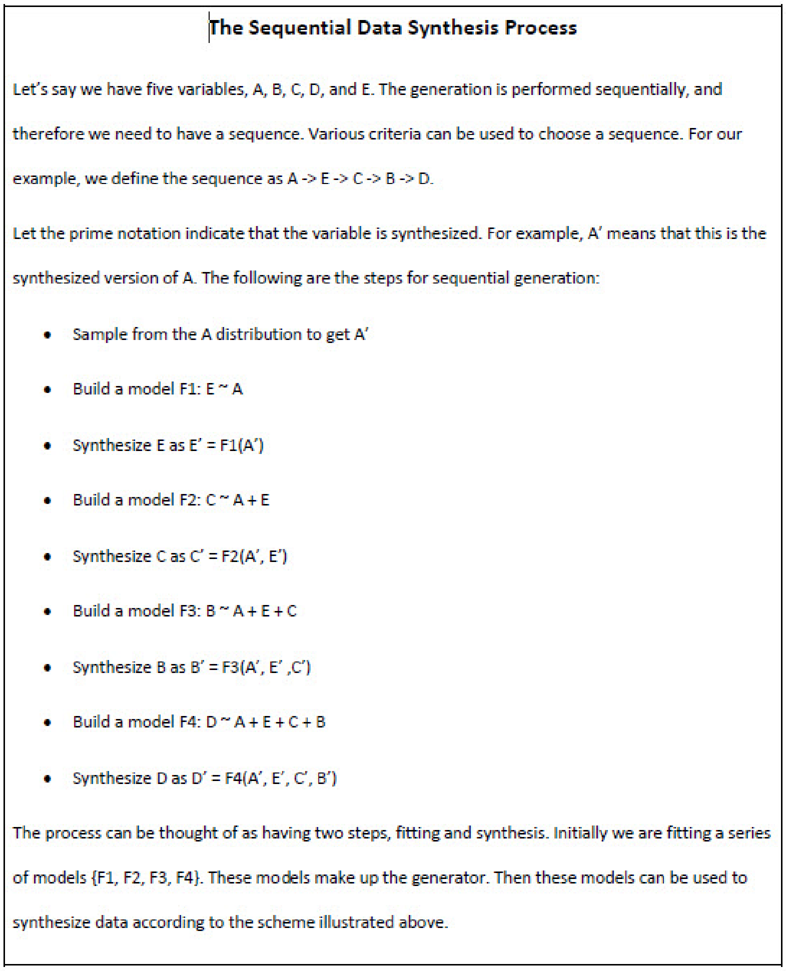
# Appendix A: Additional Details on Synthesis Method

## A. Sequential Synthesis Method

A description of the general sequential synthesis algorithm is provided in Figure 1 [1].



**The Sequential Data Synthesis Process**

Let's say we have five variables, A, B, C, D, and E. The generation is performed sequentially, and therefore we need to have a sequence. Various criteria can be used to choose a sequence. For our example, we define the sequence as A -> E -> C -> B -> D.

Let the prime notation indicate that the variable is synthesized. For example, A' means that this is the synthesized version of A. The following are the steps for sequential generation:

- Sample from the A distribution to get A'
- Build a model F1: E ~ A
- Synthesize E as E' = F1(A')
- Build a model F2: C ~ A + E
- Synthesize C as C' = F2(A', E')
- Build a model F3: B ~ A + E + C
- Synthesize B as B' = F3(A', E' ,C')
- Build a model F4: D ~ A + E + C + B
- Synthesize D as D' = F4(A', E', C', B')

The process can be thought of as having two steps, fitting and synthesis. Initially we are fitting a series of models {F1, F2, F3, F4}. These models make up the generator. Then these models can be used to synthesize data according to the scheme illustrated above.

**Figure 1:** A description of the sequential data synthesis process using classification and regression trees. Although any set of classification and regression methods can be used.

