Stanford MEDICINE

Stanford | Center for Health Education

INTRO TO CLINICAL DATA STUDY GUIDE

MODULE 5 - HANDLING UNSTRUCTURED HEALTHCARE DATA: TEXT, IMAGES, SIGNALS

LEARNING OBJECTIVES

- Describe the ways in which clinical text is different from natural language
- Describe how simple text mining strategies can be as effective at NLP when applied to clinical text
- What are negation and context detection, and why are they important?
- Explain how de-identification is different from anonymization
- Describe the important properties of healthcare images
- Describe the important properties of healthcare signals

UNSTRUCTURED DATA

This module focuses on text, images, and signals. These are all called **unstructured data** in contrast to **structured data**, the rectangular tables found in databases.

- Unstructured data: text, images, and signals
- Structured data: the rectangular tables found in databases

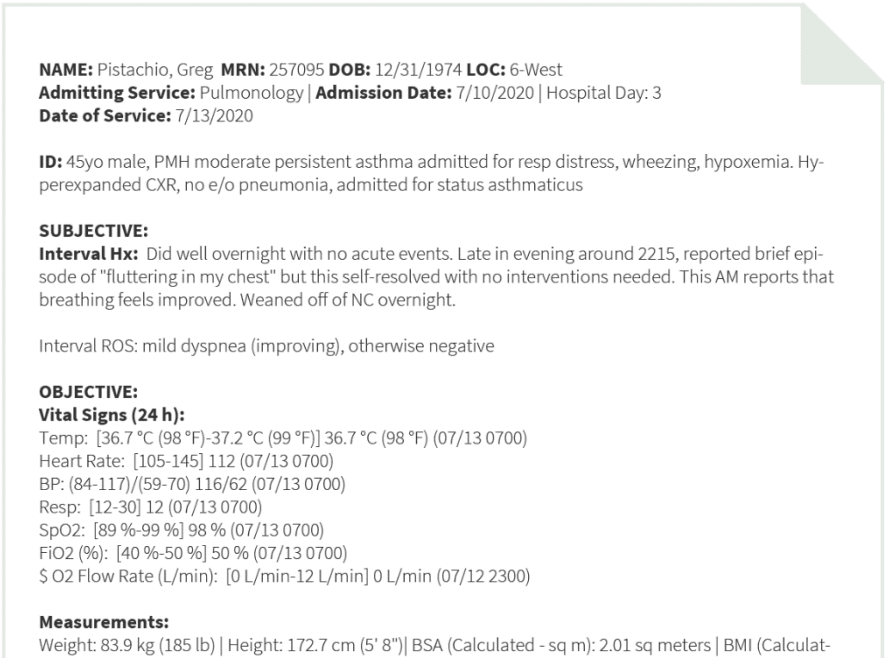
CLINICAL TEXT

Clinical text is different from ordinary natural language and the language used in scientific publication.

It is:

- Written by clinicians or other healthcare providers for clinicians and those documenting the services provided
- Used to describe patients, their pathologies, their personal, social and medical histories, and findings made during interviews and procedures

- Not written for publication and may not use full sentences



NAME: Pistachio, Greg **MRN:** 257095 **DOB:** 12/31/1974 **LOC:** 6-West
Admitting Service: Pulmonology | **Admission Date:** 7/10/2020 | Hospital Day: 3
Date of Service: 7/13/2020

ID: 45yo male, PMH moderate persistent asthma admitted for resp distress, wheezing, hypoxemia. Hyperexpanded CXR, no e/o pneumonia, admitted for status asthmaticus

SUBJECTIVE:
Interval Hx: Did well overnight with no acute events. Late in evening around 2215, reported brief episode of "fluttering in my chest" but this self-resolved with no interventions needed. This AM reports that breathing feels improved. Weaned off of NC overnight.

Interval ROS: mild dyspnea (improving), otherwise negative

OBJECTIVE:
Vital Signs (24 h):
Temp: [36.7 °C (98 °F)-37.2 °C (99 °F)] 36.7 °C (98 °F) (07/13 0700)
Heart Rate: [105-145] 112 (07/13 0700)
BP: (84-117)/(59-70) 116/62 (07/13 0700)
Resp: [12-30] 12 (07/13 0700)
SpO2: [89 %-99 %] 98 % (07/13 0700)
FiO2 (%): [40 %-50 %] 50 % (07/13 0700)
$ O2 Flow Rate (L/min): [0 L/min-12 L/min] 0 L/min (07/12 2300)

