**Abstract:** Healthcare industry rely on electronic heath record. Much of the elements are captured in HER or used in insurance claim are in structured form, however, most of EHR includes information about patients visit in clinic through the healthcare system will use XML, narrative text or images- which are in unstructured form. So, it is necessary to de-identify regardless of its format before it can be used for secondary purpose.

**Challenges during de-identification:**

* **De-identifying personal information:** Photographs, videos contain a very high level of information that can be easily used to identify an individual. De- identification seeks to remove this sensitive information while still allowing specific users of the resulting videos.

One of the most complex challenge lies in the range of issues that arise in attempting to de-identify the multimedia content. Multimedia content contains very rich information that can be easily used to identify a specific individual or make it possible to impose constraints on a person’s identity- for example, allowing an observer to conclude that a person is a tall young man.

Evaluating the effectiveness of multimedia de-identification is a multidimensional problem, including:

1. The precision and accuracy of identifying objects requiring de-identification. Google reports that its completely automatic system can blur 89% of faces and 94-96% of license plates.105 Nevertheless, journalists have criticized Google for leaving many faces unblurred. 106 Journalists have also criticized Google for blurring the faces of religious effigies. 107,108 In some contexts, it may be unacceptable to blur or otherwise adulterate certain objects, symbols, or individuals.
2. The reversibility of the transformation. Care must also be taken if pixilation or blurring are used for obscuring video, as technology exists for de-pixelating and de-blurring video by combining multiple images. As an alternative to pixilation or blurring, some researchers have developed systems that can replace faces with a composite face or with a face that is entirely synthetic.
3. The visual quality of the resulting imagery. Blurring and pixilation have the disadvantage of creating a picture that is visually jarring and could potentially affect one’s interpretation of the scene.
4. The effectiveness of the chosen identity obscuring techniques in obscuring identity. While some researchers may score the algorithms against face recognition software, other factors such as clothing, body pose, or geo-temporal setting might make the person identifiable by associates.

* **Challenge with Genomic data:** One of the challenges with genomic data is that it is possible to learn phenotypic information directly. When such information can be ascertained with certainty, it can then be used in a re-identification attack. For example, predictions (varying in accuracy) of height, facial morphology, age, body mass index, approximate skin pigmentation, eye color, and diagnosis of cystic fibrosis or Huntington's chorea from genetic information have been reported.

Combined genomic and clinical data can be quite complex, with free form textual or structured representations, as well as clinical data that are cross-sectional or longitudinal, and relational or transactional.

Several areas will require further research to minimize risks of re-identification of data used for genomic research. For example, improved methods for the de-identification of genome sequences or genomic data are needed. Sequence de-identification methods that rely on generalization that have been proposed thus far will likely result in significant distortions to large datasets. There is also evidence that the simple suppression of the sequence for specific genes can be undone relatively accurately.

* **Challenge with text de-identification:** When dealing with free-text data, the edge for detecting identifiers, in particular direct identifiers, must be extraordinarily high. It is not sufficient to catch 80%, or even 90%, of the names contained in a document.

The proportion of the personal identifiers that are found in the text is referred to as **recall**. In the above case, the recall was 90%. In healthcare, where the inclusion of even one name is considered a breach, we must have very high recall. In fact, it is necessary to take an “all or nothing” approach to evaluation. Finding nine of 10 names does not give us a 90% success rate; rather, it is a 100% failure rate.

While it may seem like the solution is to simply cast a wide net to capture all possible identifiers, care must be exercised or else the quality and usefulness of the data will be negatively impacted. **Precision** is the measure of how many identifiers are correctly detected. In other words, how often is information that is not identifying being redacted or randomized?

The measures of recall and precision are further challenged by the idiosyncrasies of medical information. Drug and disease names often have many variants. A heart attack can be referred to as a cardiac arrest, coronary infarction, myocardial infarction or simply by the abbreviation M.I. Drugs can have multiple brand names as well as their generic name.

It is apparent that the de-identification of unstructured medical data requires expert know-how and ability.

* **Automated text de-identification:** Automated text de-identification applications are mostly based on two different groups of methodologies: pattern matching and machine learning. Many systems combine both approaches for different types of PHI, but the majority uses no machine learning and relies only on pattern matching, rules, and dictionaries.

Main challenge with pattern matching de-identification methods are the already mentioned requirement for developers to craft many complex algorithms in order to account for different categories of PHI, and the required customization to a particular dataset. As such, PHI pattern recognition performance may not be generalizable to different datasets (i.e. data from a different institution or a different type of medical report). Another challenge with pattern matching de-identification methods is the need for developers to be aware of all possible PHI patterns that can occur, such as unexpected date formats.

**References:**

* <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3129641/>
* <file:///C:/Users/sukhdeep%20kaur/Downloads/NIST.IR.8053.pdf>
* <https://privacy-analytics.com/de-id-university/blog/dealing-data-variety-healthcare/>