ufjf | FRAMENET BRASIL

Vital Strategies

# Leverage data linkage and frame-based textual analysis for the identification of candidate cases prone to suffer from gender-based violence in Recife, Brazil.

Final Insights Report

2023 Data to Safeguard Human Rights Accelerator

# Summary

# Executive summary

The Brazilian Unified Health System (Sistema Único de Saúde – SUS) has different information systems that track different types of information: hospitalization, violence notification, medical records, mortality, and others. Throughout these different systems, there is no single ID that allows for the identification of the same individual in different databases. This creates challenges for tracking individuals at risk and helping victims before violence escalates. Moreover, not every field in those systems is parameterized. Some of them allow for health teams to input open text, which presents an additional challenge for data processing.

Attentive to these problems, Vital Strategies Brasil partnered with FrameNet Brasil to link data from different databases and analyze open-text fields with a goal to identify patterns that could help health professionals conduct early identification of gender-based violence in routine medical appointments.

From these analyses, we have found that:

- Generally, women who are victims of GBV tend to have more records in e-SUS AB, meaning they visit primary healthcare units more often.
- Distributing the records over time, we noticed there is a strong correlation between SINAN notifications and e-SUS records, providing evidence that GBV cases could be identified based on medical records.
- There is a systematic increase in the number of visits to the doctor from around 60 days before the SINAN notification to 200 days after.
- Evaluating patterns between clusters of records, we noted that e-SUS records that are known to be from women who are suffering violence are overlapped by cases that were labeled as likely associated with violence, validating the hypothesis about the relevance of the dates for the relationship between records.

These insights indicate that early identification is feasible and serve as a baseline for future studies on GBV developed by Vital Strategies and FrameNet Brasil.

# Introduction

This is the final insights report for the 2023 cohort of the Data to Safeguard Human Rights Accelerator program, promoted by Patrick J. McGovern Foundation. In this project, Vital Strategies Brazil and FrameNet Brasil partnered to leverage data linkage and frame-based textual analysis for the identification of candidate cases prone to suffer from gender-based violence, informing policy makers and health care teams of territories and situations in which these cases can occur.

The project was conducted using data from Recife, the capital municipality of the state of Pernambuco, in northeast Brazil. Recife has one of the biggest populations and GDPs of its region. The city's Municipal Health Department partnered with Vital Strategies Brazil and shared data on violence, hospitalization, deaths and digitized medical records from Primary Healthcare services in order to gather in-depth insights on the gender-based violence (GBV) scenario in the city.

The majority of available health data is categorical, meaning most epidemiological analyses rely on objective data and statistical insights. However, an important part of the doctor-patient relationship is the narratives shared during routine appointments. In Brazil, this data is stored in open text fields in medical records from the primary health care database, called e-SUS-AB. This data has been linked to other data sources, allowing a more complete picture of a GBV victim.

Open text fields are rarely analyzed due to the complexity of working with language. Computational linguistics methods — FrameNet Brasil's expertise — are used to annotate text for meaning representations in large sets of open text data. Through semantic analysis of samples of the data, FrameNet's team has modeled frames and lexical items covering the healthcare and violence domains. These frames and lexical items were used to create semantic representations of open text fields, which were then fed into a machine learning model used to look for GBV patterns.

During the execution of this project, data linkage, data exploration and semantic analysis provided relevant insights on the association between violence notifications and primary healthcare visits, indicating early identification of GBV victims is possible. This finding will fuel the further development of the machine learning model and can help deepen the understanding of the health consequences of GBV to women in Brazil.

This report is divided into three sections, apart from this introduction: a methodological chapter, describing the databases used, the data linkage process, and the methods to extract knowledge using AI from open text fields; a chapter containing the main results and findings from the work so far; and a final chapter, with the main insights and paths to the continuity of this work.

# Methods

This section describes the tools and methods used in this project. Firstly we describe the datasets used and their purposes; then we explain the data linkage and anonymization protocols; next, we describe the semantic analysis strategies; and, finally, we describe the process of modeling GBV phenomena into the artificial intelligence model built for this project.

## Datasets

Violence is a multi-factor phenomenon and can be traced through different types of data. In Brazil, different public health information systems track information relevant to understanding GBV: hospitalization (SIH database), violence notification (SINAN-Violência database), medical records from primary health care services (e-SUS AB database), and mortality (SIM database). Unfortunately, there is no single ID that allows for the identification of the same individual in different databases.

Each of these databases has unique characteristics regarding their purposes. The hospitalization database - SIH - registers data on hospital stays in the public health system to ensure services receive funds to cover the procedures performed on patients. Therefore, it is designed as a payment system, but it can provide relevant epidemiological information.

In Brazil, health workers have an obligation to notify the government whenever they identify cases of violence against women and attempted suicide. This data is registered on SINAN, a database that keeps records of events of mandatory notification. Due to its definition, this system is the main reference we have when searching for GBV data, and, linked to the other databases, it provides a powerful diagnosis about violence against women, but also on data quality.

Recently, there has been an effort to digitize medical records in Brazil through e-SUS AB, a system designed to collect and store data generated by primary healthcare services. This system is being implemented throughout the country and it stores great volumes of data since it keeps records from routine medical appointments.

Finally, the Mortality Information System database - SIM - has almost universal coverage and keeps records of every death in the country and general information on the cause of death and victim's profile.

## Data linkage and anonymization

The first step to analyzing the paths of victims is to link datasets from these different databases, identifying the presence of victims of violence and tracking their way through the public health services available. To conduct such linkage, it is necessary to access identified health data, which imposes ethical and bureaucratic challenges.

In order to access such information, Vital Strategies Brasil signed a cooperation agreement with the municipality of Recife, assuring high standards of data protection protocols in order not to expose any patient. Only the technicians responsible for linking and anonymizing the data were allowed to access identified records and, after these steps, all the analyses were conducted through anonymized records. Besides that, all the uploaded data was encrypted using a password known only by the technicians involved in this process.

There are two main methods for pairing datasets: probabilistic and deterministic. Both have gray areas, meaning both can identify false positives or false negatives. Probabilistic methods estimate the probability of cases being recognized as equal across datasets. To avoid the gray area using the probabilistic methods, we need humans to evaluate and define whether or not multiple records correspond to the same person. Therefore, we decided to apply deterministic methods, based on rules built around the quality and reliability of the data we were working with. Vital Strategies Brazil has a linkage algorithm, which has been used in other projects with similar purposes such as identifying underreporting and improving data quality among health information systems.

The deterministic algorithm uses rules from a combination of key variables and was developed regarding the fields available in each of the analyzed databases. The quality of the data is essential to define these rules. The first step was the pre-processing of the data for correction and standardization of variables such as name; mother's name; date of birth; street, and neighborhood, all used in the rules for record comparison. The main changes in the data were: the removal of punctuation marks, accents, repeated blanks, and prepositions; conversion of letters to uppercase; removal of numbers from variables that should be exclusively composed of letters and vice versa, removal of terms that indicate the lack of information (do not know, unknown, among others), replacement of double letters by a single one, standardization of date formats, standardization of terms used in public places ("R." was replaced by "Street", "Av." by "Avenue" etc.).

Then, new fields containing textual information (e.g. names and addresses) were created for standardization and comparison, through the following steps:

1. Parsing (separation of the fragments into first name, second name, and so on);
2. Substringing (parts of the fragment such as "Maria" → "Mari"; "Oliveira" → "Veira").

The new variables underwent a new change to their Soundex code, which transforms words into codes capable of capturing phonetic relations in comparisons, such as the use of "s" or "c" and the use or not of double consonants. Greater detail about the relationship of the data in Oliveira et al. (2016).

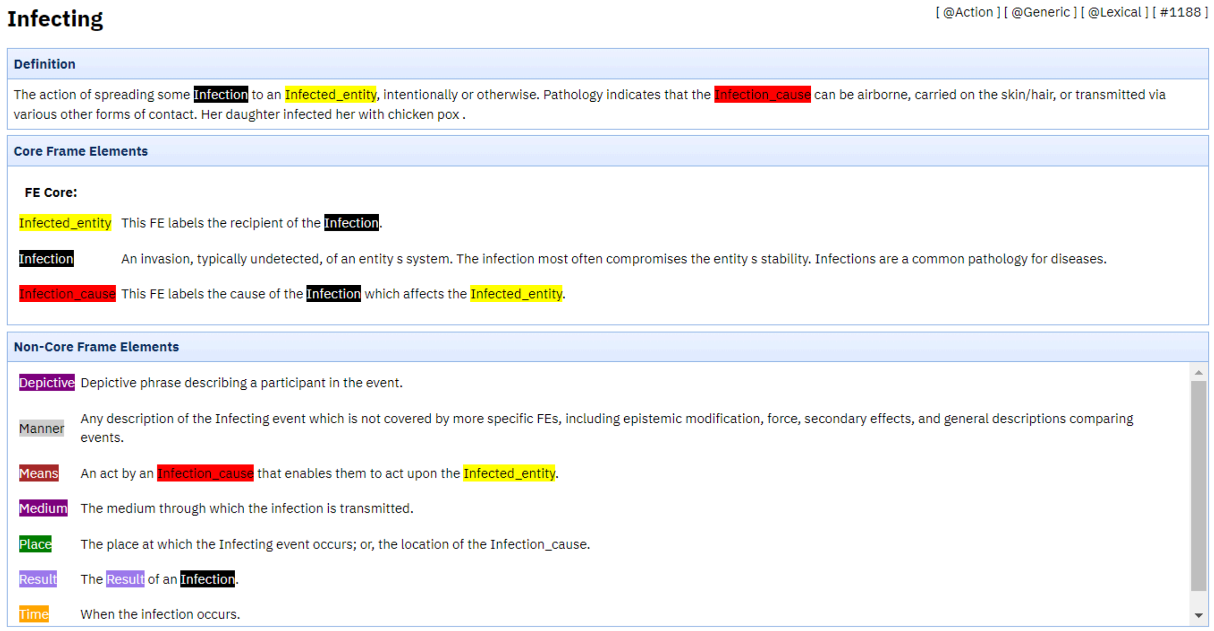
After the linkage, the anonymization process was executed through a combination of manual, automatic, and semi-automatic methods. The framework uses a combination of named entity recognition (NER) AI models, regular expressions, and fuzzy search to identify potential personal identifiable information (PII). When multiple methods agree that a text span contains PII, that part of the text is replaced with a special anonymization tag. When there is no agreement about a span, a tool was used to display the span and a technician had to evaluate whether the span contained sensitive information or not. The tool also allowed for manual search and tagging of PII, for the cases where the NER models and search algorithms failed.

