the Assistant Secretary for Planning and Evaluation. (1999, November 3). *Standards for privacy of individually identifiable health information* (NPRM). Department of Health and Human Services. Retrieved July 4, 2023, from <https://aspe.hhs.gov/reports/nrpm-standards-privacy-individually-identifiable-health-information>

Oganian, A. (2014). *v*-dispersed synthetic data based on a mixture model with constraints: UNESCO chair in data privacy international conference, PSD 2014 (J. Domingo-Ferrer, Ed.), 200–212.

Oganian, A., & Domingo-Ferrer, J. (2017). Local synthesis for disclosure limitation that satisfies probabilistic *k*-anonymity criterion. *Transactions on Data Privacy*, 10(1), 61–81.

Palley, M. A., & Simonoff, J. S. (1987). The use of regression methodology for the compromise of confidential information in statistical databases. *ACM Transactions on Database Systems*, 12(4), 593–608. <https://doi.org/10.1145/32204.42174>

Piironen, J., & Vehtari, A. (2016). *On the hyperprior choice for the global shrinkage parameter in the horseshoe prior* (2). arXiv: 1610.05559 [stat.ME].

Piironen, J., & Vehtari, A. (2017a). Comparison of bayesian predictive methods for model selection. *Statistics and Computing*, 27(3), 711–735. <https://doi.org/10.1007/s11222-016-9649-y>

- Piironen, J., & Vehtari, A. (2017b). *Sparsity information and regularization in the horseshoe and other shrinkage priors* (1). arXiv: 1707.01694 [stat.ME].
- Pistner, M., Slavković, A., & Vilhuber, L. (2018). Synthetic data via quantile regression for heavy-tailed and heteroskedastic data: UNESCO chair in data privacy international conference, PSD 2018 (J. Domingo-Ferrer & F. Montes, Eds.), 92–108.
- Purdam, K., & Elliot, M. (2007). A case study of the impact of statistical disclosure control on data quality in the individual UK samples of anonymised records. *Environment and Planning A*, 39(5), 1101–1118. <https://doi.org/10.1068/a38335>
- R Core Team. (2023). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. <https://www.R-project.org/>
- Raab, G., Nowok, B., & Dibben, C. (2016, December 1). *A simplified approach to generating synthetic data for disclosure control* (6). arXiv: 1409.0217 [stat.ME]. <https://doi.org/10.48550/arXiv.1409.0217>
- Raab, G. M., Nowok, B., & Dibben, C. (2017). *Guidelines for producing useful synthetic data*. arXiv: 1712.04078 [stat.AP].
- Raghunathan, T., Solenberger, P., Berglund, P., & van Hoewyk, J. (2016). *IVEware: Imputation and variance estimation software*. Version 0.3. University of Michigan Institute for Social Research - Survey Research Center. Retrieved August 4, 2023, from <https://src.isr.umich.edu/wp-content/uploads/iveware-manual-Version-0.3.pdf>
- Raghunathan, T. E., Lepkowski, J. M., Van Hoewyk, J., & Solenberger, P. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology*, 27(1), 85–95.
- Raghunathan, T. E., Reiter, J. P., & Rubin, D. B. (2003). Multiple imputation for statistical disclosure limitation. *Journal of Official Statistics*, 19(1), 1–16. <https://www2.stat.duke.edu/courses/Spring06/sta395/raghunathan2003.pdf>
- Ray, S., Banerjee, A., Swift, A., Fanstone, J. W., Mamalakis, M., Vorselaars, B., Wilkie, C., Cole, J., Mackenzie, L. S., & Weeks, S. (2022). A robust COVID-19 mortality prediction calculator based on lymphocyte count, urea, C-reactive protein, age and sex (LUCAS) with chest X-rays. *Scientific Reports*, 12, 18220. <https://doi.org/10.1038/s41598-022-21803-2>
- Reiter, J. P. (2002). Satisfying disclosure restrictions with synthetic data sets. *Journal of Official Statistics*, 1–19. <http://www.stat.duke.edu/~jerry/Papers/jos02.pdf>
- Reiter, J. P. (2003). Inference for partially synthetic, public use microdata sets. *Survey Methodology*, 29(2), 181–188. <https://doi.org/10.2307/1504821>
- Reiter, J. P. (2004). Simultaneous use of multiple imputation for missing data and disclosure limitation. *Survey Methodology*, 30(2), 235–242.

