Medical Data De-ID
A Canadian Perspective

Dr. William Parker, MD, BSc, FRCPC, DABR

Session 3: International session (40 min) Chairperson
Founder @ SapienSecure.io
University of British Columbia
Cardiovascular Radiologist
Disclosures

Founder of SapienSecure

A healthcare intelligence company.

Focused on Medical Data Indexing, Extracting and De-ID.

Many of the examples and tools discussed in this presentation are from using SapienSecure v2 software.
Nova Scotia

Privacy commissioner calls for changes at N.S. Health after staff found snooping

Health authority issues apology, says it plans to accept most recommendations

Josefa Cameron · CBC News · Posted: Feb 08, 2023 9:25 AM PST | Last Updated: February 8
having all employees, consultants, and sub-contractors sign confidentiality contracts prohibiting data linking and/or re-identification;

- only allowing authorized staff to access and use data on a “need-to-know” basis;

- ensuring all employees, consultants, and sub-contractors working with the data receive adequate privacy and security training;

- developing and maintaining data privacy, security, and usage standard operating procedures that specifically prohibit re-identification;

- developing and maintaining strictly enforced retention, destruction and storage policies;

- developing and maintaining role-based data access policies and processes, which are enforced and periodically audited;

- maintaining records of all signed data-sharing agreements and confidentiality agreements, and making those available to the data custodian on request;

- maintaining a proactive program for monitoring privacy, confidentiality and security polices and procedures, a mandatory and on-going training program for all individuals, and a breach protocol that is regularly updated and tested;

- ensuring that external and internal privacy reviews and audits are regularly conducted and that any identified gaps are mitigated; and

- prohibiting data linking and re-identification.
Abstract
The application of big data, radiomics, machine learning, and artificial intelligence (AI) algorithms in radiology requires access to large data sets containing personal health information. Because machine learning projects often require collaboration between radiologists and non-radiologists, it is crucial to consider the implications for radiation exposure and other patient safety issues.
Canadian Association of Radiologists White Paper on De-identification of Medical Imaging: Part 2, Practical Considerations

William Parker, MD, BMSc\(^1,2\)\(^\circ\), Jacob L. Jaremko, MD, PhD, FRCPC\(^3\), Mark Cicero, MD, BESc, FRCPC\(^4,5\), Marlene Azar, MSc\(^6\), Khaled El-Emam, PhD\(^7\), Bruce G. Gray, MD, FRCPC\(^8\), Casey Hurrell, PhD\(^9\)\(^\circ\), Flavie Lavoie-Cardinal, MSc, PhD\(^10\), Benoit Desjardins, MD, PhD, FACR\(^11\), Andrea Lum, MD, FRCPC, CCPE\(^12\), Lori Sheremeta, LLB, LLM\(^13\), Emil Lee, MD, FRCPC\(^14\), Caroline Reinhold, MD, MSc\(^15,16\), An Tang, MD, MSc, FRCPC\(^17\), and Rebecca Bromwich, PhD, LLM, LLB\(^18\)
A. Patient Journey

Patient sees his physician → Physician sends the patient for medical imaging → Images are produced → Images are sent to a picture and communication server (PACS)

Medical imaging refers to any of the tests below:
- MRI
- CT
- PET
- ULTRASOUND
- MAMMOGRAPHY
- X-RAY

Radiologist sends the report to the referring physician → Report is produced
De-identification

Process of transforming direct and indirect identifiers, and possibly implementing additional controls so that the likelihood of data subjects being correctly identified from the information is very small under the circumstances of use or disclosure.

Anonymization

Process of permanently removing all protected health information from the data set.

Pseudonymization

Process of transforming direct identifiers and possibly implementing additional controls so that data subjects cannot be correctly identified from the information under the circumstances of the use or disclosure.

Encryption

Process of using a computer algorithm to obfuscate the patient information, making it unintelligible and random looking. The original information can be recovered using a decryption algorithm.
Classifying Variables

Direct identifiers:
- Patient health number
- Medical images
- Accession number
- Registration number
- City
- Medical registration number
- Postal code
- Hospital
- Phone number
- Name
- Date of birth
- Sex

Indirect identifiers:
Under PIPEDA, personal information includes any factual or subjective information, recorded or not, about an identifiable individual. This includes information in any form, such as:

- age, name, ID numbers, income, ethnic origin, or blood type;
- opinions, evaluations, comments, social status, or disciplinary actions; and
- employee files, credit records, loan records, medical records, existence of a dispute between a consumer and a merchant, intentions (for example, to acquire goods or services, or change jobs).
Prosecutor risk assumes that an adversary is informed of the individuals within a data set.

For example, if a teenager’s parents know that their child has participated in a survey and the results are to be released in de-identified form, the risk of the parents attempting to re-identify their child’s responses would qualify as prosecutor risk.

Journalist risk assumes that the adversary does not know whose data are in the data set, thus the risk of this assessment is significantly decreased.