After that, the linked database, without any personal identification, was shared with the computational linguistics team, responsible for performing the semantic analysis.

## Semantic features extraction

The inherent complexities of language, such as ambiguity, variability, and more generally, ethical concerns, are often the reason why natural language processing (NLP) can be challenging. While modern NLP technologies address some of these issues, they often lack interpretability. On the other end of the spectrum, traditional models like bag-of-words provide interpretability but suffer from high dimensionality and information loss. Given the importance of interpretability for acquiring valuable insights in this project, we have opted for FrameNet-based semantic structures rather than relying solely on textual data, because they capture information such as context, common sense, and cultural nuances.

A FrameNet is a network of frames. According to Frame Semantics (Fillmore, 1982), a frame represents a scene, including its participants, their props, and how they interact. For instance, the frame `Infecting` (see Figure 1) represents a situation where there are three main participants: the Infection, the Infected_entity, and the Infection_cause. Other elements may also appear, such as Place and Time, providing more context and information about the scene. These scenes are also associated with lexical items that evoke them, i.e., in a sentence, words bring forward the background knowledge expressed by frames. In the case of `Infecting`, that would happen, for example, with the verb *infect*.



**Infecting**

[ @Action ] [ @Generic ] [ @Lexical ] [ #1188 ]

**Definition**

The action of spreading some Infection to an Infected_entity, intentionally or otherwise. Pathology indicates that the Infection_cause can be airborne, carried on the skin/hair, or transmitted via various other forms of contact. Her daughter infected her with chicken pox .

**Core Frame Elements**

**FE Core:**

| | |
|---|---|
| Infected_entity | This FE labels the recipient of the Infection. |
| Infection | An invasion, typically undetected, of an entity s system. The infection most often compromises the entity s stability. Infections are a common pathology for diseases. |
| Infection_cause | This FE labels the cause of the Infection which affects the Infected_entity. |

**Non-Core Frame Elements**

| | |
|---|---|
| Depictive | Depictive phrase describing a participant in the event. |
| Manner | Any description of the Infecting event which is not covered by more specific FEs, including epistemic modification, force, secondary effects, and general descriptions comparing events. |
| Means | An act by an Infection_cause that enables them to act upon the Infected_entity. |
| Medium | The medium through which the infection is transmitted. |
| Place | The place at which the Infecting event occurs; or, the location of the Infection_cause. |
| Result | The Result of an Infection. |
| Time | When the infection occurs. |

**Figure 1. Frame `Infecting`**

Given the project's focus on GBV and its work with data from the public health system, the frames for the Violence and the Healthcare domain had to be modeled before they could be used for semantic information extraction.

## Modeling the Healthcare and Violence Domains

In FrameNet, a domain is portrayed through a set of scenes closely related to that aspect of reality, such as Healthcare and Violence. These scenes are represented by frames connected via a series of typed relations, forming a network of frames (e.g. `Infecting` and `Pathogens are connected` via Using). These frames are then evoked by lexical units (LUs) that appear as words or expressions in texts where that domain is a relevant topic. These LUs can be further related to each other by qualia relations (Pustejovsky, 1995). Qualia relations capture, for example, the relation between a medical specialization and the body part this specialization focuses on (*e.g.* "ophthalmologist" and "eye"). When describing a domain, we report the number of frames and how they relate, as well as the number of lexical items and their qualia relations.

The computational modeling of a Frame-semantic specific domain followed a nine-step methodology, as outlined by Costa (2020), involving the analysis of textual data related to that domain. This methodology was expanded and adapted to model the Healthcare domain and was described by Dutra et al (2023). It was later used to model the Violence domain.

The Healthcare domain consists of a total of 33 frames, including 17 new frames created specifically for this project, and 16 that already existed in the FrameNet Brasil database. Among these, 6 frames had to be modified to align more closely with the project's goals. The Frame-to-Frame (F-F) relations defined between the frames (TORRENT et al., 2022) that modeled the domain are shown in Figure 2. Associated to these frames, 4101 lexical units and 3743 qualias relations were established.

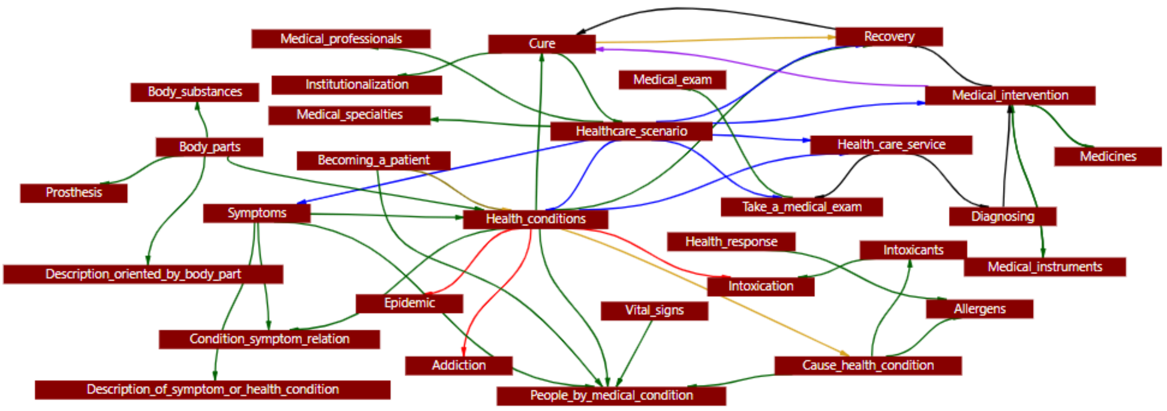


**Figure 2. F-F Relations in the Healthcare Domain**

A similar process was carried out for the creation of the Violence domain. In this instance, a bigger number of the existing frames from the FrameNet Brasil database were used, totaling 45 frames. Out of these frames, only one required modification. Moreover, 6 new frames had to be created to address the needs of the domain, increasing the total to 51 frames. The F-F relations established among the frames that modeled the domain are illustrated in Figure 3. Associated with these frames, 2069 lexical units were created, and 1627 qualia relations between them.

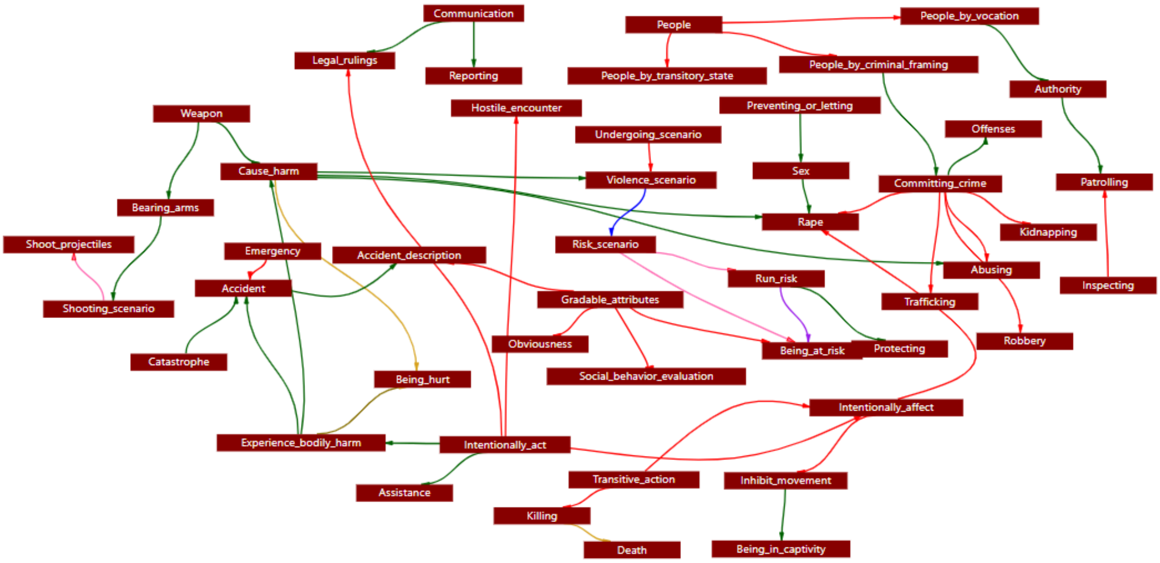


**Figure 3. F-F Relations in the Violence Domain**

## Text annotation

With the frames and their relations defined, the next stage of the process was carried out: *corpora* annotation. The annotation in this project consisted of tagging meaningful lexical units (LUs) for the frames they evoke and where the frame elements of these frames are located in each sentence. This annotation was executed through two methods: manual and automatic.