In this study partial synthesis was performed on the trial dataset. The partial synthesis ensured that potentially identifying information in the dataset (called the quasi-identifiers [2]) were synthesized. Quasi-identifiers are the variables that are knowable by an adversary and can be used for re-identification attacks [2]–[4]. Such information is knowable because it is in the public domain (e.g., in obituaries or registries, such as voter registration lists), or is known by an adversary who is an acquaintance of someone in the dataset (e.g., a neighbor or relative). Only synthesizing the quasi-identifiers to protect against identification risks is consistent with the clinical trial data anonymization guidelines from the European Medicines Agency [5] and Health Canada [6].

The process for partial synthesis is illustrated in Figure 2. In this example we have four variables overall and two of them are quasi-identifiers (Q1 and Q2). During the fitting step two models are built in sequence for each of the quasi-identifiers. Then the non-quasi-identifiers are used as inputs during the synthesis process to generate the synthetic quasi-identifiers. This process ensures that relationships between the synthetic quasi-identifiers and the other non-synthesized (input) variables are maintained. The input data into this process is tabular, and so is the synthetic version that is produced.

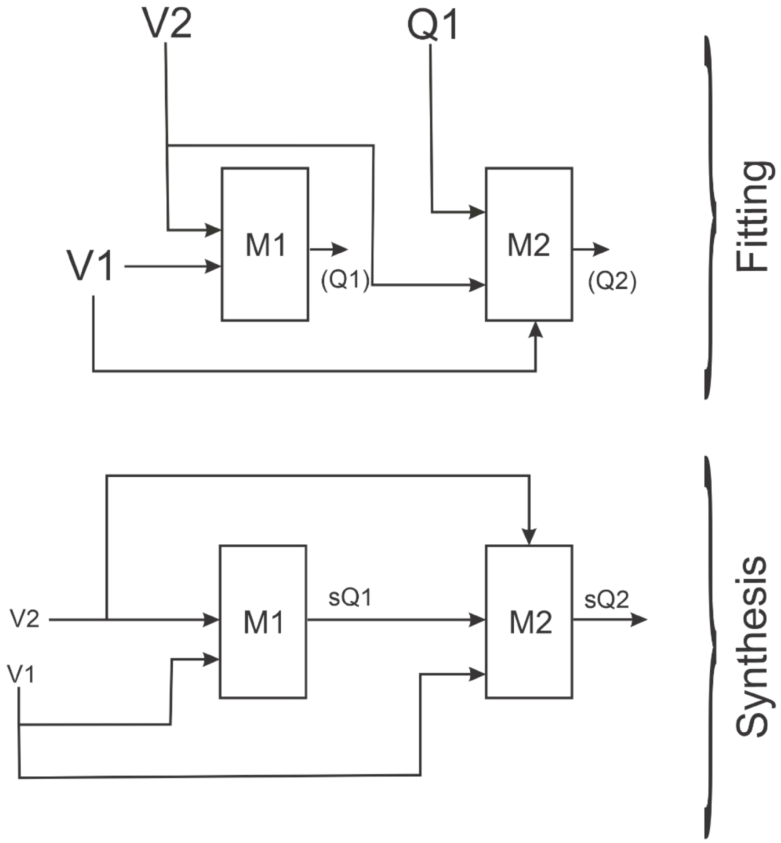


**Figure 2:** A schematic diagram illustrating the partial synthesis process where only the quasi-identifiers are synthesized. The Q1 and Q2 variables are the original quasi-identifier values, and sQ1 and sQ2 are the synthesized quasi-identifier values.

# Appendix B: Definition of Quasi-identifiers for Clinical Trial Data

## B. Introduction

The conclusion of this analysis is that a certain type of quasi-identifier contributes the most to the risk of identifying individuals: the public quasi-identifiers. Therefore, the public quasi-identifiers should be synthesized.

Public quasi-identifiers are the information that would be publicly known by an adversary about individuals. This information includes the demographics and socio-economic status variables. One exception to that rule is death information due to a serious adverse event (SAE). Data protection methods when applied to clinical trial datasets should not perturb SAE counts as that can affect the interpretation of the results.

In the following we will use the term “depersonalized” to apply to a dataset that has been transformed to protect patient privacy using any of a number of privacy enhancing techniques, such as data synthesis.

## B.1 Definitions

To start off with we will provide some definitions.

