Methods for Risk Evaluation part 3

Jingchen (Monika) Hu

Vassar College

Statistical Data Privacy

Outline

Outline

Recap

- Lectures 8 & 9:
  - Identification disclosure
    - ★ Matching based method
    - ★ Record linkage method

Plan for this lecture

- Two general types of disclosure: identification and attribute (Lecture 1)

1. Identification disclosure: The intruder correctly identifies records of interest in the released synthetic data
2. Attribute disclosure: The intruder correctly infers the true confidential values of the synthetic records using information from the released synthetic data

- In this lecture, we focus on attribute risk evaluation methods, with illustrations to the synthetic CE from Lectures 4 & 5 and the synthetic ACS from Lecture 5

Overview

- Attribute disclosure:
  - The intruder correctly infers the true, confidential values of the synthesized variables in the released synthetic data
  - Exist in partially synthetic data and fully synthetic data
- We will introduce two general approaches
  - The CAP risk measure
  - Classification-based risk measure

Overview

- In each method, it is assumed that an intruder knows certain characteristics for the targeted individual (one or multiple individuals), which we call the **key** variables
- It is further assumed that the intruder wishes to use the synthetic datasets to infer other characteristics for that individual, which we call the **target** variable(s)
- The approaches however differ with regards to assumptions about the process used by the intruder to infer the target variable(s), and the other information available

Outline

Background

- The correct attribution probability (CAP) measure, proposed by Elliot (2014) and Taub et al. (2018)
- The CAP measures the probability that an intruder can correctly predict the value of the target variable for an individual, by using the empirical distribution of this variable among synthetic observations with the same key variables
- For simplicity, our presentation focuses on categorical data, but as discussed in Elliot (2014) the idea could be extended to continuous variables by deciding on a threshold for when two different numerical values should be considered matches

Notation

- We present the measure following closely the notation and description of Baillargeon and Charest (2020)
- Let $\mathbf{Y}$ denote the confidential dataset of n records and p variables, and y_{ij} represents the j-th variable of the i-th record
- Consider a specific sensitive variable, l, that is synthesized for privacy protection
- Then, all possible values for this variable are the **targets** and denoted as $T_1, \cdots, T_G$

Notation

- Now the intruder attempts to predict the value of y_{il} using some or all of $\mathbf{Y}^{-l}$, the set of variables other than l in $\mathbf{Y}$
- We refer to the possible combination of these variables as **keys**, denoted as $K_1, \cdots, K_H$
- Each record $\mathbf{y}_i$ in the confidential dataset is thus associated with a single key $K(\mathbf{y}_i)$ and a single target $T(\mathbf{y}_i)$

Notation

- Now consider a synthetic dataset $\mathbf{Z}$ with the same keys and targets as the confidential dataset $\mathbf{Y}$
- Let n denote the number of records and r the number of variables in $\mathbf{Z}$
- Note that $\mathbf{Z}$ can be fully synthetic or partially synthetic
- Here we consider a single synthetic dataset $\mathbf{Z}$
- When $m > 1$ synthetic datasets are simulated, one would evaluate the CAP on each synthetic dataset and summarize the CAP across multiple synthetic datasets

Individual CAP

The correct attribution probability (CAP) of record $\mathbf{y}_0$ in confidential dataset $\mathbf{Y}$ with synthetic dataset $\mathbf{Z}$ is given as:

$$CAP_{\mathbf{y}_0}(\mathbf{Z}) = \frac{\sum_{i=1}^{n} I[T(\mathbf{z}_i) = T(\mathbf{y}_0), K(\mathbf{z}_i) = K(\mathbf{y}_0)]}{\sum_{i=1}^{n} I[K(\mathbf{z}_i) = K(\mathbf{y}_0)]}, \tag{1}$$

if $\sum_{i=1}^{n} I[K(\mathbf{z}_i) = K(\mathbf{y}_0)] \neq 0$ and 0 otherwise.

Discussion question: What does $CAP_{\mathbf{y}_0}(\mathbf{Z})$ represent?