Since the manual annotation task is relatively time-consuming, it could only be carried out using a sample of the text fields from the public health systems. This annotation serves two purposes: (i) to validate the model for the two domains and (ii) to constitute a training dataset for an automatic labeling neural model that can be used for all other text fields and even new ones.

For the automatic labeling of text fields, we trained a variation of the LOME model using existing *full-text* annotation of frames and the newly annotated sentences of the Violence and Healthcare domain. Using this trained model, we were able to obtain the frames and frame elements of each sentence in each text field from e-SUS AB and SINAN.

## Model features

In this project, we chose to use the e-SUS records as inputs to the models, while records from SINAN played a crucial role in classifying these inputs. Information from the other two systems can be used to further classify cases (e.g. *Did the violence notification lead to the victim's death?*) and find unreported cases.

Because of that, we focused on building semantic features for the e-SUS records, taking into consideration its fields. For each record, the semantic representation consists of a transformation based on all sentences from the 7 open text fields that form the electronic medical record and the ICD code assigned to that record by a health professional. First, LOME automatically labels each sentence - including the description of the ICD code.

Then, a second algorithm is used to find, based on LOME annotations, the lexical units within that sentence. Since LUs represent words and expressions, these can be especially valuable in cases where the analysis requires more fine-grained information. While frames and frame elements are to characterize a situation to the extent required by the analysis, the actual choice of words may become important in other cases, such as when describing symptoms.

These annotations are transformed into feature vectors indicating the frequency of occurrence for each frame, frame element, and lexical unit that appeared, as well as whether a given frame or frame element co-occurred with another. For each LU that appeared at least once in this representation, we verify whether it has a qualia relation to any other LU within that same vector. Subsequently, the vector is expanded to represent those relations.

The next step in the pipeline consists of finding the TF-IDF (Term Frequency-Inverse Document Frequency) representation of a vector, *i.e.,* to weight the counts of features by their significance and frequency for each individual vector. The TF-IDF transformed vector serves as the fundamental representation of a record in our analysis.

## Principal Component Analysis (PCA)

In comparison to a method such as bag-of-words, FrameNet semantic features exhibit considerably smaller dimensionality. However, the number of dimensions remains substantial for efficient machine learning. Even after basic preprocessing operations, we were left with 40333 features, consisting of frames, their elements, the co-occurrence of those two entities, and, most importantly, lexical units (LUs), along with the type of qualia relation between two of the LUs whenever applicable.

To reduce the number of features to a more feasible amount, we used principal component analysis (PCA). Each principal component represents a combination of the features and by retaining 4000 (nearly a 10x reduction in size) we were able to still preserve over 90% of the original data variance.

The combination of text features with PCA is a popular technique in literature, often referred to as Latent Semantic Analysis (LSA). LSA is used to reduce the dimension of traditional text features because different forms in text may have highly similar semantic properties. For instance, some synonyms or inflections for gender and number. When incorporating FrameNet semantic information, rather than grouping words with semantic similarities as a single feature, we are grouping frames, frame elements, and lexical units that exhibit very similar patterns of evocation.

When analyzing the data, we used this smaller, principal component representation instead of the basic TF-IDF representations.

# Analysis

## Gender-Based Violence in Recife: An Overview

This section explores data on violence against women available on SINAN, the main data source for GBV cases in Brazil, in order to understand the context of violence in Recife.

From the data available at SINAN and the population data made available by the Brazilian Institute of Geography and Statistics (IBGE), we can calculate the notification rate per 100.000 inhabitants in Recife. This rate helps us to understand the evolution of notifications through the years and compare Recife with other cities in Brazil:

**Table 01: Number of records, number of women, female population and notification rate per 100 thousand female inhabitants per year. Recife, 2016 a 2022. SINAN-Violência.**

| Year | Number of notifications | Number of women | Population | Notification rate per 100.000 female inhabitants |
|---|---|---|---|---|
| 2016 | 1419 | 1380 | 876505 | 157,4 |
| 2017 | 1779 | 1721 | 881165 | 195,3 |
| 2018 | 1843 | 1780 | 886107 | 200,9 |
| 2019 | 1968 | 1922 | 890946 | 215,7 |
| 2020 | 1554 | 1494 | 895699 | 166,8 |
| 2021 | 2183 | 2102 | 900372 | 233,5 |
| 2022 | 2815 | 2668 | 900372 | 296,3 |
| **Total** | **13561** | **12607** | **6231166** | **202,3** |

**Table 02: Ranking of the ten Brazilian capitals with higher notification rates per 100 thousand women. Recife, 2011-2021, SINAN-Violência.**

| Capitais | 2011-2021 | Ranking |
|---|---|---|
| Campo Grande (MS) | 558,1 | 1 |
| Vitória (ES) | 468,0 | 2 |
| Palmas (TO) | 446,5 | 3 |
| Curitiba (PR) | 393,5 | 4 |
| Rio Branco (AC) | 271,3 | 5 |
| Boa Vista (RR) | 243,7 | 6 |
| Rio de Janeiro (RJ) | 198,1 | 7 |
| Belém (PA) | 195,8 | 8 |
| João Pessoa (PB) | 195,7 | 9 |
| **Recife (PE)** | **191,2** | **10** |

A descriptive analysis of the Violence Notification System (SINAN) gives us interesting insights into the quality of the records in Recife. Although the notification of violence is mandatory in Brazil, when identified by health professionals, there is still a great level of non-notified cases.

The major notifiers are Hospitals (52.2%) followed by Basic Health Units (22.9%) and Emergency Services (20.9%). This data indicates that most cases of violence are being identified too late, only when cases have more severe consequences for women.

**Table 03: Proportion of notifications by notifying units. Recife, 2016-2022, SINAN - Violência.**

| Notifying Unit | % |
|---|---|
| Hospitals | 52.2% |
| Basic Health Units | 22.9% |
| Emergency Services | 20.9% |
| Others | 4% |

From the data registered in SINAN, most notifications are for adult women: 37.4% of cases happened to those between 30 and 59, and 23.5% against women from 20 to 29 years old. On the racial profile of victims, 69.7% are identified as black and 20.1% are white. However, 8.2% of the entries do not have information on the race/color of the victim.

Regarding the type of violence most present, the majority of cases (75.5%) are for interpersonal violence against 18.1% of self-inflicted violence[1]. There is a high prevalence of self-inflicted violence among young girls: 33.6% of these cases are of violence against girls between 10 and 19 years old.

Within the interpersonal violence cases, the majority of cases registered are for physical violence (51.9%), followed by psychological (33.7%) and sexual (26.4%). Cross-referencing this data with the age of victims, there is a high proportion of sexual violence against girls from 10 to 19 years old (39%) and of negligence against girls from 0 to 9 years old (64.4%). Among adult and elderly women, physical and psychological violence prevails.

**Table 04: Proportion of notification by type of violence and age. Recife, 2016-2022, SINAN - Violência[2].**

| Type of violence | 0 to 9 years old | 10 to 19 years old | 20 to 29 years old | 30 to 59 years old | 60+ | Total |
|---|---|---|---|---|---|---|
| Physical | 17.3 | 40.4 | 61.3 | 66.3 | 46.5 | 51.9 |
| Psychological | 12.1 | 18.8 | 32.1 | 50.2 | 44.1 | 33.7 |
| Sexual | 23.9 | 39.0 | 28.1 | 21.3 | 9.0 | 26.4 |
| Negligence | 64.4 | 8.0 | 1.0 | 1.9 | 23.9 | 12.9 |
| Other types | 3.7 | 27.4 | 28.2 | 30.5 | 28.8 | 25.4 |
| No type of violence registered | 1.9 | 1.8 | 1.7 | 1.7 | 2 | 1.7 |

According to the data, the main aggressor was an intimate partner (39%), followed by other family members (27.1%) and other people known to the victim (9.9%). This indicated that most GBV notified in Recife is domestic violence.

When a case of GBV is identified, the professional caring for the victim may suggest a referral to other services to protect the woman or attend to her necessities in the best way possible. 37.9% of cases did not receive any referral for other services. The main referral was for another healthcare service (24.6%), followed by Child Protection Services (23.6%).

[1] 6.4% of cases were registered as undefined/does not apply.
[2] The proportions don't add up to 100% because one entry can identify more than one type of violence.

**Table 05: Proportion of types of referrals registered. Recife, 2016-2022, SINAN - Violência.**

| Types of referrals | % |
|---|---|
| No referral | 37,9% |
| Healthcare services | 24,6% |
| Child Protection Services | 23,6% |
| Women’s Police Station | 15% |
| Others | 23,9% |

Analyzing the deaths of women registered in Recife from 2016-2022 on the Mortality Information System (SIM), approximately 6% are related to external causes. Among these, 12.5% are classified as death by aggression. The majority is classified under “Other external causes” (48.6%).