### B.1.1 Population, Real and Depersonalized Samples

To make this analysis more general, we will refer to the original dataset as the “Original Sample” and the depersonalized version of that dataset as the “Depersonalized Sample”. The depersonalized sample can be created using data synthesis, for example.

There is also the concept of the population. The Original Sample is assumed to be drawn from a population. We will discuss how the population is defined further below. The relationships between these datasets are illustrated in Figure 3.

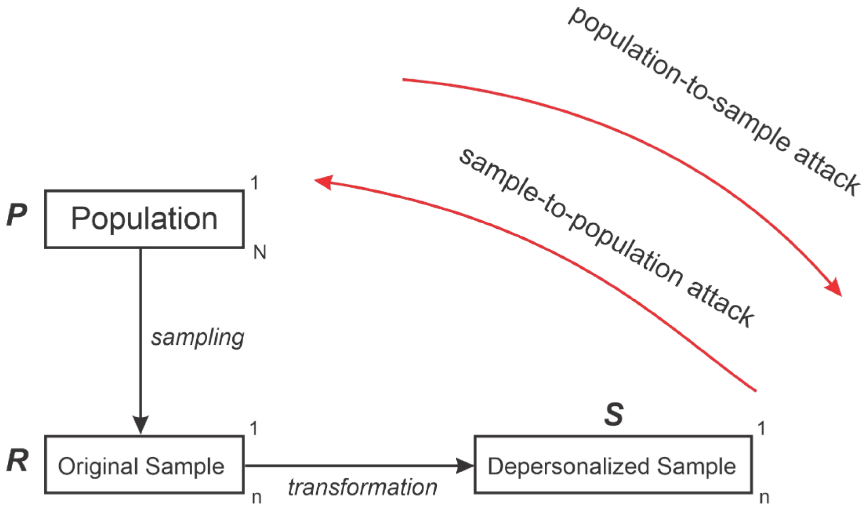


**Figure 3:** Directions of attack. The population has N records and the samples have n records.

### B.1.2 Directions of Attack

When an adversary attacks a depersonalized sample, she can do so in one of two directions [7], [8]. The direction for these types of attacks is illustrated in Figure 3.

The adversary can start from an acquaintance. This may be, for example, a relative, a co-worker, or a neighbor. The adversary will have some background information about that acquaintance and then use that background information to find the record that matches. This is called a *population-to-sample attack*.

An adversary can also start from the depersonalized sample and select a record to match against someone in the population. The adversary will need a population registry to match against.[1] This would be an identified database[2] of people in the population that the same quasi-identifiers as in the depersonalized sample. Typically these registries exist, such as voter registration lists [9]. Or the adversary can try to construct one from, say, social media [10]. This is called a *sample-to-population attack*.

### B.1.3 Quasi-identifiers

The adversary matches an acquaintance with depersonalized sample records or matches a depersonalized sample record with population registry. In both cases the matches are performed using the quasi-identifiers. The quasi-identifiers are variables that are present in the depersonalized sample record and also either known to the adversary or that are in the population registry.

There are two types of quasi-identifiers: (a) acquaintance quasi-identifiers, and (b) public quasi-identifiers. Public quasi-identifiers are a subset of acquaintance quasi-identifiers.

[1] In practice, population registries may be incomplete (such as the voter registration list). However, for the purposes of our analysis we will make the conservative assumption that a complete population registry exists.

[2] An identified database is one that has identities of individuals in it, such as a name, address, and possibly SIN.

Acquaintance quasi-identifiers are those known about an acquaintance. An adversary can know many things about an acquaintance, including their medical history. Public quasi-identifiers are those typically included in a population registry. This is the kind of information that exists in voter registration lists and that *many* people share about themselves on public social media. It also includes what can be inferred from public information. For example, income can be inferred via a person's ZIP code, which is easily obtainable for most people. Public quasi-identifiers are typically demographic and socio-economic information, as well as major events (e.g., births and deaths).

A population-to-sample attack is potentially more potent because the adversary will have more quasi-identifiers to work with. However, sampling can be protective in this case because the attacker's acquaintance may or may not be in the depersonalized sample.