Measurements:
Weight: 83.9 kg (185 lb) | Height: 172.7 cm (5' 8")| BSA (Calculated - sq m): 2.01 sq meters | BMI (Calculat-

What makes clinical text valuable:

- It can augment the billing codes that have been assigned to medical records
- It contains a description of what happened to the patient
- It is used for biosurveillance to detect and monitor disease outbreaks
- It is used for improving the set of words that are used to refer to diseases
- It is used in clinical decision support
- It is used to add codes to the patient's medical record for querying and reporting

The value derived from clinical text may depend on the medical condition under consideration. Some conditions are accurately represented by codes. In other cases, a significant fraction of the patients would only be identified through the analysis of the text in the clinical notes.

Clinical text can enable clinical research by facilitating the construction of study cohorts. It has even been used as part of an automated procedure to discover new knowledge by mining the text for particular patterns.

There are many important challenges to using clinical text:

- Be ungrammatical.
- Have misspellings and concatenations

- Contain short telegraphic phrases, acronyms, and abbreviations
- The quality of the sources can vary widely

A very important problem in analyzing clinical text is the problem of **negation**.

- **Negation**: Refers to the use of a phrase to indicate that the patient does not have a condition
- The problem of negation: the analysis of clinical text needs to detect when a term mentioned is inside a negation. Roughly 40% of the content of clinical text is stated as a negation.

A related problem is called **context**.

- **Context:** Refer to a condition that a patient had before, or that a patient's family member has or had
- The analysis needs to detect the context of a term mentioned

Finally, there is pervasive fear, misunderstanding, and confusion around security, privacy, anonymization, and de-identification. This has artificially increased the burden to obtaining data access to clinical text.

PRIVACY AND DE-IDENTIFICATION

Textual data can contain Protected Health Information (PHI), which cannot be shared without the patient's permission.

Terminology

- **Anonymization:** Data cannot be linked to a specific person
- **De-identification**: Removal of identifiers that constitute protected health information

Methods for de-identification:

1. Safe Harbor: Requires the removal of 18 specific items
2. Statistical Method: A statistician validates and documents that the statistical risk of re-identification is very small
3. "Hiding in plain sight": Replaces text that looks like identifiers with realistic-appearing surrogate information.

Visit the US Government's Health and Human Services website for more detail on statistical method and the safe harbor method.

NATURAL LANGUAGE PROCESSING

Natural Language Processing (NLP) is an area of artificial intelligence that develops methods to allow computers to process human language.

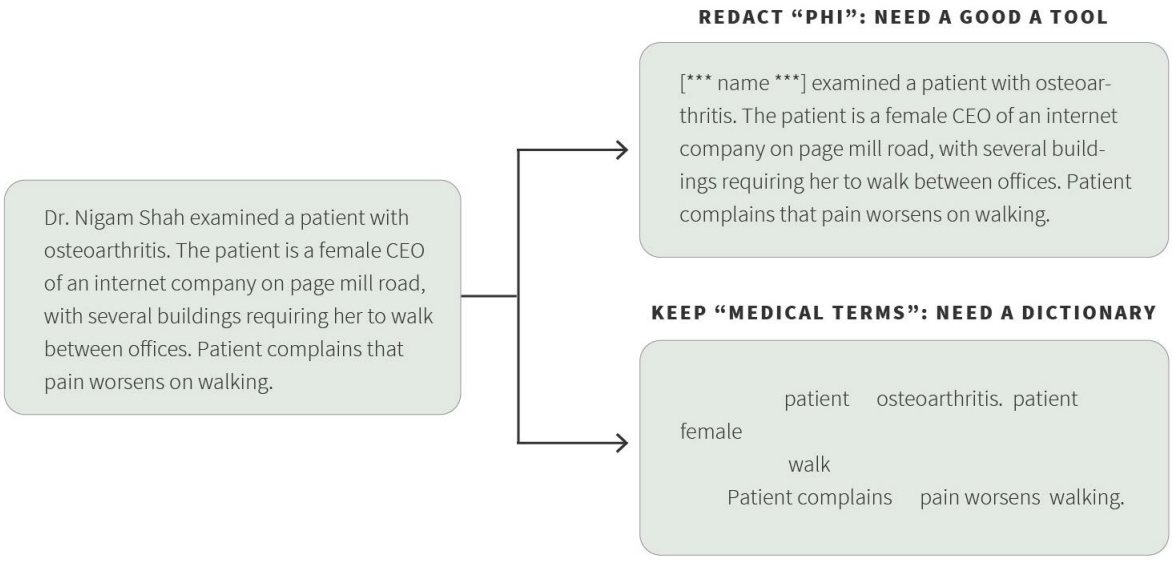
NLP consists of the following steps:

1. Tokenization: the purpose of this step is to detect words and identify where sentences begin and end
2. Parsing: determines the grammatical structure of each sentence and tags each word with its part of speech, such as noun or verb
3. Named entity recognition: identifies words or phrases of interest, and assigns appropriate category labels
4. Section detection: identifies boundaries between different sections of a document

When NLP tools are used on clinical text, they typically require some adaptation of the usual steps in an NLP system in order to work. The NLP tools need to address negation and context problems.

PRACTICAL APPROACH TO PROCESSING CLINICAL TEXT

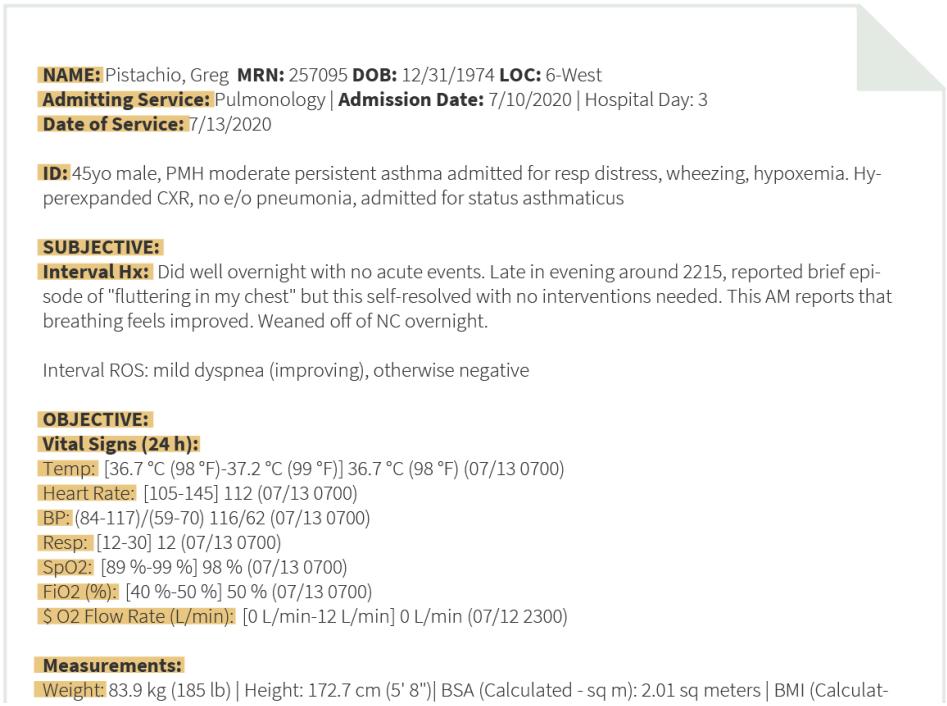
The ultimate goal is not to understand the contents of the document but to identify features in the text that will enable the downstream analyses that answer questions.



There are two main approaches for considering the issue of PHI:

1. Remove the PHI. This requires locating all the PHI in the clinical text. If you can do this accurately, then the PHI can be removed or skipped, and the desired features can be extracted from the remaining text.
2. Keep only medical terms that are useful for analysis downstream; all other terms will be passively filtered out. Use a **knowledge graph**, a dictionary of terms from a computed-based representation of medical information, to extract the features.
    a. A knowledge graph can also help you decide which terms are ambiguous
    b. An **ambiguous term** is a term with multiple meanings
    c. Another text processing problem is how to find similar terms that might be missing from your dictionary.

We described the important problem of detecting negation and context in clinical text. To detect negation and context in clinical text, there are software packages that have been designed, such as Negex and Context.



NAME: Pistachio, Greg **MRN:** 257095 **DOB:** 12/31/1974 **LOC:** 6-West
Admitting Service: Pulmonology | **Admission Date:** 7/10/2020 | Hospital Day: 3
Date of Service: 7/13/2020

ID: 45yo male, PMH moderate persistent asthma admitted for resp distress, wheezing, hypoxemia. Hyperexpanded CXR, no e/o pneumonia, admitted for status asthmaticus

SUBJECTIVE:
Interval Hx: Did well overnight with no acute events. Late in evening around 2215, reported brief episode of "fluttering in my chest" but this self-resolved with no interventions needed. This AM reports that breathing feels improved. Weaned off of NC overnight.