Deaths from aggression prevail among girls from 10 to 19 years old and women from 20 to 29. For both of these groups, aggression was the cause of more than half of deaths by external causes:

**Table 06: Proportion of external causes of death by age. Recife, 2016-2022, SIM.**

| Cause of death | 0 to 9 years old | 10 to 19 years old | 20 to 29 years old | 30 to 59 years old | 60+ | Total |
|---|---|---|---|---|---|---|
| Ext Others | 77.4 | 23.7 | 15.4 | 38.2 | 56.4 | 48.6 |
| Ext W00-W19 Falls | 3.2 | 0.9 | 2.1 | 6.7 | 36.3 | 25.3 |
| Ext X85-Y09 Aggression | 9.7 | 55.3 | 50.3 | 25.7 | 1.1 | 12.5 |
| Ext V01-V99 Transportation accidents | 6.5 | 10.5 | 14.9 | 12.7 | 3.9 | 6.8 |
| Ext X60-X84 Self-inflicted injuries | 0 | 6.1 | 13.8 | 14 | 1.8 | 5.3 |
| Ext W65-W74 Accidental drowning | 3.2 | 0.9 | 0.5 | 0.9 | 0.2 | 0.5 |
| Ext X00-X09 Exposure to smoke, fire, or flames | 0 | 1.8 | 0 | 0.9 | 0.3 | 0.5 |
| Ext X40-X49 Poisoning, intoxication by or exposure to harmful substances | 0 | 0.9 | 3.1 | 0.9 | 0 | 0.5 |

From this data, we are able to perform a general diagnosis on the present scenario of violence, identifying risk factors - such as age and race - and generally understand the characteristics of gender-based violence in Recife. However, more than half of GBV cases are being notified by hospitals, indicating that violence is being identified when injuries are more severe.

Our hypothesis is that it is possible to identify GBV earlier, through routine appointments conducted in primary health care facilities. Women attend such appointments to consult with doctors on a series of everyday issues and maybe the records from these appointments can provide insights on early signs of violence.

The information present on e-SUS is more qualitative, however, with written observations made by nurses and doctors through text. In order to analyze text, we used semantic analysis techniques to unveil patterns “hidden” in open text fields.

## The relation between primary health care visits and violence notifications

To identify cases prone to suffer from GBV, this project focuses on the elaboration of a computational model that extracts data from primary health care records and uses it as input to build a machine learning model. This model computes a score to identify how similar a given medical visit is to those of known cases of GBV.

The choice of primary health care visits as the main observation point has to do with the number of records of this type, which is considerably higher than in other systems. Cases that were not reported in SINAN may likely be identifiable in one or more e-SUS AB records, based on the patient history or the health professional observations, which are recorded in text fields. It is important to note that, although notification is compulsory, there is a relevant amount of underreporting due to several factors - lack of sensitivity to identify GBV, lack of resources to notify, difficulty in reporting to the health professional the real cause, etc.

It is a major challenge to identify underreporting in health systems that do not have national coverage, are built with records from compulsory notification, come from services that do not perform active search for cases, and that, in part, depend on the sensitivity of health professionals. Identifying e-SUS records from patients who are victims of violence becomes even more challenging due to the system’s purpose, which is to collect information from the entire primary healthcare system. This means the main source of information are the open textual fields filled by health professionals. Therefore, SINAN records served as the main markers to target cases of GBV. These records represent higher-risk cases, which lead to notification and sometimes death. The database linkage was essential to allow the identification of e-SUS AB records belonging to women who suffered violence or  women who died due to violent causes.

Note, however, that the same individual can have multiple records in SINAN, *i.e.*, multiple instances of GBV, but there's no obvious way to organize their primary care visits into more or less related to each SINAN record. This creates a challenge in the analysis phase because not all medical records will be related to the episode of violence: each woman can go to different medical visits, regarding different health issues. Moreover, it's unlikely - but not impossible - that a primary care visit many years before a GBV episode is more related to that episode than more recent visits. In that sense, medical visits and violence episode dates become relevant to the task of qualifying the relation between entries from different public health systems.

In this section, we explore important aspects of e-SUS AB entries and their relation to SINAN records. We'll present insights obtained in that process and discuss how they impact some choices made when developing the machine learning model.

## Distribution of e-SUS AB and SINAN records over time

One way to explore the relation between records from the two different information systems is to see how their entries are distributed over time and compare those two distributions.

Figure 04 shows the distributions of e-SUS AB records over time. Each bar represents the number of primary care visits by month. The x-axis is limited to the time period of 2016-2023 due to the availability of data for this project. There is a clear increase in the number of records over time, indicating a gradual increase in the number of units using the information system. There are also seasonal variations: there is a decrease in the overall number of visits near the last three months of the year (the beginning of summer in the region), and there is a notable decrease in entries from March of 2020, due to the COVID-19 pandemic, a public health emergency that demanded most of the attention from all of the primary healthcare staff and some services were interrupted. Also, patients became reluctant to go after public health services.

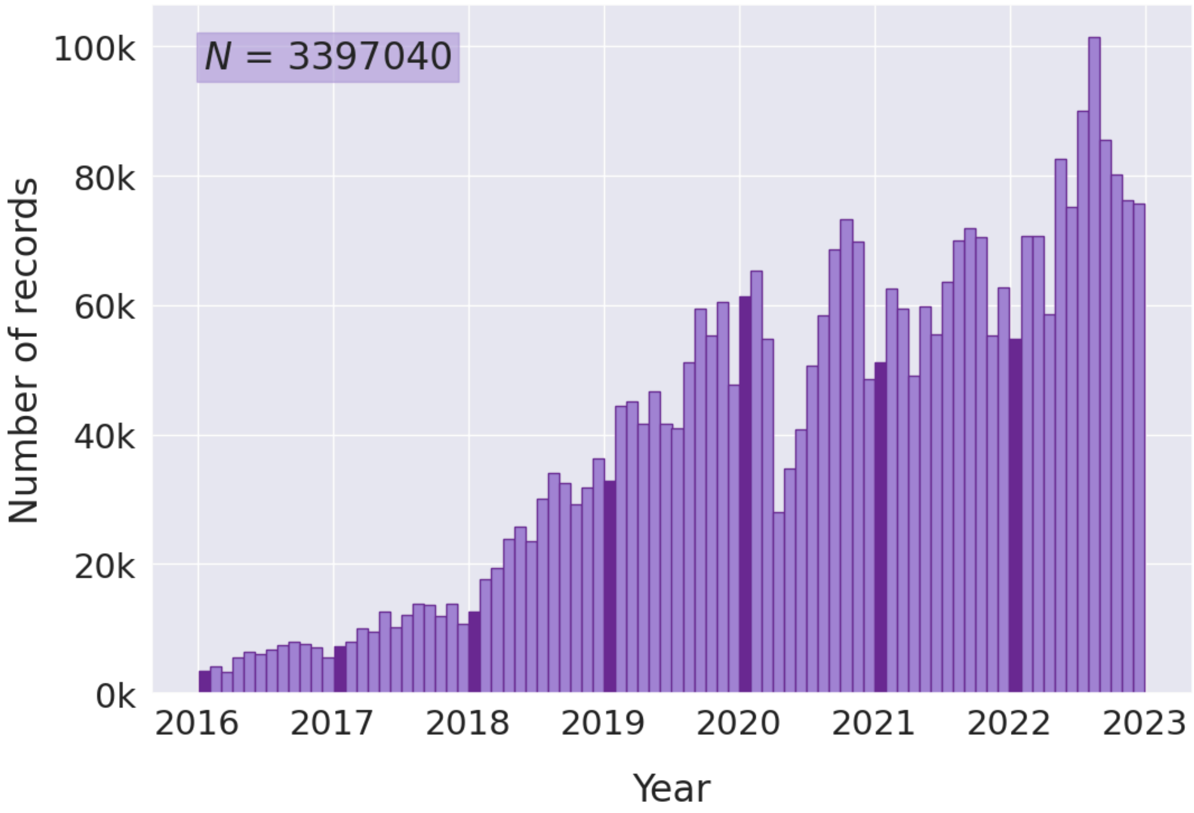


**Figure 04: Distribution of e-SUS AB records over time (January is highlighted for each year)**

For comparison, Figure 05 shows a similar chart, for the same period, but for the number of SINAN records. The highly reduced number of records at the beginning of 2020 shows how the pandemic similarly impacted the number of entries in SINAN. This is, however, the only significant similarity. SINAN records are distributed in time very differently from e-SUS records: they are much more uniform, especially from 2017 to 2021, with the exception of the COVID-related impact.

There's a smaller number of reports in 2016 and there seems to be a significant increase in reporting starting from the end of 2021 onwards.

It is expected that e-SUS will have a higher number of records in comparison to SINAN - Violência, since routine appointments with a physician are far more frequent than cases of violence. Also, the e-SUS system is under implementation and it is expected that the number of records will increase in time. SINAN - Violência has been in place for longer and it is expected that, after implementing this system, notifications will stabilize in time, if conditions remain similar and violence does not generally increase in a specific location.

Finally, this chart is considerably more noisy than the one above due to differences in the number of records: while 05 was plotted based on 13 thousand data points, 04 was computed from almost 3,4 million records.