For example, “if only a sample of de-identified rows from an original data set is released, this would qualify as journalist risk.”

Because terminology has been somewhat inconsistent across statutes and in general use within jurisdictions, it is most useful to focus on the concept of identification risk as follows. This risk consists of 2 criteria: (a) the ability to correctly match a record to a real person, and (b) by doing so learn something new about that person.
Step 1.

Labwork
Clinical Notes
Medications
Medical Imaging
Pathology
Billing
Administrative Coding
Faxes and Forms

Intelligently Aggregated Data
Step 2.

De-Identification Processes

- Hand-writing Redaction
- Date Shift Consistency
- Medical Record Consistency
- PHI NLP
- Forms OCR Redaction
- Image Defacing
- Image Burnt-in Text Redaction
- Differential Privacy
- K- Anonymity
- DICOM De-ID
Medical Record Value Consistency

The unique identifiers, dates, and locations are not usually important specifically in research. But in relationship to other records in the dataset they can be essential to the success of the project.

Medical Record Number or Date consistency software can help create safe datasets that retain value.

<table>
<thead>
<tr>
<th>PatientID</th>
<th>Gender</th>
<th>Age</th>
<th>Zip code</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>55998</td>
<td>M</td>
<td>19</td>
<td>15723</td>
<td>Negative</td>
</tr>
<tr>
<td>88557</td>
<td>F</td>
<td>35</td>
<td>15674</td>
<td>Positive</td>
</tr>
<tr>
<td>55868</td>
<td>F</td>
<td>35</td>
<td>15674</td>
<td>Positive</td>
</tr>
<tr>
<td>44551</td>
<td>M</td>
<td>45</td>
<td>15623</td>
<td>Negative</td>
</tr>
<tr>
<td>58524</td>
<td>M</td>
<td>45</td>
<td>15623</td>
<td>Negative</td>
</tr>
<tr>
<td>25584</td>
<td>F</td>
<td>61</td>
<td>15633</td>
<td>Negative</td>
</tr>
<tr>
<td>58744</td>
<td>F</td>
<td>61</td>
<td>15643</td>
<td>Positive</td>
</tr>
<tr>
<td>87524</td>
<td>M</td>
<td>19</td>
<td>15762</td>
<td>Positive</td>
</tr>
<tr>
<td>87384</td>
<td>M</td>
<td>19</td>
<td>15762</td>
<td>Negative</td>
</tr>
<tr>
<td>17583</td>
<td>F</td>
<td>19</td>
<td>15762</td>
<td>Positive</td>
</tr>
</tbody>
</table>

M: male; F: female
Image De-Facing

Multi-planar (CT and MRI) hold enough spacial information to reconstruct the face of the patient. Many big data projects are focused on neuro-based problems and require imaging. Simple defacing algorithms exist open source, and from businesses that will automate this.
Burnt-in Text Redaction
Medical Reports are treasure troves of medical information that can be game changing for a big-data project. Extraction of information from reports should not compromise privacy. PHI NLP is becoming more and more common, and with transformers, can be extremely accurate. Can be locally trained and re-trained at institutions to improve accuracy over time.

Catheter over guidewire advancement of a 5-French soft tip guiding catheter then followed, with digital subtraction biplane angiograms performed in the right common carotid and right external carotid artery territories.
Differential Privacy

Enough Indirect Identifiers in Combination Can Directly Identify an Individual

<table>
<thead>
<tr>
<th>Name</th>
<th>Postal code</th>
<th>Age</th>
<th>Sex</th>
<th>Dx</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jacquie</td>
<td>S7A 4D1</td>
<td>23</td>
<td>F</td>
<td>Pneumonia</td>
</tr>
<tr>
<td>Patricia</td>
<td>S7A 2X2</td>
<td>27</td>
<td>F</td>
<td>Pneumonia</td>
</tr>
<tr>
<td>George</td>
<td>A4T 1D7</td>
<td>18</td>
<td>M</td>
<td>COVID-19</td>
</tr>
<tr>
<td>Albert</td>
<td>A4T 4E3</td>
<td>13</td>
<td>M</td>
<td>Pneumonia</td>
</tr>
<tr>
<td>Sienna</td>
<td>S7A 1G7</td>
<td>25</td>
<td>F</td>
<td>COVID-19</td>
</tr>
</tbody>
</table>
Hand-writing Redaction

Hand writing is one of the more challenging data-mineable sources, and littered with personal information. My research projects focus on identifying and removing this data. There are simple software tools that can analyze files and detect the presence of handwriting.
Final words...

- True healthcare improvements will occur with increased mobility and movement of medical data.

- Patient privacy does NOT need to be compromised.

- Technology exists to export, aggregate and de-ID medical data almost instantly, and more organizations need to adopt these tools to empower innovation.
THANK YOU

Questions?

Dr. William Parker, MD, BSc, FRCPC, DABR

Session 3: International session (40 min) Chairperson
Founder @ SapienSecure.io
University of British Columbia
Cardiovascular Radiologist