## B.2 Risk Model for the Identification of Clinical Trial Participants

In this section we will formulate the basic risk model specific to clinical trials and illustrate it under different assumptions. Our main conclusion is that the baseline risk of identification of participants in clinical trials under the population-to-sample attack is lower than generally accepted thresholds. Therefore, the focus should be on managing sample-to-population attacks, where the relevant quasi-identifiers are public quasi-identifiers. These are the quasi-identifiers that we synthesized in our study.

Without loss of generality, in this analysis we focus on identification risk to participants in the US. One main reason is that there is much more data that can be used by us to conduct meaningful risk analyses from the US. The US population that we use is 330 million. We also assume the attacker performs exact matching when performing attacks in either direction.

### B.2.1 Evaluation of Population-to-sample Risk

We show that the population-to-sample risk is low under different assumptions about the adversary knowledge. We examine three assumptions: (a) the adversary knows that a target individual participated in an industry sponsored trial, but not knowing which one, (b) the adversary knows which specific trial the target individual participated in, and (c) the adversary knows the disease being studied for the trial that the target individual participated in, but not the specific trial or sponsor of the trial.

#### B.2.1.1 Background

An adversary needs to know the population from which the target individual was selected from. For example, if all that the adversary knows is that the target individual participated in a clinical trial then the relevant population is all individuals who participated in a trial. We will make a series of different assumptions about what the adversary knows and evaluate the risk of a successful population-to-sample attack. For our purposes we will always assume that the adversary knows that the target individual has participated in a clinical trial.

Based on a commonly used methodology for evaluating risk, there are three attacks that can occur on a dataset [7], [8]. The first is a deliberate attack on the clinical trial dataset by the adversary. The overall risk of matching the acquaintance with a depersonalized sample record under a deliberate attack model can be expressed as:

$$pr\left(b \mid attempt, a\right) \times pr\left(attempt \mid a\right) \times pr\left(a\right) \tag{1}$$

where $pr(a)$ is the probability that the adversary knows someone in the population (has an acquaintance), $pr(attempt \mid a)$ is the probability of the adversary attempting to identify a record in the dataset given that the adversary knows someone in the population, and $pr(b \mid attempt, a)$ is the probability of identification given that the adversary will attempt to identify a record and knows someone in the population. For our purposes we make the conservative assumption that the probability of an adversary attempting an attack on the data is one: $pr(attempt \mid a) = 1$. Therefore, we can simplify the model to:

$$pr(b \mid a) \times pr(a) \qquad (2)$$

The second attack is when an adversary inadvertently identifies a record while working with a dataset. Under worst case assumptions, that risk is the same as in equation (2).

The third attack is when a breach occurs. This is modeled as:

$$pr(b \mid breach, a) \times pr(breach \mid a) \times pr(a) \qquad (3)$$

The probability of a breach occurring will by definition be some number smaller than one.

The maximum of the probabilities from these three attacks is taken as the overall risk of identification [7], [8], which will be the same as equation (2). This equation gives us the probability of a successful population-to-sample attack.

This number needs to be smaller than the commonly used risk threshold of 0.09 by the European Medicines Agency (EMA) [5] and Health Canada [6] for a dataset to be considered to have a low risk of identification.

Furthermore, the value for $pr(a)$ can be expressed as [7]:

$$pr(a) = 1 - (1 - v)^{150} \qquad (4)$$

where $v$ is the prevalence in the population that we are looking at, and 150 is the Dunbar number (see the literature review in [7]), which is the average number of “friends” or acquaintances that a person has. For example, if the population is all individuals who participated in clinical trials in the US, then $v$ is the proportion of individuals in the US who have participated in clinical trials.

An additional parameter that we will need is the probability of identification under a population-to-sample attack [7] (also see the appendix of that reference for the derivation):

$$pr(b \mid a) = \frac{1}{N} \sum_{i=1}^{n} \frac{1}{f_i} \qquad (5)$$

where $f_i$ is the size of the equivalence class in the depersonalized sample that record $i$ is in. An equivalence class is the group of records in the depersonalized sample that have the same values on the quasi-identifiers. For example, if the quasi-identifiers are age and sex, then an equivalence class would be “50-year-old males”, and for a depersonalized sample record $i$ that has these values on age and sex, $f_i$ is the number of 50 year old males in the depersonalized sample.