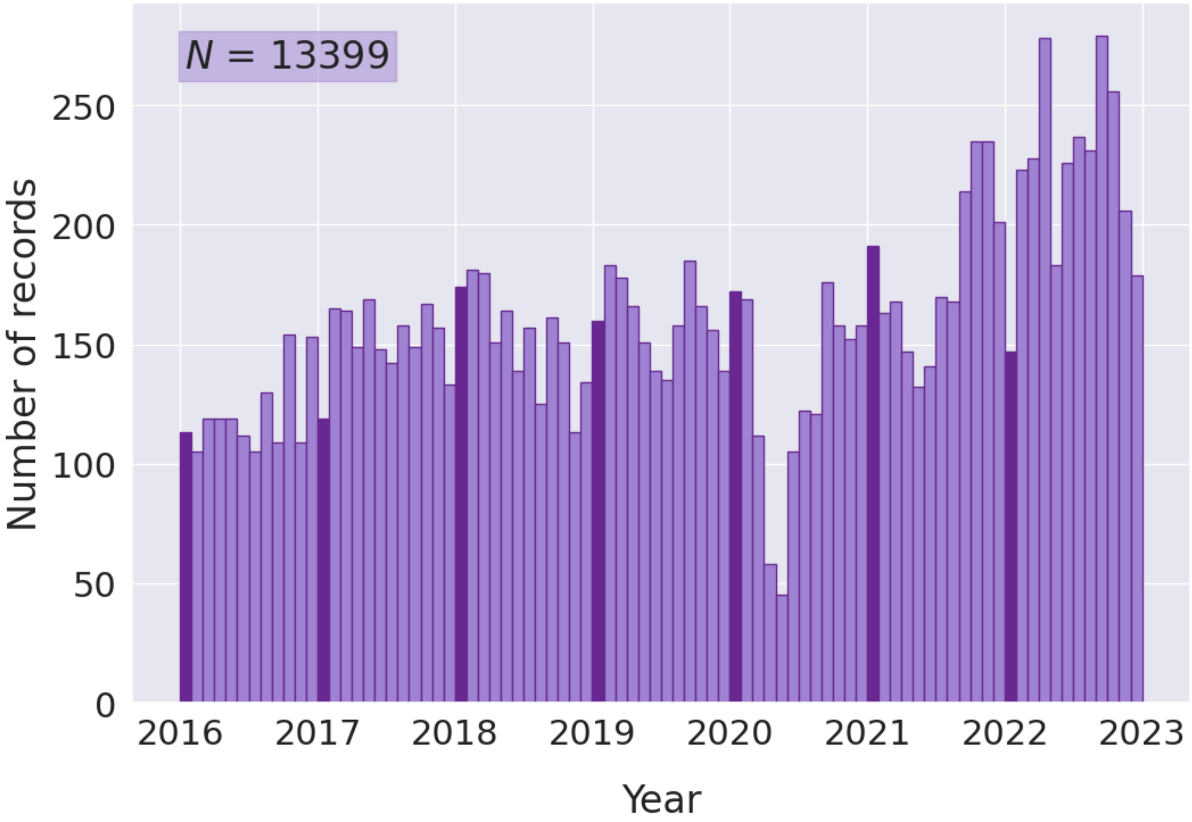


**Figure 05: Distribution of SINAN records over time (January is highlighted for each year)**

Those two relatively simple distributions already provide essential information about the data. The distribution of violence notifications is somewhat stable over time. However, the primary care visit distribution is not. An important takeaway is that, when analyzing a more recent case of GBV, the model will likely have more input data on primary care visits. It is also possible that violence cases from the first years of the time period of this project, namely 2016 and 2017, will be less likely to be identified in the e-SUS AB records, considering how the system was much less adopted at the time. The COVID-19 pandemic can also impact the model, considering the decrease in entries during this period.

It is also important to note that the volume of records depends on the history of implementation of each system. The e-SUS AB is an information system strongly influenced by the characteristics of each municipality, such as socioeconomic factors, geographical location, population density, urbanization, level of computerization, etc. In 2013 the Ministry of Health proposed to offer a new health information system to meet the different computerization and organization needs of municipalities, aiming to modernize the technological platform, with support for care management, optimization of data collection, interface with the different systems used for basic care and improving the level of detail in health records. In this process, local realities have a direct impact on the completeness of the system.

SINAN Violências, on the other hand, was implemented in 2006. Since this date, municipalities have been receiving resources and training to register more and more qualified information about violence in the country. Recife has a good healthcare system, considering it is a capital city with a high GDP when compared to other municipalities in the region. All this helps explain why there is a stabler number of records in SINAN when compared to e-SUS AB in the same period.

## Distribution of e-SUS AB records per individual

An important aspect of e-SUS data for the success of this project is how frequent primary care visits are in comparison to other events such as violence notifications, hospitalizations, and death. However, although the notification is a rare and unique event (in most cases, one episode of violence is notified once), violence is not. Gender-based violence can assume many forms - psychological, sexual, physical, etc. - and can be episodic or ongoing.

Domestic violence tends to follow a cyclic pattern: it starts with small, but constant conflicts that increase until an episode of high tension, permeated by insults, humiliation, intimidation and threats. In time, the threats start to materialize into physical aggression. After the violent episodes, comes a “honeymoon” phase, in which the aggressor may apologize or promise to improve and never do it again. However, the episodes continue and tend to become more serious with time (LUCENA et al., 2016).

One of the premises adopted for this project is that, albeit subtle, some signs related to GBV can be found in electronic medical records, based on the patient's narrative and/or the health professional's observations. Intuitively, the more information we have on a patient’s history, the more those signs are likely to appear. For the success of this type of approach, it is necessary to understand the amount of information that can be expected from a given individual and how that information is distributed over the individual's history and in relation to the whole population. To visualize how the information is distributed over the entire population, outliers were removed (people with more than 35 records) and the counts of e-SUS records per person were plotted as a histogram, represented in Figure 06. The figure shows two separate distributions, one for the whole population and another for women with a SINAN record. Each bar represents a number of entries for a given person. For example, the first bar represents all women with a single primary care visit record, totaling almost 30% of the population when looking at all women.

It is easy to see how different the full population is from the group of women who have been reported as victims of GBV. While the former has a lot of individuals with few e-SUS AB records, in the latter, only around 16% of women have a single registered visit to primary care units. The decrease is also smoother for women who are known victims of GBV, meaning that these women tend to have more records in e-SUS AB. This reinforces the hypothesis that women suffering from GBV deal with this problem for longer periods of time and that it has an impact on their health overall, which can be identified by grievances shown in routine appointments.

This characteristic of the distribution is positive for the main goal of this project because it shows that, on average, there's more data about victims than random individuals from the full population. However, this could also mean that it is easier to use this data for training, but when working with new data, it might be hard to identify possible cases of GBV because some individuals have a very small number of visits to primary care units.

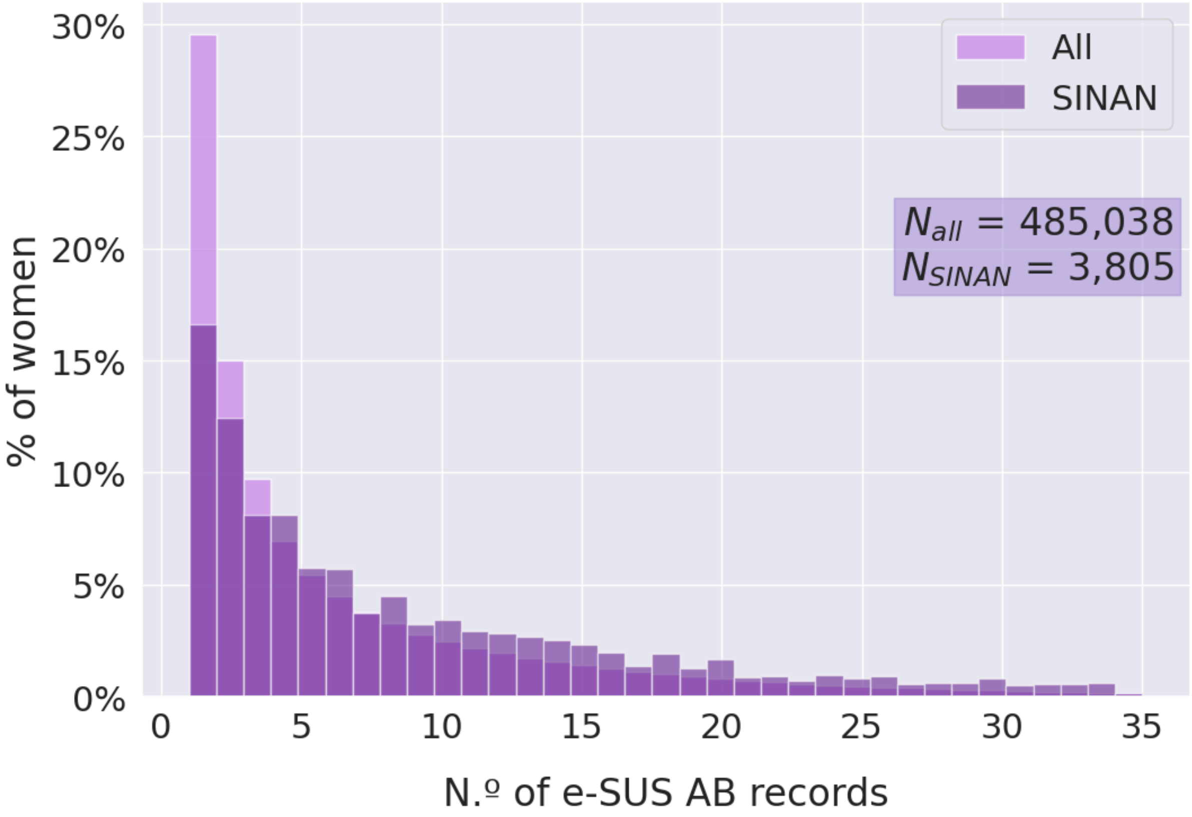


**Figure 06: Percentage of women with different number of e-SUS records**

This issue is even more obvious when the same data is plotted as a cumulative frequency chart, like in Figure 07. More than 50% of all women have five or less records in e-SUS AB. Depending on the distribution of those records over time, it could be hard to make inferences about GBV characteristics that are manifested over time, instead of a single episode.