- Reiter, J. P. (2005a). Releasing multiply imputed, synthetic public use microdata: An illustration and empirical study. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *168*(1), 185–205.
- Reiter, J. P. (2005b). Using CART to generate partially synthetic, public use microdata. *Journal of Official Statistics*, *21*(3), 441–462.
- Reiter, J. P. (2023). Synthetic data: A look back and a look forward. *Transactions on Data Privacy*, *16*(1), 15–24.
- Reiter, J. P., & Kinney, S. K. (2012). Inferentially valid, partially synthetic data: Generating from posterior predictive distributions not necessary. *Journal of Official Statistics*, *28*(4), 583–590.
- Reiter, J. P., & Mitra, R. (2009). Estimating risks of identification disclosure in partially synthetic data. *The Journal of Privacy and Confidentiality*, *1*(1), 99–110. <https://doi.org/10.1198/016214505000000619>
- Reiter, J. P., Wang, Q., & Zhang, B. E. (2014). Bayesian estimation of disclosure risks for multiply imputed, synthetic data. *Journal of Privacy and Confidentiality*, *6*(1), 17–33.
- Ripley, B. D. (1996). *Pattern recognition and neural networks* (8th ed.). Cambridge University Press.
- Rubin, D. B. (1987, June). *Multiple imputation for nonresponse in surveys* (D. B. Rubin, Ed.). John Wiley & Sons, Inc. <https://doi.org/10.1002/9780470316696>
- Rubin, D. B. (1993). Discussion: Statistical disclosure limitation. *Journal of Official Statistics*, *9*(2), 461–468. <http://www.jos.nu/Articles/article.asp>
- Rubin, D. B. (2004). The design of a general and flexible system for handling nonresponse in sample surveys. *The American Statistician*, *58*(4), 298–302. <https://doi.org/10.1198/000313004X6355>
- Ruppert, D., Wand, M. P., & Carroll, R. J. C. (2003). *Semiparametric regression*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511755453>
- Saito, M., Saito, S., Koyama, M., & Kobayashi, S. (2020). Train sparsely, generate densely: Memory-efficient unsupervised training of high-resolution temporal GAN. *International Journal of Computer Vision*, *128*, 2586–2606. <https://doi.org/10.1007/s11263-020-01333-y>
- Sakshaug, J. W., & Raghunathan, T. E. (2010). Synthetic data for small area estimation: UNESCO chair in data privacy international conference, PSD 2010 (J. Domingo-Ferrer & E. Magkos, Eds.).
- Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., & Chen, X. (2016, June 10). *Improved techniques for training GANs* (1). arXiv: [1606.03498](https://arxiv.org/abs/1606.03498) [cs.CV]. <https://doi.org/10.48550/arxiv.1606.03498>

- Sauer, A., Chitta, K., Müller, J., & Geiger, A. (2021, November 1). *Projected GANs converge faster* (1). arXiv: 2111.01007 [cs.CV]. <https://doi.org/10.48550/arxiv.2111.01007>
- Schafer, J. L. (1997). *The analysis of incomplete multivariate data*. CRC Press. <https://doi.org/10.1201/9781439821862>
- Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological methods*, 7.
- Shah, A. D., Bartlett, J. W., Carpenter, J., Nicholas, O., & Hemingway, H. (2014). Comparison of Random Forest and Parametric Imputation Models for Imputing Missing Data Using MICE: A Caliber Study. *American Journal of Epidemiology*, 179(6), 764–774. <https://doi.org/10.1093/aje/kwt312>
- Shokri, R., Stronati, M., Song, C., & Shmatikov, V. (2017, March 31). *Membership inference attacks against machine learning models* (2). arXiv: 1610.05820 [cs.CR]. <https://doi.org/10.48550/arxiv.1610.05820>
- Shumailov, I., Shumaylov, Z., Zhao, Y., Gal, Y., Papernot, N., & Anderson, R. (2023). *The curse of recursion: Training on generated data makes models forget* (2). arXiv: 2305.17493 [cs.LG].
- Si, Y., & Reiter, J. P. (2011). A comparison of posterior simulation and inference by combining rules for multiple imputation. *Journal of Statistical Theory and Practice*, 5(2), 335–347. <https://doi.org/10.1080/15598608.2011.10412032>
- Si, Y., & Reiter, J. P. (2013). Nonparametric Bayesian multiple imputation for incomplete categorical variables in large-scale assessment surveys. *Journal of Educational and Behavioral Statistics*, 38(5), 499–521. <https://doi.org/10.3102/1076998613480394>
- Smith, J. W., Everhart, J., Dickson, W., Knowler, W., & Johannes, R. (1988). Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. *Proceedings of the Annual Symposium on Computer Application in Medical Care*, 261–265.
- Snoke, J., Raab, G. M., Nowok, B., Dibben, C., & Slavkovic, A. (2018). General and specific utility measures for synthetic data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 181(3), 663–688. <https://doi.org/10.1111/rssa.12358>
- Stadler, T., Oprisanu, B., & Troncoso, C. (2022). Synthetic data – anonymisation groundhog day. *31st USENIX Security Symposium (USENIX Security 22)*, 1451–1468. <https://www.usenix.org/conference/usenixsecurity22/presentation/stadler>
- Stan Development Team. (2018). rstanarm: Bayesian applied regression modeling via stan. R package version 2.17.4. <http://mc-stan.org>
- Stan Development Team. (2024). RStan: The R interface to Stan [R package version 2.26.22]. <https://mc-stan.org/>
- Strack, B., Deshazo, J. P., Gennings, C., Olmo, J. L., Ventura, S., Cios, K. J., & Clore, J. N. (2014). Impact of HbA1c measurement on hospital readmission rates: Analysis