Average CAP version 1

The average CAP of confidential dataset $\mathbf{Y}$ with synthetic dataset $\mathbf{Z}$ is given by:

$$\overline{CAP}_{\mathbf{Y}}(\mathbf{Z}) = \frac{1}{n}\sum_{i=1}^{n} CAP_{\mathbf{y}_i}(\mathbf{Z}), \tag{2}$$

where $CAP_{\mathbf{y}_i}(\mathbf{Z})$ is the individual CAP for record $\mathbf{y}_i$

Average CAP version 2

- Recall that any observation, say $\mathbf{y}_i$, for which the key $K(\mathbf{y}_i)$ does not appear in the synthetic dataset was assigned a CAP value of 0
- An alternate version of the average CAP, also suggested in Taub et al. (2018), removes these records from the average, so that the divider in the definition on previous page becomes $1/(n - |A|)$, where A is the set records for which the key does not appear in $\mathbf{Z}$ and $|A|$ is its cardinality

Comments on average CAP

- The average CAP can be interpreted as the proportion of correct predictions for an intruder who is predicting the target variable of each observation in the confidential dataset by sampling from the empirical distribution of the target variable in the synthetic data, conditional on the value of the key variable

Procedure for calculating average CAP

1. For each confidential record $\mathbf{y}_0$ in the confidential dataset $\mathbf{Y}$, find the collection of records in the synthetic dataset $\mathbf{Z}$ whose key is the same as $\mathbf{y}_0$. Among these records in the collection, find the subset of records whose target is the same as $\mathbf{y}_0$. From these, calculate the average CAP for record $\mathbf{y}_0$
2. Repeat step 1 for every confidential record $\mathbf{y}_i$ $(i = 1, \cdots, n)$ in the confidential dataset $\mathbf{Y}$.
3. Calculate the average CAP for $\mathbf{Y}$ as the average of the individual CAPs

Example of the ACS sample

We use the CAP measure to evaluate the attribute disclosure risks of a synthetic ACS sample created in Lecture 5, where `DIS` and `HICOV` are synthesized and the other variables remain unsynthesized. The synthesis model is the DPMPM model with the NPBayesImputeCat R package. We assume that the intruder knows the values of `SEX`, `RACE` and `MAR` of the record of interest, and want to predict the disability status `DIS` for this individual, using the synthetic dataset.

Example of the ACS sample

- Load datasets

```
ACSdata <- data.frame(readr::read_csv(file = "ACSdata.csv"))
n <- dim(ACSdata)[1]
ACSdata_syn <- data.frame(readr::read_csv(file = "ACSdata_syn.csv"))
## make sure variables are in the same ordering
ACSdata_syn <- ACSdata_syn[, names(ACSdata)]
ACSdata_con <- ACSdata
```

Example of the ACS sample: individual CAP

```
CalculateIndividualCAP <- function(condata_i, syndata,
                                   key.vars, target.vars){
  condata_i <- condata_i
  syndata <- syndata
  m <- nrow(syndata)

  match_key <- eval(parse(text=paste("condata_i$",key.vars,
                                     "==syndata$",
                                     key.vars,sep="",
                                     collapse="&")))
  match_key_target <- (eval(parse(text=paste("condata_i$",key.vars,
                                             "==syndata$",key.vars,
                                             sep="",collapse="&")))&
                         eval(parse(text=paste("condata_i$",target.vars,
                                               "==syndata$",
                                               target.vars,sep="",
                                               collapse="&"))))
```

Example of the ACS sample: individual CAP

```
  if (sum(match_key) > 0)
    CAP_i <- sum(match_key_target) / sum(match_key)
  else
    CAP_i <- 0

  return(CAP_i)
}
```

Example of the ACS sample: individual CAP

- A loop to go through the dataset

```
n <- nrow(ACSdata_con)
CAP_syn_all <- rep(NA, n)
for (i in 1:n){
  CAP_syn_all[i] <- CalculateIndividualCAP(condata_i = ACSdata_con[i, ],
                                           syndata = ACSdata_syn,
                                           key.vars = c("SEX",
                                                        "RACE",
                                                        "MAR"),
                                           target.vars = c("DIS"))
}
```

Example of the ACS sample: average CAP

- Calculate the average CAP

```
mean(CAP_syn_all)
```

```
## [1] 0.7228838
```

Example of the ACS sample: average CAP of the confidential data

- Results on the confidential data
  - Average CAP is 0.7224124
- See hidden code for the calculation

