PRIVACY ANALYTICS
an IQVIA company

How to de-identify (almost) anything

Table of Contents

PRIVACY ANALYTICS
an IQVIA company

Statistical de-identification is enjoying a rise in interest, driven by increased regulation and skyrocketing data demands fed by tokenization and linkage technology, as well as AI initiatives.

Organizations are increasingly conducting research and development built on a complete view of the patient, whether as personalized medicine or via broader applications of AI tools. This creates more demand for linked, multi-domain, multi-modal data about patients that's demonstrably not personal data or personal health information. There is also a spike in data de-identification demand for unstructured and other emerging data sources, from the more familiar text and images to audio and video and even things like physical tissue samples and whole genome sequences.

As organizations (and data privacy solution providers) introduce and scale up their solutions for statistical de-identification, the resulting workflows can be highly contrasting, and the requirements introduced to demonstrate or maintain the de-identification can likewise be quite differentiated. But underpinning these differing results is a consistent set of considerations, building from core methodological concepts consistent with several standards and frameworks [2, 3, 4, 5, 6]. This guide explores de-identification concepts, starting with familiar tabular data and scaling to more complex
data types.

[1] International Organization for Standardization. (2022). Information security, cybersecurity and privacy protection – Privacy enhancing data de-identification framework(ISO/IEC Standard No. 27559:2022). https://www.iso.org/standard/71677.html

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

[3] Office for Civil Rights, "Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule," Department of Health and Human Services, Washington, DC, 2012.

[4] Subcommittee on Disclosure Limitation Methodology, "Working paper 22: Report on statistical disclosure control," Office of Management and Budget, 1994.

[5] Health System Use Technical Advisory Committee and the Data De-Identification Working Group, "'Best Practice' Guidelines for Managing the Disclosure of De-Identified Health Information," Canadian Institute for Health Information, 2010.

[6] Information Commissioner's Office, "Anonymisation: Managing Data Protection
Risk Code of Practice," Information Commissioner's Office, 2012.

Statistical de-identification: hiding in a crowd

Data is considered de-identified when the data subjects (the people described in the data that need their identities protected) can effectively blend in with similar patients in the data—when they can hide in a crowd of patients that look alike. Put another way, data is de-identified when the identifiability of patients (the extent to which they stick out in the crowd) is low enough. Tabular data identifiability can be visualized as building a table of all possible patients in the dataset, filtering that table using known information about a specific patient, like their name, age, or ZIP code, and then checking how many patients fit those criteria.

The referenceable fields are broadly called **identifiers** and range in power. Direct identifiers, like patient names, referenceable patient ID numbers, and patient addresses, are highly likely to be useful in a re-identification attempt, as they're either intrinsically associated with a patient's identity or readily referenceable without special access to data. Indirect identifiers, like demographics, are less powerful than direct identifiers and don't render patients identifiable independently. However, when combined, they may let an adversary narrow down the size of the crowd.

Fields that an adversary can't practically use to narrow down an individual's identity are called non-identifiers. Non-identifiers can vary depending on the situation but may include specific test results that would vary over time, like a blood pressure reading.

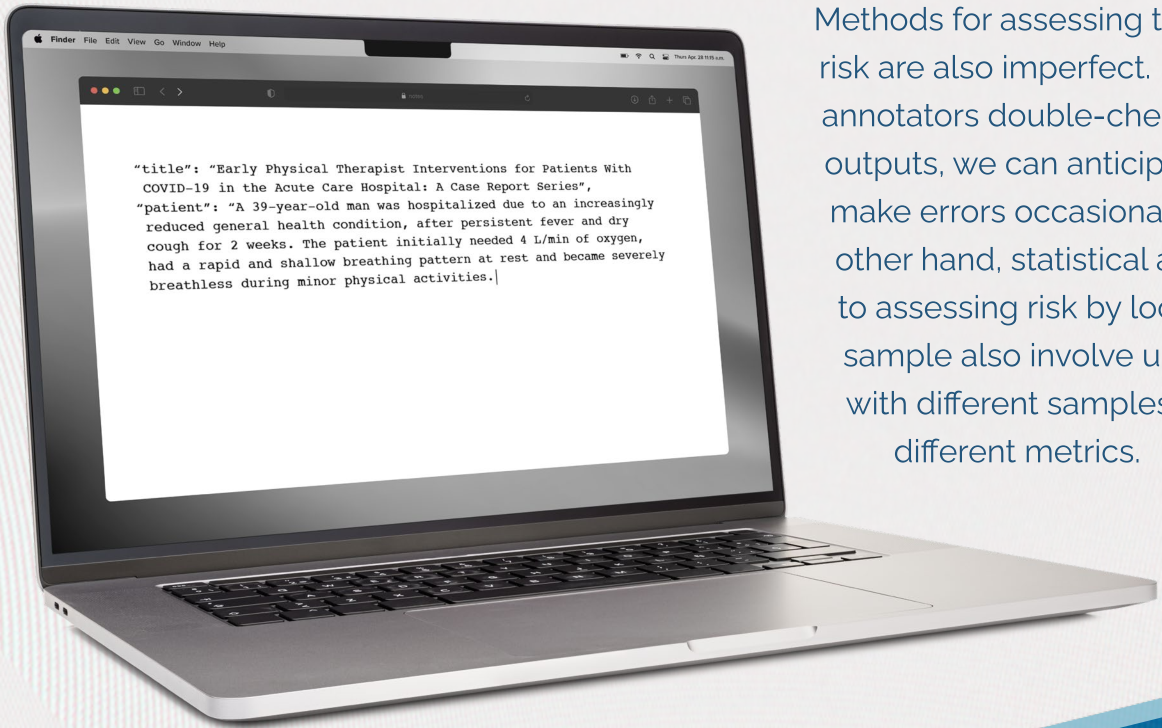
It's important to make sure the groups are set using identifiers that fit a particular threat model. An adversary knows or can learn those identifiers about a patient, making them attackable in practice. The set of identifiers can vary with the threat model and with the context of the data release, which could render some information not referenceable.

