# Dataset Description

## Dataset Overview

**Organization Name:** Soroka

**Number of Records:** 1000

**Data Domain:** Oncology

## Cohort Characterization

**Number of Distinct Patients:** 1000

**Medical Indications Covered:** breast cancer patients from soroka

## Text Content Details

**Minimum Text Length (total terms):** 26

**Average Text Length (total terms):** 147.19

**Maximum Text Length (total terms):** 1239

## Dates

**Date Range of Records:** 2015-01-26 to 2024-11-04

## Data Source

**Original Data Source:** free text fileds, Camillion system(מחלה\_נוכחית ,סיכום\_המחלה, בדיקה\_גופנית ,דיון\_תכנית\_והמלצות)

## Language Analysis

**Hebrew Count:** 523

**English Count:** 27

**Mixed Count:** 450

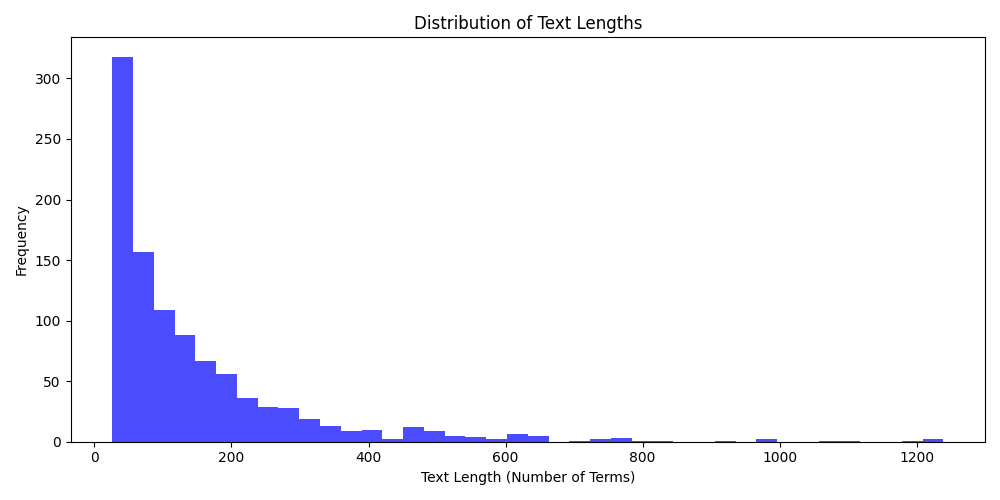
**Hebrew Ratio:** 52.30%

**English Ratio:** 2.70%

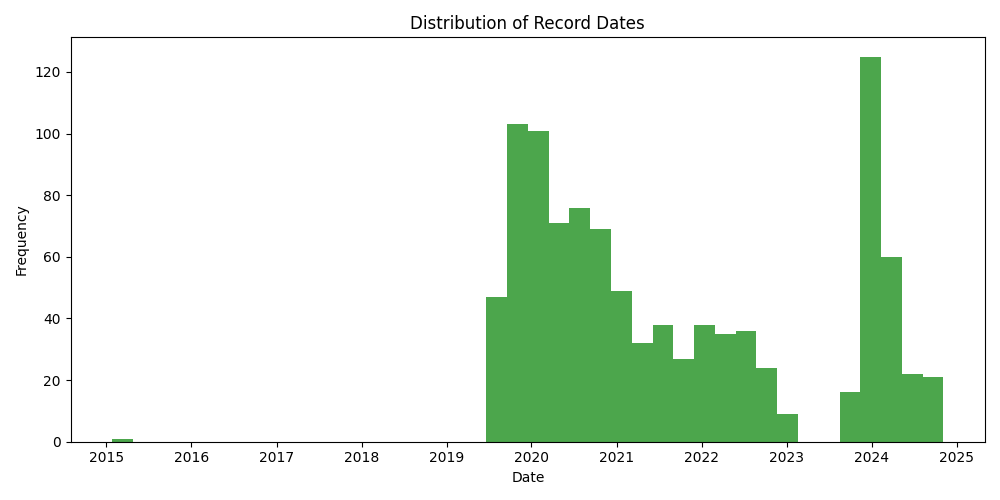
**Mixed Ratio:** 45.00%

*Note: Mixed texts are classified as such when more than 10% of the text words belong to a secondary language.*

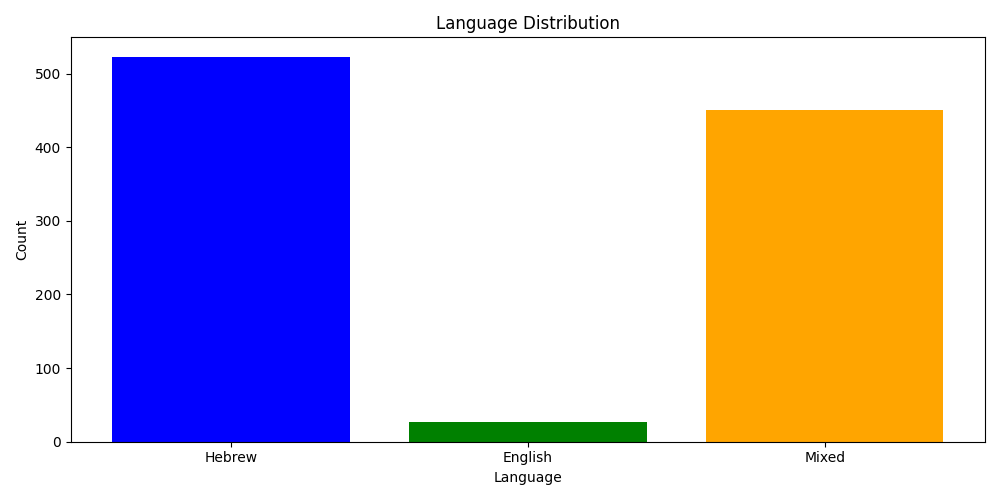
## Distribution of Text Lengths



## Distribution of Record Dates



## Language Distribution



*Note: Mixed texts are classified as such when more than 10% of the text words belong to a secondary language.*