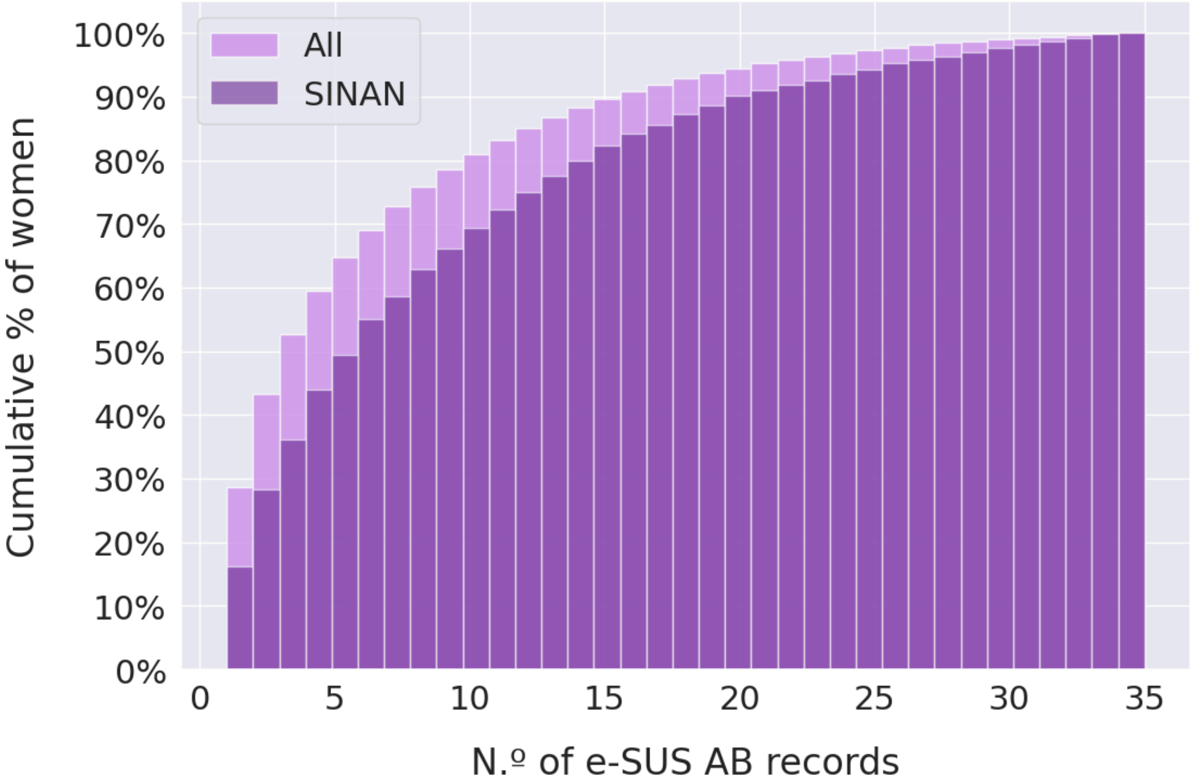


**Figure 07: Cumulative percentage of women with different number of e-SUS records**

Another important observation about the data in Figures 06 and 07 is that it shows that victims of GBV have more primary care visits, but not why. Possible explanations for that are: (i) frequent visits to the doctor could indicate victims' attempts to seek help; (ii) GBV directly increases the need of medical care; (iii) GBV indirectly increases the need of medical care, by deteriorating women's health; and (iv) victims who visit Primary Healthcare Units more often are more likely to be identified and reported by health professionals.

If explanation (i) or (iii) hold true, then it is possible that certain cases can be identified before a would-be notification date. If, instead, only (ii) is true, then GBV cases may be more easily identified only after they occur. Finally, it is also possible that the main explanation for this difference is in the notification process, and not in GBV itself. In that case, women that suffer GBV with more primary care visits have more events in which the violence can be brought up and/or be identified by health professionals. Thus, the notification likelihood increases. A more in-depth study is required in order to draw a more refined analysis from this data and evaluate these explanations.

## Primary care visits in relation to date of violence

One of the assumptions made for the development of this project was that signs of GBV can be observed in medical records from primary care visits. Finding exactly what those signs are and how frequent they are is the main goal of this project. However, other aspects of the data can serve as proxies to that. In the previous section, we argued that having more information on an individual could increase the chances of identifying characteristics related to GBV. Here, this idea is extended by asking the question: is there an increase in primary care visits near the date of GBV notifications?

This question can be answered using the linked data and temporal information of e-SUS AB and SINAN records. First, all e-SUS records belonging to the same victim are grouped together. For each of those records, the date of the primary care visit is used to compute the difference in days between it and the violence episode. For individuals with more than one SINAN notification, medical visit records have multiple differences, one for each notification.

After that, a weight is assigned to each (e-SUS, SINAN) record pair. The weight is based on the date of the primary care visit. This is an important step to account for the increase of e-SUS AB records over time, discussed in subsection 2.1. Records in earlier years have more weight simply because they are less frequent. For that reason, when discussing this data, there are no absolute count values, only weighted ones.

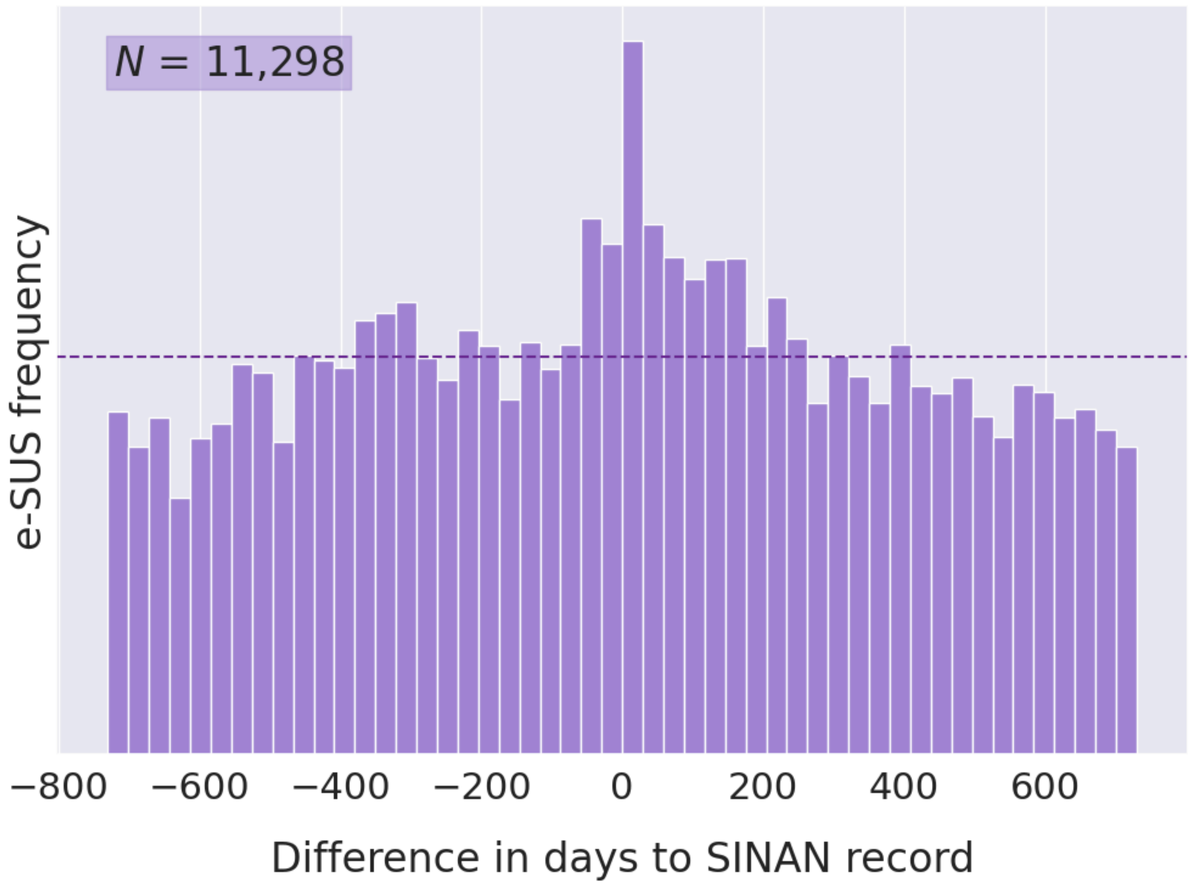


**Figure 08: Frequency of e-SUS records by difference (in days) to date of violence episode**

Finally, these weighted pairs are used to build a histogram (Figure 08). It represents the distribution of e-SUS AB records in terms of their temporal difference to the violence notifications. In total, 11,298 (e-SUS AB, SINAN) record pairs were used. The bar at the 0 mark in the x-axis represents the exact day of violence, while negative ones indicate earlier days and positive ones, later days. The dotted horizontal line represents the average of the distribution and is useful for comparison. Spikes considerably over that line indicate an increase in medical visits in that timespan before or after the notification.

The maximum difference shown in the plot is two years before or after the notification. Only records between 2018 and 2020 were considered to prevent the graph extremities from being affected by lack of data: for example, a violence notification in 2016 would have way less past medical record information simply because the earliest date in the e-SUS AB dataset is 01/01/2016. The same also applies to those notifications near the end of 2022.

The highest point of the distribution is near the day of violence. This information is expected and shows that, at least in that time span, there's a strong correlation between SINAN notifications and e-SUS records. It provides evidence that GBV cases could be identified based on medical records. If the distribution was uniform-like, then violence characteristics in electronic medical records would be even harder to identify.

Even more interesting is how this increase in visits is not limited to days very close to violence episodes. Figure 08 shows that there is a systematic increase in the number of visits to the doctor from around 60 days before the SINAN notification to 200 days after. This type of insight is useful because it limits the project's search space (at least for the initial analysis) to a time span where it is known that the number of e-SUS AB of the population increases. In more practical terms, these 260 days represent a window where GBV becomes more apparent in public health systems, and thus easier to identify.

Another region with higher points on the graph is the one close to day 400. These will have to be carefully analyzed in the future.