If, after filtering the table, the group of patients with matching identifiers is too small (or consists of one patient alone!), then the data may not be appropriately de-identified. The precise cutoff for "too small" is based on the specific use case and context of the de-identified data release. This is guided by existing precedent and scientific research. Groups of patients can be made larger by modifying identifiers to reduce the identifiability but keep some of the utility. For example, generalizing an exact date to a broader date window or replacing a ZIP9 or ZIP5 location code with a less precise one.

Outside of the box(es): Extending from tables to text

Moving from tables to text, the concepts remain consistent. It's still a question of how big the groups are, but now the identifiers aren't neatly organized in tables. Instead, identifiers need to be detected, where a tool or person using a well-defined process flags identifiers as they arise in natural language captured as text.

No detection process is perfect; they all have some residual risk, which is hopefully very small. Rules-based approaches may encounter text features not anticipated when setting the rules, leading to leaks. AI or machine-learning approaches may encounter edge cases where they make the wrong decision. Human annotators will occasionally make errors or have inconsistencies in their judgment. A robust process minimizes errors as much as possible, but that risk cannot be considered zero.



```
"title": "Early Physical Therapist Interventions for Patients With
 COVID-19 in the Acute Care Hospital: A Case Report Series",
"patient": "A 39-year-old man was hospitalized due to an increasingly
 reduced general health condition, after persistent fever and dry
 cough for 2 weeks. The patient initially needed 4 L/min of oxygen,
 had a rapid and shallow breathing pattern at rest and became severely
 breathless during minor physical activities.
```

Methods for assessing the residual risk are also imperfect. If human annotators double-check all outputs, we can anticipate they will make errors occasionally. On the other hand, statistical approaches to assessing risk by looking at a sample also involve uncertainty, with different samples producing different metrics.

These uncertainties can be accounted for with a strong statistical approach whereby you:

1. Carefully select representative text samples to use to evaluate detection,

2. Use a robust method to account for the uncertainty in your overall assessment of risk and

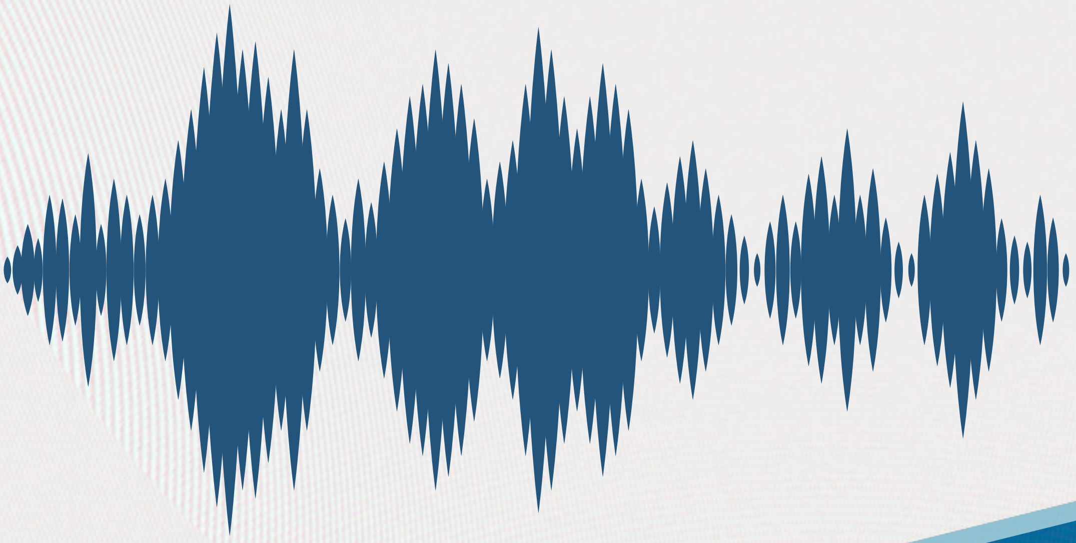
3. Ensure that even the more identifiable scenarios in that "window" of uncertainty stay compliant.