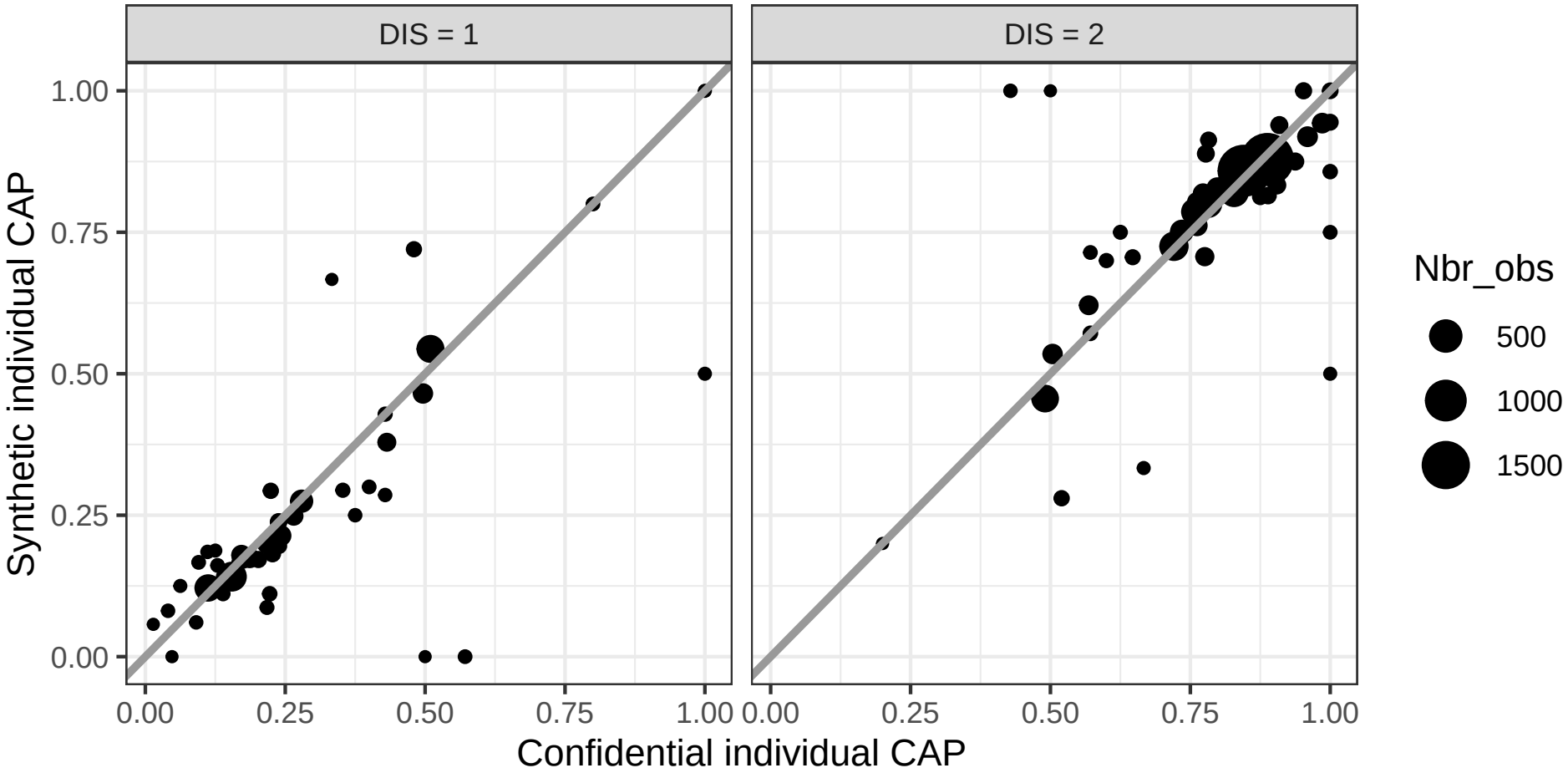
Discussion question: What is your finding?

```
## [1] 0.7224124
```

Example of the ACS sample: record-level individual CAP

- It is also useful here to consider the individual CAP values, because some observations have higher risk than others
- To better understand the risk for individual observations, we create a scatterplot comparing the individual CAP for each record before and after the synthesis process

Example of the ACS sample: record-level individual CAP



Discussion question: What does the plot tell you? What do you think of the CAP statistic?

Outline

Background and notation

- As indicated earlier, a weakness of the CAP statistic is that it uses a very simple model to predict the values of the target variable
- To tackle this challenge, researchers have suggested evaluating the attribute disclosure risk using more general classification approaches (Choi et al. (2017), Kaur et al. (2021))
- Any classifier can be used for this purpose, and the general procedure is as follows
- We use the same notation of confidential dataset $\mathbf{Y}$, synthetic dataset $\mathbf{Z}$, among other things
- Moreover, we deal with a single synthetic dataset $\mathbf{Z}$ here for illustration purpose

Discussion question: What kind of classification algorithm you have heard about and used in the past?