One important thing to note is that, despite showing the medical care visits before and after the notification as different types in the plot, there's no evidence to make concrete claims about them at this level of analysis. For example, while it is possible that e-SUS records before the notification represent victims' attempts at seeking help, it is also possible that at that time, the patient has already suffered violence, but of lower risk, and thus, the episode was not notified. That is, day 0 in this analysis does not represent the only possible day of violence, especially considering SINAN's underreporting. The goal here is to merely show the rippling effect of the notification itself, not of the actual GBV. By measuring how far it goes (the 260 days), we can make more informed decisions about the final model.

## Identifying relevant textual patterns for GBV identification

One of the primary goals of this project is to provide insights about GBV for policy makers. We have discussed the complexities of finding causality between data points and showed that there's a certain time span where those causal relations could be easier to identify. However, a more important insight for those workers at primary care units is what types of conditions, events, and symptoms they should be looking for to help women in vulnerable conditions.

One of the central hypotheses of this project is that this fine-grained type of information can only be obtained by considering parameterized and text fields. For instance, in medical records from e-SUS AB, health professionals can classify the patient's condition using the International Classification of Diseases (ICD) codes. This is an important piece of information, but often it does not describe all symptoms, causes, and consequences. This is where the patient's history and the health professional observations, both stored as text, can provide a much needed level of detail.

In this section, we first present an analysis to answer the question "*Can semantic features be used to distinguish e-SUS AB records related to violence from others?*". We then present the methods used to train a classification model to identify patterns .

## Visualizing different types of e-SUS records

The idea that e-SUS records of victims subject to GBV are considerably different from those not under that condition is sensible from a purely informational point of view. However, when analyzing data from real use of health information systems, data entry can be impaired by a multitude of external factors. Health professionals may be working under stressful and demanding conditions, users may lack the proper training to use the systems and, very importantly, there is the fear of reprisal when dealing with sensitive situations such as GBV. These are all factors that can't be controlled, but should be accounted for when working with this kind of data.

Considering this context, a separate analysis was conducted to evaluate the feasibility of training a classification model for this type of data. First, the different classes of e-SUS records had to be formally defined in order to separate cases we know are associated with violence and cases that are not. As previously mentioned, causality is hard to establish between the records of the same individual, and for that reason, some of these definitions partially rely on approximation rules. These rules can be expanded and updated as the project progresses. So far, e-SUS records were separated into four different groups:

1. **violence:** any record with an ICD code for aggression or within two days of a SINAN notification or SIH/SIM record with the same code;
2. **not violence**: certain ICD codes that have a small probability of being associated with violence, e.g. COVID-19, parasitic diseases, tumors, and some congenital malformations;
3. **likely violence**: any record within 30 days within of a notification of violence that doesn't have an ICD code for aggression;
4. **undefined**: any record that does not fall into one of the previous categories. Not all undefined records are equally likely to be related to violence and this should be more explored in the future.

With those classes properly defined, we can check how well they are separated from one another. Because each record, after the application of a Principal Component Analysis - PCA, is represented by a vector with four thousand dimensions, a dimension reduction needs to be used to plot the data. The two most popular techniques to do that are t-distributed stochastic neighbor embedding (t-SNE) and Uniform Manifold Approximation and Projection (UMAP). We opted for UMAP because its embedding initialization tends to yield better results. After transforming and sampling the data, the 2D plot shown in Figure 09 was generated. These two dimensions are not meaningful for other analyses in this project, they just express how similar or different the two data points are.

This projection shows that indeed, e-SUS records that received the **violence** label are restricted to a specific region of the graph. The only significant overlap of classes is that between **violence** and **likely violence**, which is expected and shows that our inference about the relationship between records based on the date of each record is relevant. It also demonstrates how rules can be used as an initial filter for the early identification of GBV.

Those insights affirm the hypothesis that there is a substantial difference in those e-SUS records of victims of GBV and that early identification is feasible. However, the projection also makes explicit some of the challenges when working with this data.

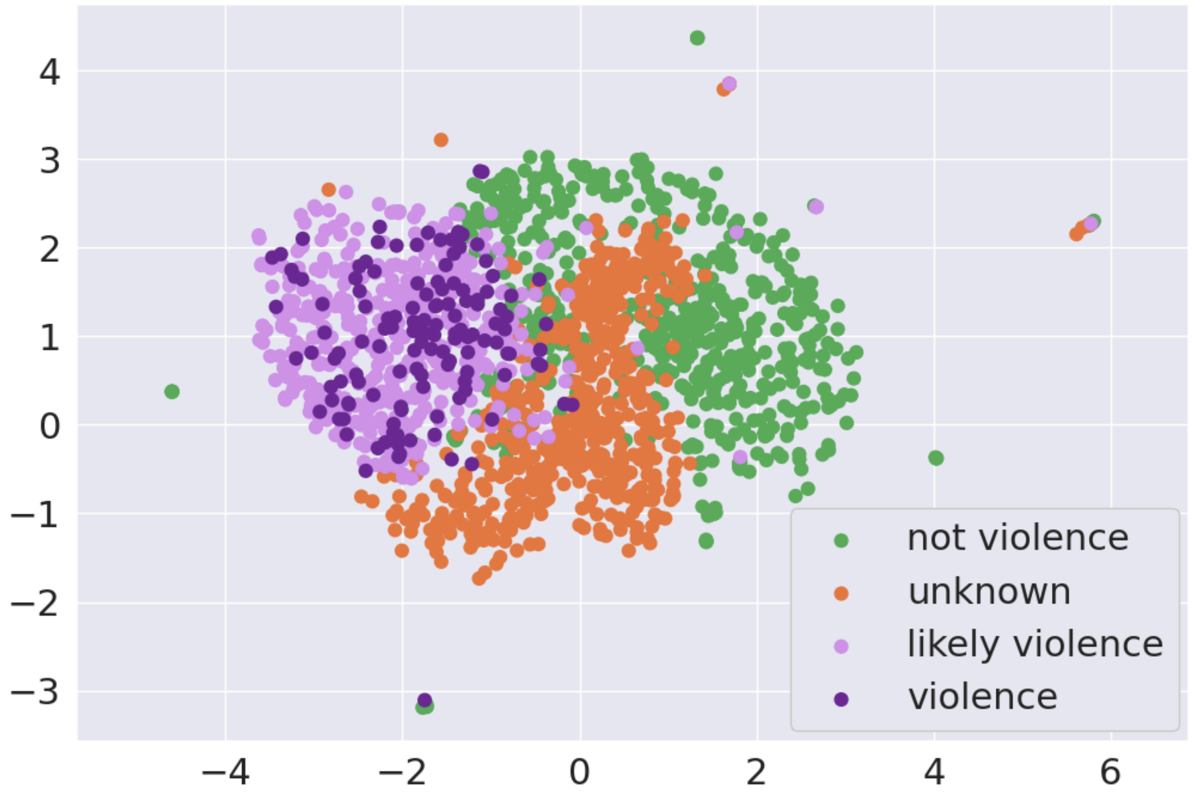


**Figure 09: UMAP projection of e-SUS records color-coded by class**

First, cluster shapes are irregular, especially when looking at the **not violence** and **unknown** classes. Second, but also very important, there is no significant variation in density on the projection. That is, the clusters in the projection are not in separate regions of the space and are very concentrated. Since many clustering algorithms assume that clusters are either convex, isotropic, or dense, the choice of techniques is limited. In fact, some traditional clustering algorithms were evaluated in a preliminary analysis and none performed well.

The fact that data clusters are not easily separable makes sense because of the text genre of an electronic medical record. Different conditions may share symptoms and treatments, and health professionals may use similar language structures to describe different situations. Naturally, some records may represent more prototypical cases of GBV or a medical condition, but, near the cluster boundaries, these linguistic and medical differences are harder to define. It highlights the fact that not all records can be easily separated into violence or not, there is a gray area that is difficult to unveil. These boundaries represent data points that may require additional professional input in order to qualify the information and extract the most value out of these records.

## Ranking feature importance

With the information that the different classes of e-SUS records represent different regions in the UMAP projection space — and by extension in the principal component space — a classification model was trained to distinguish between violence cases and others. This model can be used to classify new records, but is also useful to evaluate feature importance.

In the context of this project, feature importance refers to how significant each FrameNet entity is to the model to evaluate an e-SUS record. Note that the classification model takes the principal components of those semantic features as input. However, the principal components can be "reverted" back to the original FrameNet entities and thus, one can estimate the importance of these interpretable features.

Not all machine learning models allow for easy estimation of importance. For this first analysis, we decided to look for insights using a Support Vector Machine (SVM) model. Linear SVMs have relatively few hyper-parameters and because they are linear, can be easily used to estimate importance.

Before training the final model, a grid search was used to find the best values for **C** and the kernel type. The selected value of **C**=50 and a linear kernel were chosen based on their recall and F-score. Recall was the main scoring method because a high recall value indicates the model is able to identify most cases related to GBV (even if with more false positives). The notion behind it is that it is better to have false positive cases for GBV than false negatives. The F-score was only used to identify the point of diminishing returns for recall increase.

For the final training, only records classified as **violence** or **not violence** were considered. The other two classes were not included because they could bias the model. Therefore, the training dataset was significantly smaller, consisting of only 408 examples, split into the two classes. The **not violence** class is significantly larger, but it was undersampled to prevent biases. Because the model only had four false positive records and one false negative in the test set, we decided to compare how it evaluates **unknown** and **likely violence** records in comparison to manual evaluation. Table 07 shows the results of this evaluation.