We now need to define the population, which will depend on the assumption we make about what the adversary knows. We will consider examine three different assumptions.

#### B.2.1.2 Assumption 1: Industry Sponsored Trial

Here we focus on clinical trials that lead to approvals of FDA-regulated products, and therefore we limit our analysis to clinical trials sponsored by the life sciences industry. We do not consider investigator-initiated trials or those funded by governments and foundations under this assumption.

If we start off by asking what is the probability that an adversary would identify a target individual who has participated in an FDA-regulated trial, then we are interested in calculating the probability that an adversary would know someone who participated in an FDA-regulated trial which resulted in an approved product.

The 2019 snapshot from the FDA noted that there 46,391 individuals who participated in clinical trials submitted as part of approved New Molecular Entities (NMEs) and original biologics.[3] NMEs and original biologics are medications made of new molecular structures that have not been approved by the FDA before.

Only 40% of these patients were recruited in the US (see the snapshot report). These trials would have been ongoing for multiple years before approval.

Therefore, for this population we have $v=\frac{\left(46,391\times 0.4\right)}{330,000,000}=0.00005623$. If we use this in equation (4) , then the probability of an adversary knowing someone who has participated in one of these trials in the US is 0.008.

However, only considering NMEs ignores clinical trials for line extensions. It has been estimated that industry’s R&D direct costs for these post-NME approvals is one fourth than that of an NME [12]. If we extrapolate that to the number of participants, then we can recompute the risk value above as $v=\frac{\left(46,391\times 0.4\times 1.25\right)}{330,000,000}=0.00007028$. If we use this in equation (4), then the probability of an adversary knowing someone who has participated in one of these trials in the US is 0.01.

If we sum the FDA snapshot numbers from 2015 to 2019 inclusive, then the overall baseline risk is $pr\left(a\right)=0.0553$. This assumes that a dataset was for a trial which was part of an FDA approval over that five year period.

This means that if a life sciences company is sharing a clinical trial dataset for a trial that was submitted as part of an FDA approval from 2015-2019, the estimated probability of an adversary knowing a participant is 0.0553.

[3] See <https://www.fda.gov/drugs/drug-approvals-and-databases/drug-trials-snapshots>

This number is already lower than the commonly used risk threshold of 0.09. From equation (1), irrespective of the value of the $pr(b|a)$ in equation (5), the overall population-to-sample risk will already be below the threshold.

### B.2.1.3 Assumption 2: Specific Trial

The above is an industry-wide estimate. We can also compute the probability of successful identification for a specific trial as well since risk assessments are performed on a per trial basis. In this case we are assuming that the adversary knows that the target individual has participated in a specific trial.

If we have a study with 5,500 participants, the estimated probability of an adversary knowing someone who has participated in that trial (in the US) from equation (4) is 0.0025. This estimate is shown in Figure 4 for different trial sizes.

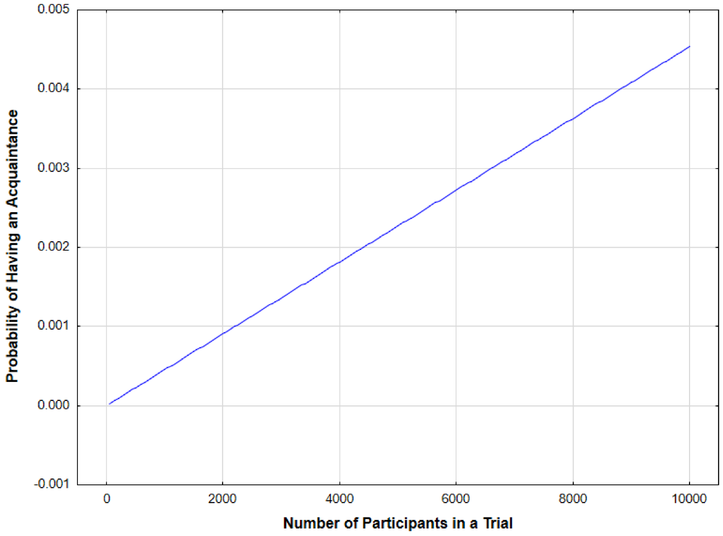


**Figure 4:** The probability of an adversary having an acquaintance in a specific trial (y-axis) against the size of the trial (x-axis).