Procedure

1. Using only the synthetic dataset $\mathbf{Z}$, obtain a classifier to predict the value of the target variable given the values of the keys
2. For each confidential record $\mathbf{y}_0$ in the confidential dataset $\mathbf{Y}$, use the classifier obtained in Step 1 to predict the values of the target variable
3. Calculate the average attribute disclosure risk as the proportion of observations for which the prediction in Step 2 is correct. One could also modify this to penalize more for certain kinds of errors

Example of the CE sample

We evaluate the attribute disclosure risks of a synthetic CE sample in Lecture 4, where `Expenditure` is synthesized using a Bayesian simple linear regression synthesizer with `Income` as a predictor. Assume the intruder has access to the synthetic dataset, and the value of `Expenditure` for all records.

Example of the CE sample

- Loading datasets

```
CEdata <- data.frame(readr::read_csv(file = "CEdata.csv"))
n <- dim(CEdata)[1]
CEdata_syn <- data.frame(readr::read_csv(file = "CEdata_syn_SLR.csv"))
```

Example of the CE sample: use the k-NN classifier

- W will use the simple k-NN classifier, with the default value of $k = 3$, i.e., we consider 3 nearest neighbors for each record
- We refer the reader to James and R. (2021) for details about this classifier

Example of the CE sample: use the k-NN classifier

- We first make predictions for the target variable using the synthetic dataset

```
library(FNN)
library(ggplot2)
CE_kNN_pred_syn <- FNN::knn.reg(train = CEdata_syn[, c("UrbanRural",
                                                       "Income",
                                                       "Race",
                                                       "KidsCount")],
                                test =  CEdata[, c("UrbanRural",
                                                   "Income",
                                                   "Race",
                                                   "KidsCount")],
                                y = CEdata_syn[, c("Expenditure")])
```

Example of the CE sample: use the k-NN classifier

- To enable before and after synthesis comparison of attribute disclosure risk at the record-level, we also make predictions using the confidential dataset

```
CE_kNN_pred_con <- FNN::knn.reg(train = CEdata[, c("UrbanRural",
                                                   "Income",
                                                   "Race",
                                                   "KidsCount")],
                                test =  CEdata[, c("UrbanRural",
                                                   "Income",
                                                   "Race",
                                                   "KidsCount")],
                                y = CEdata[, c("Expenditure")])
```

Example of the CE sample: compare file-level MSE

- We can first compare the mean-square error (MSE) of the predictions from the k-NN classifier trained on the synthetic and the confidential datasets

```
CE_kNN_MSE_syn <- mean((CEdata$Expenditure -
                          CE_kNN_pred_syn$pred)^2)
CE_kNN_MSE_syn
```

```
## [1] 147374578
```

```
CE_kNN_MSE_con <- mean( (CEdata$Expenditure -
                           CE_kNN_pred_con$pred)^2)
CE_kNN_MSE_con
```

```
## [1] 80887543
```

Discussion question: What does the result tell us?