Interval ROS: mild dyspnea (improving), otherwise negative

OBJECTIVE:
Vital Signs (24 h):
Temp: [36.7 °C (98 °F)-37.2 °C (99 °F)] 36.7 °C (98 °F) (07/13 0700)
Heart Rate: [105-145] 112 (07/13 0700)
BP: (84-117)/(59-70) 116/62 (07/13 0700)
Resp: [12-30] 12 (07/13 0700)
SpO2: [89 %-99 %] 98 % (07/13 0700)
FiO2 (%): [40 %-50 %] 50 % (07/13 0700)
S O2 Flow Rate (L/min): [0 L/min-12 L/min] 0 L/min (07/12 2300)

Measurements:
Weight: 83.9 kg (185 lb) | Height: 172.7 cm (5' 8")| BSA (Calculated - sq m): 2.01 sq meters | BMI (Calculat-

Many types of clinical notes have section headings that can be useful in processing the text. One way to find section headings is to look for everything that occurs from the beginning of a line up to the first colon character.

Mining clinical text:

- Pre-process the entire collection of documents containing clinical text

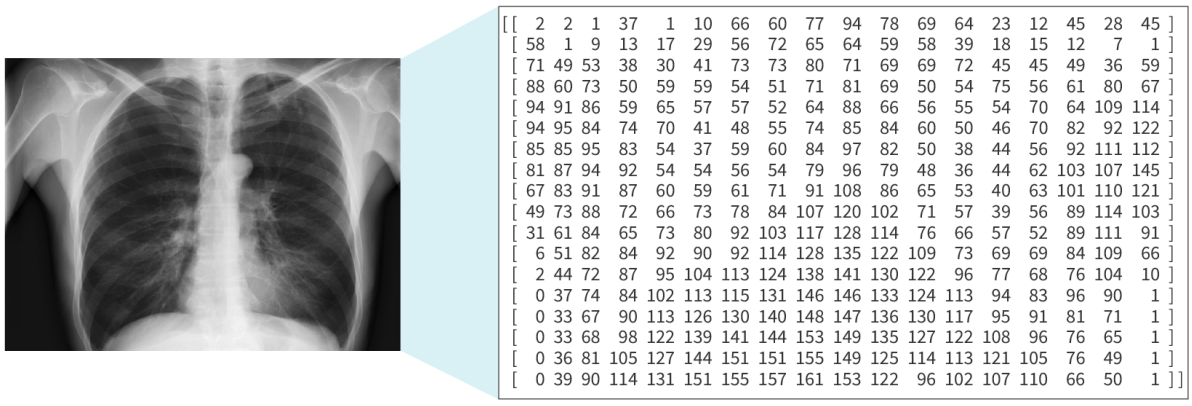
    - Use knowledge graph and negation/context detection to find important terms
    - The output of this step is indexed positive, present mentions of diseases, drugs, devices and procedures
        - Indexed: a record of what string was mentioned where
        - Positive: negations are omitted
        - Present: personal and family history of the condition are omitted
- Answer a clinical research question: Use the knowledge graph to find synonyms of terms relevant to that question; use the timeline to help resolve ambiguous terms
- Count present, positive mentions about the patient: Use the temporal information, the aggregated event and drug mentions, and contextual filters to create a patient-feature matrix and construct patient cohorts for further statistical analysis. Using the temporal information is crucial.

IMAGES

Medical images serve several important goals:

1. Diagnosis
2. Disease staging and response to treatment
3. Guiding surgical interventions

Images capture important details about anatomic structures and physiological processes. However, images are large and require interpretation by human or machine to turn the low-level image elements into meaningful features for downstream prediction tasks.



```
[[  2   2   1  37   1  10  66  60  77  94  78  69  64  23  12  45  28  45 ]
 [ 58   1   9  13  17  29  56  72  65  64  59  58  39  18  15  12   7   1 ]
 [ 71  49  53  38  30  41  73  73  80  71  69  69  72  45  45  49  36  59 ]
 [ 88  60  73  50  59  59  54  51  71  81  69  50  54  75  56  61  80  67 ]
 [ 94  91  86  59  65  57  57  52  64  88  66  56  55  54  70  64 109 114 ]
 [ 94  95  84  74  70  41  48  55  74  85  84  60  50  46  70  82  92 122 ]
 [ 85  85  95  83  54  37  59  60  84  97  82  50  38  44  56  92 111 112 ]
 [ 81  87  94  92  54  54  56  54  79  96  79  48  36  44  62 103 107 145 ]
 [ 67  83  91  87  60  59  61  71  91 108  86  65  53  40  63 101 110 121 ]
 [ 49  73  88  72  66  73  78  84 107 120 102  71  57  39  56  89 114 103 ]
 [ 31  61  84  65  73  80  92 103 117 128 114  76  66  57  52  89 111  91 ]
 [  6  51  82  84  92  90  92 114 128 135 122 109  73  69  69  84 109  66 ]
 [  2  44  72  87  95 104 113 124 138 141 130 122  96  77  68  76 104  10 ]
 [  0  37  74  84 102 113 115 131 146 146 133 124 113  94  83  96  90   1 ]
 [  0  33  67  90 113 126 130 140 148 147 136 130 117  95  91  81  71   1 ]
 [  0  33  68  98 122 139 141 144 153 149 135 127 122 108  96  76  65   1 ]
 [  0  36  81 105 127 144 151 151 155 149 125 114 113 121 105  76  49   1 ]
 [  0  39  90 114 131 151 155 157 161 153 122  96 102 107 110  66  50   1 ]]
```

Images are produced by the transduction of some kind of physical energy.

Each image is a two-dimensional rectangular array of values, where the value is a measure of signal intensity. Images can also be three-dimensional, with the third dimension corresponding either to time or to space.

There are several ways to categorize medical images:

1. By the imaging modality--typically based on the kind of physical energy that is being detected:
    - **Visible light**: Photographs or videos
    - **X-ray**: High-frequency electromagnetic radiation, which is differentially absorbed by bones, air cavities, fluids, and tissues. X-rays, CTs, and related imaging techniques use ionizing radiation, which can cause injury to the exposed tissue in a dose-dependent way.
    - **Ultrasound**: Uses sound waves with frequencies higher than can be detected by the human ear. These waves propagate through, and are reflected from, tissue depending on the tissue density. Ultrasound has an advantage that it does not harm the tissue.
    - **Magnetic resonance and nuclear medicine**: Involve the use of magnetic resonance, in which the magnetic properties of atomic nuclei are measured when brief electromagnetic pulses are applied. The density of the tissue can be calculated. The location of specific chemical tracers can also be found. Magnetic resonance also does not harm the tissue.
2. By structural versus functional:
    - **Structural images** capture the spatial location and organization of anatomic structures.
    - **Functional images** identify activity or change over time.

The most common images accessible for data mining are radiology images.

The life cycle of radiology images is:

1. Image acquisition
2. Image storage and management: The images are moved from the image acquisition device to a system that stores and organizes the images for retrieval. The storage system also records metadata.
    a. Digital Imaging and Communications in Medicine (DICOM), is a very widely used standard for the storage and transmission of medical images. DICOM applies to all the imaging modalities described earlier.
    b. Medical images are stored in Picture Archiving and Communication Systems (PACS). These representation standards and storage systems allow for image

Terminology

- **Wearables**: Electronic devices that use improved sensor technology to allow individuals to monitor physical activity, exercise and sleep

compression to reduce the amount of space occupied.

3. Processing and interpretation
    a. Traditionally, the content of a medical image is interpreted by a specialist and the text of this report is stored in the EMR.
    b. Research these days concerns the processing of medical images by computer with the goals to:
        i. Automate routine tasks, and highlight important features to ease the task of the radiologist or pathologists
        ii. Automatically produce an expert-level textual description of the contents of the image

SIGNALS

A signal is created by biomedical equipment that measures some physiological value, and transduces it into an electrical signal, usually a voltage. Voltage is then converted into a numerical value in digital format for computer analysis.

Signals are important because they provide continuous, real-time physiological. They are a time series of regularly-spaced values, converted into a digital format.

Signals are:

- composed of measured values at regularly-spaced intervals as determined by the sampling frequency of the sensor
- typically captured in clinical contexts in which continuous monitoring is important, such as in the intensive care unit
- important for wearable devices, such as those that track heart rate

The research goals in using signals are similar to that for images: automatic **feature detection** and automatic construction of an **interpretation**.

Most commercially available devices use proprietary algorithms for feature detection and interpretation of the signal

What are the major issues with using signals?

- We do not know how accurate algorithms are, or which features are crucial for their interpretations to be correct
- Accuracy of consumer devices is uncertain; there are risks associated with missing real health problems, and falsely reporting problems that do not exist
- The true value of these devices and apps to health has not yet been demonstrated
- Privacy concerns