**Table 07: Model accuracy on e-SUS records labeled as likely violence and unknown**

| Class | No. of records | violence prediction (model) | violence evaluation (manual) |
|---|---|---|---|
| likely violence | 26 | 24 (96%) | 12 (46%) |
| unknown | 136 | 74 (54%) | 4 (3%) |

These results show that the model still requires further calibration: because recall was used to select the best parameters, the model ends up labeling a significant number of cases as possibly related to GBV. Although it is important to identify as many records that may be related to GBV, precision must achieve a reasonable threshold for the model to be useful for those monitoring it.

Of the **likely violence** records incorrectly labeled by the model as violence, a considerable amount was related to routine care visits for infants/children. While these women have a SINAN record, human evaluators found no signs that could raise GBV suspicions in their medical records. This seems to be a form of overgeneralization by the model, wherein it mistakenly linked semantic structures associated with motherhood to indicators of GBV. This issue will have to be further investigated.

To better contextualize these results and to understand what features are most important for the model, health professionals and policy makers, we use the model feature importances to compute a rank of features. Figure 10 shows the estimated importance of the top 15 most relevant semantic features. The larger the bars, the more relevant a feature is to identify GBV cases. Note, however, that the features do not necessarily correlate with the **violence** class: an important feature may strongly correlate with the **non violence** class, and because of that it is also relevant.

Interestingly, the most important feature for the model was the lexical unit for "*laboratory exam*", which, being a commom demand in routine medical care, could be related to non-GBV records. However, this had to be explored further, considering that GBV cases could also demand exams. The other four relevant features seem to be related to medication: the lexical unit for "*prescription*" and "*administer*", the frame `Medicines`, and that same frame appearing as the action of another frame.

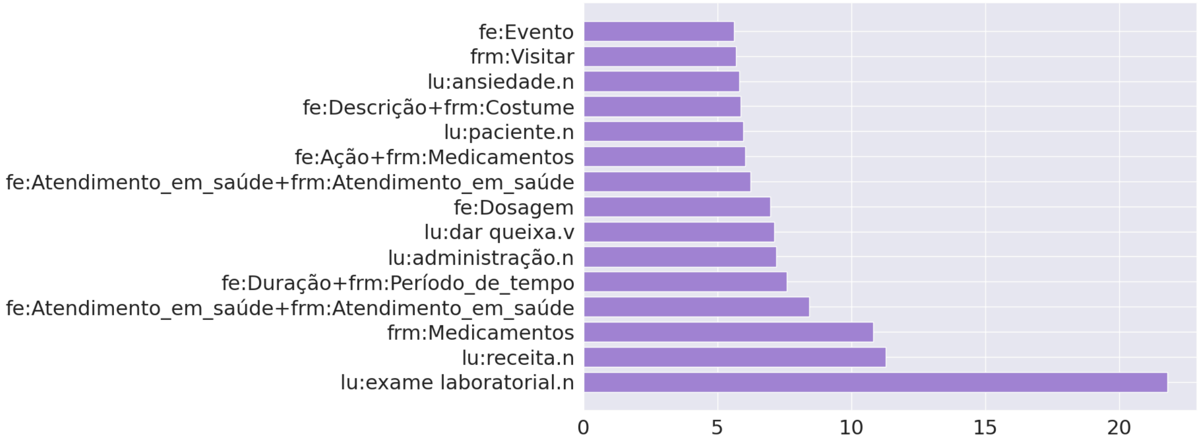


**Figure 10: Top-15 features for the SVM models, ranked by importance.**

One interesting feature that also appears among the 15 most important is the lexical units for "*anxiety*", which is the LU most likely to be related directly to violence. The lexical unit for “*complaint*” (“dar queixa”) requires further investigation, since it can be associated with complaining about a health condition or filing a complaint about violence to the police, filling a police record.

In this section, we have shown some of the most important features for the model to classify e-SUS records. Since the list is very long because each of the thousands of features has an important value, an interesting alternative is to search specifically for certain features. Those could be situations or symptoms that the medical literature has shown to be related to GBV, or others where more evidence is needed.

Due to its complexity, the model needs constant improvement with permanent input from gender, health, and violence specialists in order to reach its full potential in recognizing patterns of GBV. The Accelerator was an opportunity to explore the potential of this data, but the work is ongoing. In addition, a dashboard is being built to visualize the analyses and serve as a tool for local public managers.

# Takeaways and future work

This project aimed to identify patterns associated with GBV in medical records with a goal to help health professionals identify early signs of violence through signs and symptoms not usually associated with GBV. From the data analyzed, the main insights so far have been:

- From the analysis of the integrated SINAN and e-SUS databases, we noted that, generally, women have few e-SUS AB records, while among women with a violence notification, only around 16% have a single registered visit to primary care units. This means women who are victims of GBV tend to have more records in e-SUS AB and visit primary healthcare units more often.

- Distributing the records over time, we noticed there is a strong correlation between SINAN notifications and e-SUS records, providing evidence that GBV cases could be identified based on medical records.

- There is a systematic increase in the number of visits to the doctor from around 60 days before the SINAN notification to 200 days after.

- Evaluating patterns between clusters of records, we noted that e-SUS records that received the **violence** label are restricted to a specific region of the graph. The only significant overlap of classes is that between **violence** and **likely violence**, validating the hypothesis about the relevance of the dates for the relationship between records.
- These findings suggest that rules can be a useful first filter for the early identification of GBV, reinforcing the hypothesis that there is a substantial difference in those e-SUS records of victims of GBV and that early identification is feasible.

The data analysis approach adopted in this project will serve as a baseline for future studies on GBV developed by Vital Strategies Brasil and FrameNet Brasil. The advantages of the current approach can be further explored, while some of the shortcomings need to be addressed in order to obtain more insights.

In a more technical sense, our experience with this data approach showed that by automatically annotating frames and elements it is possible to find patterns related to GBV. Moreover, the integration of certain categorical fields with the texts (such as ICD code and sociodemographic information), can be productive. It highlights how in future projects, the combinations of categorical and text data can improve the quality of the analysis.

The approach also demonstrated the importance of an exploratory analysis of data prior to working with the model. This type of analysis is important for any future work that may rely on linked data and in the causal relation between those data points. Another aspect in which our exploratory analysis was very effective was in better understanding how different objects of analysis are manifested. For example, this preliminary work was essential to understand that although different from other cases, GBV cases may have a lot in common with other objects, especially considering the edge cases of those groups.

Finally, the active involvement of  health and violence specialists in each step of the project was essential to alleviate some of the technical shortcomings. For example, the model for GBV classification assigns importances to thousands of variables. With specialized knowledge, it is possible to more easily navigate that space to find useful insights. That expertise will also be useful to further explore classification methods, such as clustering with constraints and other machine learning models. For other projects working with complex data as this one, active participation should always be considered.

Although some of our insights in this project may be constrained to the Brazilian reality and how it interacts with the gender-based violence phenomenon, certain lessons we learned could be useful for other organizations dealing with similar challenges. We have highlighted how our analyses were only possible when data from different public health systems were linked into a single dataset. This is an important takeaway for teams and organizations that plan on working with phenomena where important evidence will be spread into different data sources.

Moreover, textual data could only be effectively analyzed after appropriate modeling. We have shown a FrameNet-based model and annotation schema can be leveraged to not only classify cases, but also understand how information in that schema is used for that end.

The insights brought by this project can subsidize further research in the fields of gender-based violence and its health consequences. Analyzing data on medical visits made by women with a notification of violence can help us further our understanding on health consequences brought by living with violence. The time patterns between records can deepen our knowledge on the violence cycle and its consequences.

Identifying that finding a pattern between health conditions displayed in medical records and violence is also an important breakthrough and, in advancing this research, can help build tools for the early identification of violence by health professionals working throughout the public health system.

# References

Costa, Alexandre Diniz. (2020) “A tradução por máquina enriquecida semanticamente com frames e papéis qualia.” (Ph.D. thesis in Linguistics. Universidade Federal de Juiz de Fora, Juiz de Fora.)

Dutra, Lívia et al. Building a Frame-Semantic Model of the Healthcare Domain: Towards the identification of gender-based violence in public health data. In: Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana (STIL), 14, 2023, Belo Horizonte/MG. Anais [...]. Porto Alegre: Sociedade Brasileira de Computação, 2023. p. 338-346. DOI: https://doi.org/10.5753/stil.2023.234145.

Fillmore, C. J. (1982). Frame semantics. In: Linguistic Society of Korea (ed.), “Linguistics in The Morning Calm”. Seoul: Hanshin, p.111-138.

Lucena, Kerle Dayana Tavares de, Deininger, Layza de Souza Chaves, Coelho, Hemílio Fernandes Campos, Monteiro, Alisson Cleiton Cunha, Vianna, Rodrigo Pinheiro de Toledo, & Nascimento, João Agnaldo do. (2016). Analysis of the cycle of domestic violence against women. Journal of Human Growth and Development, 26(2), 139-146. https://dx.doi.org/10.7322/jhgd.119238

Oliveira, Gisele Pinto de; Bierrenbach, Ana Luiza de Souza; Júnior, Kenneth Rochel de Camargo; Coeli, Cláudia Medina; Pinheiro, Rejane Sobrino. Accuracy of probabilistic and deterministic record linkage: the case of tuberculosis. Rev. Saúde Pública (50), 2016. https://doi.org/10.1590/S1518-8787.2016050006327

Pustejovsky, James.(1998) The generative lexicon. MIT press.

Torrent, Tiago Timponi et al. (2022) “Representing context in framenet: A multidimensional, multimodal approach.” In: Frontiers in Psychology, v. 13. doi: 10.3389/fpsyg.2022.838441