Example of the CE sample: compare record-level prediction

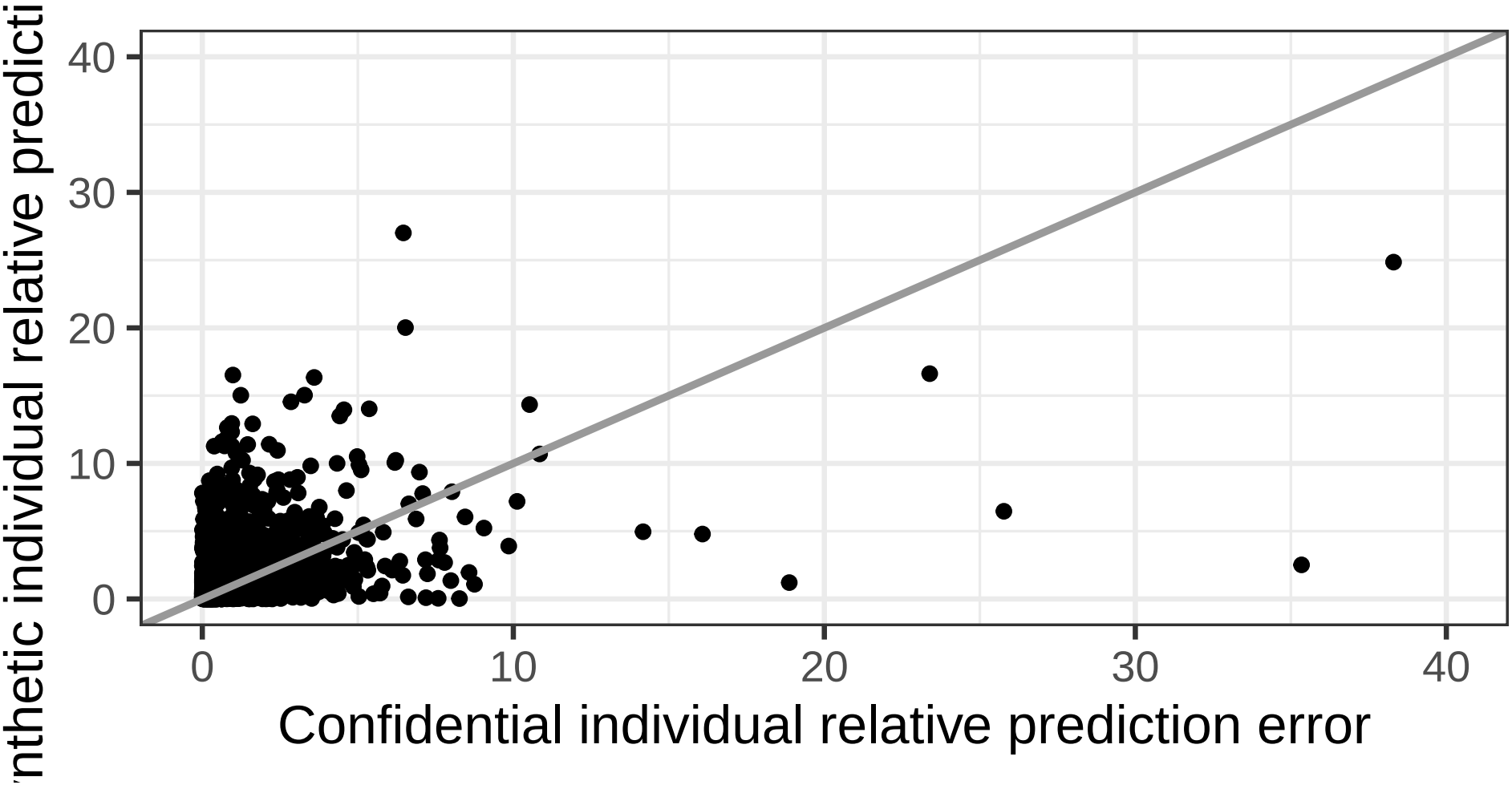
- We can also look at the individual prediction errors.
- Specifically, we calculate the record-level relative absolute prediction errors in the confidential and the synthetic datasets
- We then return the proportion of observations will have a less accurate prediction with the synthetic data compared to with the confidential data

```
relative_error_syn <- (abs(CEdata$Expenditure -
                           CE_kNN_pred_syn$pred)/CEdata$Expenditure)
relative_error_con <- (abs(CEdata$Expenditure -
                           CE_kNN_pred_con$pred)/CEdata$Expenditure)
mean(relative_error_con > relative_error_syn)
```

```
## [1] 0.3389831
```

Discussion question: What does the result tell us?

Example of the CE sample: compare record-level prediction



Discussion question: What does the plot tell us?

Outline

Summary

- Attribute disclosure risk evaluations
  - the CAP risk measure
  - classification-based risk measure

Summary

- Attribute disclosure risk evaluations
  - the CAP risk measure
  - classification-based risk measure
- No homework! But you should be working on disclosure risk evaluation for your project and be preared to present next week
- Lecture 11: Overview of differential privacy 1 - basics and mechanisms

References I

Baillargeon, M., and A. Charest. 2020. "A Closer Look at the CAP Risk Measure for Synthetic Datasets." Privacy in Statistical Databases (E-Proceedings).

Choi, Edward, Siddharth Biswal, Bradley Malin, Jon Duke, Walter F. Stewart, and Jimeng Sun. 2017. "Generating Multi-Label Discrete Patient Records Using Generative Adversarial Networks." In Proceedings of the 2nd Machine Learning for Healthcare Conference, edited by Finale Doshi-Velez, Jim Fackler, David Kale, Rajesh Ranganath, Byron Wallace, and Jenna Wiens, 68:286–305. Proceedings of Machine Learning Research. Boston, Massachusetts: PMLR.

Elliot, M. 2014. "Final Report on the Disclosure Risk Associated with the Synthetic Data Produced by the SYLLS Team." CMIST.

James, Witten, G., and Tibshirani R. 2021. An Introduction to Statistical Learning with Applications in R, Second Edition. Springer.

References II

Kaur, D., M. Sobiesk, S. Patil, J. Liu, P. Bhagat, A. Gupta, and N. Markuzon. 2021. "Application of Bayesian Networks to Generate Synthetic Health Data." Journal of the American Medical Informatics Association 28 (4): 801–11.

Taub, J., M. Elliot, M. Pampaka, and D. Smith. 2018. "Differential Correct Attribution Probability for Synthetic Data: An Exploration." Privacy in Statistical Databases, 122–37.