For the trial sizes shown in Figure 4, the baseline population-to-sample risk will be below the commonly used threshold of 0.09.

### B.2.1.4 Assumption 3: Cancer Trial

We now consider a specific disease: cancer and assume that the adversary knows that the target is participating in an oncology trial. The prevalence of cancer in the US in January 2017 is 15,760,939 individuals[4], which is 4.7% of the population. It is estimated that less than 5% of cancer patients participate in cancer trials [13], which represents 0.2% or less of the US population.

Based on that prevalence, we then have $pr(a)=0.259$, which means that a randomly selected adversary has approximately a 25% chance of knowing someone who has participated in a cancer trial.

Let us consider $pr(b|a)$. Under the worst case assumption (i.e., highest identification risk) we have $f_i=1$, which gives us $pr(b|a)={}^{n}\!/_{N}$, which is the sampling fraction. If we assume a 10,000 participant trial, then the sampling fraction is: 0.012. The overall risk of identification is $0.259\times 0.012=0.0033$, which is quite a low population-to-sample risk. If the $f_i=1$ is not correct (i.e., that any $f_i>1$) then the population-to-sample risk would be even smaller.

Therefore, the baseline probability of an adversary knowing someone who has participated in an oncology trial and successfully identifying their record is quite small.

### B.2.2 Evaluation of Sample-to-Population Risk

In this type of risk assessment, the adversary is matching a record in the clinical trial dataset with a population registry. The kinds of quasi-identifiers in population registries are demographic and socio-economic indicators. Therefore, identification risk management should focus on reducing the identification risk on these public quasi-identifiers (rather than the full set of acquaintance quasi-identifiers).

### B.2.3 Quasi-identifiers Synthesized in the Current Study

This analysis has made the case that under some common assumptions about adversary knowledge and types of clinical trial datasets, the baseline population-to-sample risk for clinical trial data is below the commonly used 0.09 threshold. Therefore, the focus for data synthesis should be on the public quasi-identifiers to protect against sample-to-population attacks. These are the types of quasi-identifiers that were synthesized in our study.

Therefore, the public quasi-identifiers selected for the N0147 trial were (in that order) age, sex, BMI, race, DFS event status, OS event status, time in days to DFS, and time in days to OS. In the case of OS, the event is death (which here is an outcome rather than an SAE), and for DFS the event was death or disease progression. All dates were converted to relative dates (consistent with a contemporary clinical trial de-identification standard [14]).

## B.3 Limitations

The model we use makes assumptions, such as using the Dunbar number. To the extent that these assumptions are reasonable, the estimates here can be relied upon.

Our analysis assumes that the adversary will perform exact matching on the quasi-identifiers. This is a common assumption in the disclosure control literature.

[4] See https://seer.cancer.gov/explorer/application.html?site=1&data_type=5&graph_type=11&compareBy=sex&chk_sex_1=1&series=9&age_range=1&advopt_compprev_y_axis_var=0

# Appendix C: Univariate Comparisons

## C. Introduction

In this appendix we present the methods and results for comparing the univariate distributions in the real and synthetic data.

### C.1 Methods

The univariate results consisted of distributions on the categories of the variables (the relevant continuous variables were categorized in the published secondary analysis study). Relative entropy (KL-divergence [15]) is often used in machine learning to compare two distributions. However, KL-divergence is difficult to interpret because it has no fixed upper bound and is not compared to a yardstick to obtain a relative interpretation. We therefore convert it to a relative value so that it can be interpreted more easily.