- of 70,000 clinical database patient records. *BioMed Research International*, 2014. <https://doi.org/10.1155/2014/781670>
- Sweeney, L. (1998a). Datafly: A system for providing anonymity in medical data. In T. Y. Lin & S. Qian (Eds.), *Database Security XI: Status and prospects* (pp. 356–381). Springer New York, NY.
- Sweeney, L. (1998b). Towards the optimal suppression of details when disclosing medical data, the use of sub-combination analysis. In B. Cesnik, A. McCray, & J.-R. Scherrer (Eds.), *MEDINFO '98: Proceedings of the Ninth World Congress on Medical Informatics*. Elsevier.
- Sweeney, L. (2002). k -anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(5), 557–570. <https://doi.org/10.1142/S0218488502001648>
- Sweeney, L., Abu, A., & Winn, J. (2013). *Identifying participants in the personal genome project by name (a re-identification experiment)* (1). arXiv: 1304.7605 [cs.CY].
- Sweeney, L., Yoo, J. S., Perovich, L., Boronow, K. E., Brown, P., & Brody, J. G. (2017). Reidentification risks in HIPAA safe harbor data: A study of data from one environmental health study. *Technology Science*.
- Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2015, December 11). *Rethinking the inception architecture for computer vision* (3). arXiv: 1512.00567 [cs.CV]. <https://doi.org/10.48550/arxiv.1512.00567>
- Taub, J., Elliot, M., Pampaka, M., & Smith, D. (2018). Differential correct attribution probability for synthetic data: An exploration: UNESCO chair in data privacy international conference, PSD 2018 (J. Domingo-Ferrer & F. Montes, Eds.), 122–137.
- Taub, J., Elliot, M., & Sakshaug, J. W. (2020). The impact of synthetic data generation on data utility with application to the 1991 UK samples of anonymised records. *Transactions on Data Privacy*, 13(1), 1–23. <http://www.tdp.cat/issues16/tdp.a306a18.pdf>
- Taylor, L., Zhou, X.-H., & Rise, P. (2018). A tutorial in assessing disclosure risk in microdata. *Statistics in medicine*, 37(25), 3693–3706. <https://doi.org/10.1002/sim.7667>
- Templ, M., & Alfons, A. (2010). Disclosure risk of synthetic population data with application in the case of EU-SILC: UNESCO chair in data privacy international conference, PSD 2010 (J. Domingo-Ferrer & E. Magkos, Eds.).
- The European Union. (2016). Regulation (EU) 2016/679 of the European Parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (text with EEA relevance). *Official Journal of the European Union*, 59(L119), 1–88.

