Leveraging data linkage and frame-based textual analysis for the identification of candidate cases prone to suffer from GBV in territories
## Data Problem

The high percentage of Gender-Based Violence (GBV) underreporting in Brazilian municipalities due to:

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>victims' fear of talking about the violence experienced;</td>
</tr>
<tr>
<td>2</td>
<td>health professionals' difficulties in identifying violence episodes</td>
</tr>
<tr>
<td>3</td>
<td>lack of integration in Brazilian public health information systems</td>
</tr>
<tr>
<td>4</td>
<td>lack of instruments to support Health Teams</td>
</tr>
</tbody>
</table>
Data Retrieval Pipeline

Data for going beyond stereotypes
Brazilian Public Health Systems store data of different types about people living in a given territory. Part of this data is parameterized, part takes the form of open text fields.
STAGE 1
Data Collection

TERRITORY A

TERRITORY B

TERRITORY C

MORTALITY RECORDS

VIOLENCE NOTIFICATIONS

e-MEDICAL RECORDS

STAGE 1
Data Collection
STAGE 2
Data Linkage & Anonymization
STAGE 3
Frame Analysis

Cloud Processing

Semantics Analysis

FrameNet Brasil
violence_frames
textual_conditions_frames

Demographic and Violence Metadata

Machine Learning Algorithm

Identification of Patterns of GBV at Territory Level

Semantics Associated with Text
Innovative inputs for prioritizing health strategies in preventing gender-based violence

Semantic frames for gender-based violence case identification in medical records;

Innovative knowledge about gender-based elements sufficient to structure alert systems to support medical professionals in identifying cases of violence that the victim does not refer;

Strengthening of the public health surveillance strategies for GBV
The municipality may require additional time to make the data available, tempering with the timeline;
The municipality may require additional time to make the data available, tempering with the timeline;

The quality of the data for some variables may show low completeness;

The data may be sparse considering the municipality population and the % using the public health system;

The quality of the data may harm the output of the analyses, impairing the usability of the final tool;
Risks

1. The municipality may require additional time to make the data available, tempering with the timeline;

2. The quality of the data for some variables may show low completeness;

3. The data may be sparse considering the municipality population and the % using the public health system;

4. The quality of the data may harm the output of the analyses, impairing the usability of the final tool;

5. The inferences provided by the AI may put victims or candidate victims in embarrassing or vulnerable situations.
<table>
<thead>
<tr>
<th></th>
<th>Mitigation Strategies</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Reordering some steps in the project timeline once we already have parts of the necessary, datasets</td>
</tr>
<tr>
<td>2</td>
<td>Adding more than one municipality to the project.</td>
</tr>
<tr>
<td>3</td>
<td>Adding more than one municipality to the project once there are national patterns, but the data quality can differ in each region.</td>
</tr>
<tr>
<td>4</td>
<td>Adding more than one municipality to the project. If they don't have medical record, it can be used for augmenting the original dataset</td>
</tr>
<tr>
<td>5</td>
<td>Elaboration of strategies for the anonymization of women with a minimum loss of potential in the final tool;</td>
</tr>
<tr>
<td>6</td>
<td>Conduction of focus groups comprising healthcare and social services professionals</td>
</tr>
</tbody>
</table>
### TIMELINE & MILESTONES

<table>
<thead>
<tr>
<th>ACTIVITIES</th>
<th>MAR. 23</th>
<th>APR. 23</th>
<th>MAY. 23</th>
<th>JUN. 23</th>
<th>JUL. 23</th>
<th>AUG. 23</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data collection - municipality will make data available through a cloud environment</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cleaning and organization of databases</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Focus groups to gather qualitative information and guarantee data usage ethics</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Linkage execution</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Tactical Roadmap Revision</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Completion of the paired database</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**OUTPUT 1:** Dataset from the linkage of public health systems of one or two Brazilian municipalities

Sample data analysis and annotation for identification of GBV-related frames and lexical items associated with them, as well as for flagging potential instances of PII in the open text fields.
Adaptation of the FrameNet Brasil database for the representation of the GBV lexicon

OUTPUT 2: Descriptive analysis of the linked dataset focused on the GBV phenomenon

Mid-Point with Program Lead

OUTPUT 3: Descriptive analysis of the linked dataset focused on the data quality and the services provided

Interim Grant Report

Annotation of text data with frames and frame elements for building the gold standard dataset

Assessment of ML methods and techniques for suitability to the problem and dataset

Test-implementation of selected algorithms

Ablation studies for defining the role of different data structures in the results
**OUTPUT 4:** Revised and expanded model of the lexicon associated with the domain of GBV, defined in terms of computationally implemented semantic frames

**Analysis**

Final implementation of selected algorithm

**OUTPUT 5:** Extendable, general framework for GBV risk classification of territories, based on public health records and their respective links and semantic information

**OUTPUT 6:** Territorialized data visualization dashboard to instruct public policy - to be available for the local partner surveillance health department

**Project wrap-up**

**Final Insight Report**

**Final Activities:** Technical report; Blog #2

**Final Grant Report** with the code library
Thank you!