Dividing KL-divergence by Shannon's entropy we get the *relative increase in entropy due to using synthetic data*, and we use it to compare the univariate distributions of the real and synthetic datasets. It is a form of normalization of the relative entropy to make it interpretable (in the same way that relative error is interpreted when computing model prediction accuracy). A value of zero means that there are no differences in the distributions. A value of one means that the entropy or uncertainty due to the use of synthetic data as opposed to the real data is twice that of using the real data.

### C.2 Results

The univariate comparisons of the distributions on the KL-divergence metric are shown in Table 1. As can be seen, all the values were less than 1%, therefore the relative increase in entropy is quite low due to data synthesis. The values that are zero in the table pertain to variables that were not synthesized in the partial synthesis process.

| **Variable** | Normalized KL-divergence metric |
|---|---|
| Age | 0.147% |
| Sex | 0.35% |
| BMI | 0.06% |
| ECOG | 0% |
| Race | 0.049% |
| KRAS | 0% |
| T Stage | 0% |
| Histology | 0% |
| Adjuvant Chemotherapy | 0.095% |
| Positive LNs | 0% |
| Adjuvant Regimen | 0% |
| Overall survival | 0.054% |
| Disease free survival | 0.017% |

**Table 1:** Comparing the real and synthetic univariate distributions on the normalized KL-divergence metric.

# References

[1] K. El Emam, L. Mosquera, and C. Zheng, “Optimizing the synthesis of clinical trial data using sequential trees,” *J Am Med Inform Assoc*, Nov. 2020, doi: 10.1093/jamia/ocaa249.

[2] K. El Emam, *Guide to the De-Identification of Personal Health Information*. CRC Press (Auerbach), 2013.

[3] K. El Emam and L. Arbuckle, *Anonymizing Health Data: Case Studies and Methods to Get You Started*. O’Reilly, 2013.

[4] K. El Emam, S. Rodgers, and B. Malin, “Anonymising and Sharing Individual Patient Data,” *BMJ*, vol. 350, p. h1139, Mar. 2015, doi: 10.1136/bmj.h1139.

[5] European Medicines Agency, “External guidance on the implementation of the European Medicines Agency policy on the publication of clinical data for medicinal products for human use,” Sep. 2017.

[6] Health Canada, “Guidance document on Public Release of Clinical Information,” Apr. 01, 2019. https://www.canada.ca/en/health-canada/services/drug-health-product-review-approval/profile-public-release-clinical-information-guidance.html.

[7] K. El Emam, *Guide to the De-Identification of Personal Health Information*. CRC Press (Auerbach), 2013.

[8] Institute of Medicine, “Sharing Clinical Trial Data: Maximizing Benefits, Minimizing Risk,” Washington, D.C., 2015.

[9] K. Benitez and B. Malin, “Evaluating Re-Identification Risks with Respect to the HIPAA Privacy Rule,” *J Am Med Inform Assoc*, vol. 17, no. 2, pp. 169–177, Mar. 2010, doi: 10.1136/jamia.2009.000026.

[10] Janice Branson, Nathan Good, Jung-Wei Chen, Guillermo Monge, Christian Probst, and Khaled El Emam, “Evaluating the Re-identification Risk of a Clinical Study Report Anonymized under EMA Policy 0070 and Health Canada Regulations,” *Trials*, vol. 21, p. 200, 2020.

[11] European Medicines Agency, “European Medicines Agency policy on publication of data for medicinal products for human use: Policy 0070.” Oct. 02, 2014, [Online]. Available: http://www.ema.europa.eu/docs/en_GB/document_library/Other/2014/10/WC500174796.pdf.

[12] CBO, “Research and Development in the Pharmaceutical Industry,” Congressional Budget Office, 2006.

[13] J. M. Unger, E. Cook, E. Tai, and A. Bleyer, “Role of Clinical Trial Participation in Cancer Research: Barriers, Evidence, and Strategies,” *Am Soc Clin Oncol Educ Book*, vol. 35, pp. 185–198, 2016, doi: 10.14694/EDBK_156686.

[14] PhUSE De-Identification Working Group, “De-Identification Standards for CDISC SDTM 3.2,” 2015.

[15] Thomas Cover and Joy Thomas, *Elements of Information Theory*. Wiley-Interscience, 2006.