- Torfi, A., & Fox, E. A. (2020, March 4). *CorGAN: Correlation-capturing convolutional generative adversarial networks for generating synthetic healthcare records* (2). arXiv: 2001.09346 [stat.ML].
- Upton, G., & Cook, I. (2014). Structural zero. <https://www.oxfordreference.com/view/10.1093/acref/9780199679188.001.0001/acref-9780199679188-e-1584>
- U.S. Census Bureau. (2013). *Synthetic longitudinal business database* (Version 2.0) [data set]. Ithaca, NY, Cornell University, Synthetic Data Server. Retrieved August 4, 2023, from <http://www2.ncrn.cornell.edu/ced2ar-web/codebooks/synlbd/v/v2>
- U.S. Census Bureau. (2018). *SIPP synthetic beta* (Version 7.0) [data set]. Ithaca, NY, Cornell University, Synthetic Data Server. Retrieved August 4, 2023, from <https://www2.ncrn.cornell.edu/ced2ar-web/codebooks/ssb/v/v7>
- U.S. Census Bureau. (2022). *Geography program: Glossary*. Retrieved December 19, 2023, from https://www.census.gov/programs-surveys/geography/about/glossary.html#par_textimage_13
- U.S. Census Bureau. (2023, January 30). *Synthetic longitudinal business database (SynLBD): Validating results*. Retrieved August 4, 2023, from <https://www.census.gov/programs-surveys/ces/data/public-use-data/synthetic-longitudinal-business-database/validating-results.html>
- van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, 45(3), 1–67. <https://www.jstatsoft.org/v45/i03/>
- Vehtari, A., Gabry, J., & Goodrich, B. (2022, February 21). Bayesian logistic regression with rstanarm. Retrieved May 9, 2022, from <https://avehtari.github.io/modelselection/diabetes.html>
- Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, 27(5), 1413–1432. <https://doi.org/10.1007/s11222-016-9696-4>
- Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S* (4th ed.) [ISBN 0-387-95457-0]. Springer. <http://www.stats.ox.ac.uk/pub/MASS4>
- Video Privacy Protection Act of 1988, Public Law 100–618, 100th Congress (1988). <https://www.govinfo.gov/app/details/STATUTE-102/STATUTE-102-Pg3195/summary>
- Wiese, M., Knobloch, R., Korn, R., & Kretschmer, P. (2019). Quant GANs: Deep generation of financial time series. *Quantitative Finance*, 20(9), 1419–1440. <https://doi.org/10.1080/14697688.2020.1730426>
- Willenborg, L., & de Waal, T. (2001). *Elements of statistical disclosure control* (P. Bickel, P. J. Diggle, S. Fienberg, K. Krickeberg, I. Olkin, & S. Zeger, Eds.). Springer New York. <https://doi.org/10.1007/978-1-4613-0121-9>

- Winkler, W. E. (2007). *Examples of easy-to-implement , widely used methods of masking for which analytic properties are not justified*. Statistical Research Division, U.S. Census Bureau.
- Woo, M.-J., Reiter, J. P., Oganian, A., & Karr, A. F. (2009). Global measures of data utility for microdata masked for disclosure limitation. *Journal of Privacy and Confidentiality*, 1(1). <https://doi.org/10.29012/jpc.v1i1.568>
- Wood, S. N. (2003). Thin plate regression splines. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 65(1), 95–114. <http://www.jstor.org/stable/3088828>
- Wood, S. N. (2006). *Generalized additive models: An introduction with R* (1st ed.). CRC Press.
- XGBoost developers. (2023). *XGBoost documentation: XGBoost parameters* (2.0.0). Retrieved May 2, 2024, from https://xgboost.readthedocs.io/en/release_2.0.0/parameter.html#parameters-for-tree-boost
- Xie, L., Lin, K., Wang, S., Wang, F., & Zhou, J. (2018, February 19). *Differentially private generative adversarial network* (1). arXiv: 1802.06739 [cs.LG].
- Yu, M., Reiter, J. P., Zhu, L., Liu, B., Cronin, K. A., & Feuer, E. J. (2017). Protecting confidentiality in cancer registry data with geographic identifiers. *American Journal of Epidemiology*, 186(1), 83–91. <https://doi.org/10.1093/aje/kwx050>
- Zeileis, A., Kleiber, C., Krämer, W., & Hornik, K. (2003). Testing and dating of structural changes in practice. *Computational Statistics & Data Analysis*, 44(1–2), 109–123. [https://doi.org/10.1016/S0167-9473\(03\)00030-6](https://doi.org/10.1016/S0167-9473(03)00030-6)
- Zeileis, A., Leisch, F., Hornik, K., & Kleiber, C. (2002). strucchange: An R package for testing for structural change in linear regression models. *Journal of Statistical Software*, 7(2), 1–38. <https://doi.org/10.18637/jss.v007.i02>
- Zhang, W., & Ray, S. (2023). From coarse to fine: A deep 3D probability volume contours framework for tumour segmentation and dose painting in PET images. *Frontiers in Radiology*, 3, 1225215. <https://doi.org/10.3389/fradi.2023.1225215>
- Zhang, X., Ji, S., & Wang, T. (2018, March 25). *Differentially private releasing via deep generative model (technical report)* (2). arXiv: 1801.01594 [cs.CR].
- Zhou, X., & Reiter, J. P. (2010). A note on Bayesian inference after multiple imputation. *American Statistician*, 64(2), 159–163. <https://doi.org/10.1198/tast.2010.09109>
- Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 67(2), 301–320. <http://www.jstor.org/stable/3647580>