Text de-identification can also benefit from hiding in plain sight or HIPS, where detected identifiers are replaced by realistic-looking surrogates rather than being redacted (i.e., rather than replacing text with "*****" or "[NAME]"). With HIPS, it can be difficult to differentiate between an identifier that was mistakenly left in the data and an identifier that was detected and replaced with a new, de-identified value. So, the overall risk level is reduced.

The voice of the patient: audio and ambient listening

With a handle on text, extending to other forms of language is a manageable jump. There are a lot of applications dealing with audio recordings (or audio tracks of video recordings) containing spoken language, including ambient listening AI applications. With transcription tools, spoken language can be converted to plain text, complete with timestamps for where words start and stop. This text can be addressed using the text techniques of the previous section, with any redaction propagated back to mute portions of audio using those timestamps. This approach doesn't allow for realistic surrogates and the HIPS effect, so all else being equal, it requires higher recall in detecting identifiers.

Audio introduces a new form of direct identifier in a patient's distinctive voice: their tone, cadence, and other attributes of speech, which in many cases can be considered highly identifiable. There are some techniques to depersonalize the sound of someone's voice to render it no longer recognizable or referenceable. In other cases, carefully constructed access controls (like restricting audio from being played except by analysts well-separated from the source of the recordings) can reduce the likelihood of spontaneous recognition of an individual's voice.



A thousand words?
De-identifying images

Moving further, the application can grow from language to images while keeping the "hiding in a crowd" concept from text. This conceptual transition retains the notion of detection but now applies it to features in an image.

Most image formats have two categorical types of information: structured metadata (sometimes called "headers") and pixel data, which is the image itself. The headers are structured and can be considered tabular for this discussion, perhaps with text entries that require the text handling from the previous section. The pixel data is where some new concepts arise.

Pixel data can have a few different identifier types:

- Burnt-in text, where direct and/or indirect identifiers like patient name, exam date, date of birth, or other details are overwritten into the pixel data instead of being stored in the headers.
- Image-captured indirect identifiers that may naturally arise, like treatment information, sex, approximate height, weight, or age, and diagnostic information.
- Image-captured direct identifiers, like the geometry of the face, tattoos, or photos of the head, face, or full body.

Burnt-in text can usually be removed. For readability, it is generally added in a way that doesn't blend in well with the anatomical part of the image, so it's more readily detected. It can be deleted by replacing it with a generous bounding box. Depending on the application, this detection can be very greedy – configured to err on the side of deleting features that may be identifying. You need to consider whether your application will be affected or not if things like earrings, buttons, or grommets are mistaken for text characters and deleted from the image. This redaction approach does lose the benefits of HIPS but can be compensated with the greedy approach.

Other indirect identifiers are typically retained and otherwise handled in the identifiability model. These identifiers are often stored in the metadata, where an adversary can reference them. As such, the analysis treats these as un-transformable when deciding which de-identification approach to take since an adversary can “undo” the transformations by looking at the image. When these data transformations aren’t enough, strengthening the data access controls around the de-identified data can further reduce risk.

Images of the head or face, where required for the application, can be challenging to transform without destroying the usefulness of the image. In these cases, the state of de-identification under a particular regulation can be a tricky topic. Best practices when data can’t be transformed include limiting access, comparison/reference, and visualization so that faces can’t be seen by analysts or compared to identified datasets. There are emerging technology tools for “de-facing” or “skull-stripping” 3D images of the head for applications like neurology, where the image of the brain is needed, but not the anatomy of the face and skull.

It's in whose genes?

Some data types are intrinsically patient-specific: the images of the head and face already discussed, fingerprints, retinal images, whole genome sequences, and (debatably) the voice. For many applications, these data assets can't practically be modified without destroying the analytic value of the data. There's no accurate analogy to generalizing a date or ZIP as there is with tables because, for images, many of the identifiability-reducing changes could render the data useless.

Some best practices limit the identifiability of these types of data. They focus on robust processes, environmental controls, and access controls to ensure the data is only accessed by authorized individuals in approved use cases (with appropriate limits on tooling used on the data and ways data can be explored or consumed). These approaches focus on mitigating re-identification risk by limiting the likelihood of a re-identification attempt being brought against the data rather than limiting the likelihood of the attempt being successful.

Conversely, there's increasing use of transformation-based tools for tokenizing data for initial analysis, where end users get access to encoded versions of the data. With those tools, a researcher may do exploratory work relatively unconstrained and find some genomic marker that correlates with an outcome, provided there are tight controls and governance on reversing the tokenization.

Ready for what's next

The landscape of data applications is constantly evolving, and privacy techniques and best practices evolve along with it. With a strong understanding of basic concepts and some creativity in applying them, data de-identification efforts can expand and scale to address emerging challenges.

Contact us for a detailed discussion on how our services and technology can enhance your data-driven initiatives while protecting patient privacy.

info@privacy-analytics.com

About Privacy Analytics

With Privacy Analytics, you get proven technology and expertise to enable timely, **usable data** that can be **safely linked** and put to work – in compliance with global regulations – and backed by **auditable proof.**

https://privacy-analytics.com/

- ✓ GDPR
- ✓ HIPAA
- ✓ CCPA

100s of other privacy and data protection laws worldwide

100s of clients served over 17 years in business

PRIVACY ANALYTICS
an IQVIA company