

## Annotation Guidelines Identification of Smokers

The goal of this annotation task is to identify patients who are **tobacco smokers** from discharge summaries provided by hospitals. We will use this data as ground truth in training an automatic smoker classification system.

The summaries you will be working with are not original patient records: they have been modified in two important ways. First, all identifying information (names, dates, locations, etc) have been changed to ensure patient privacy. Second, all direct evidence of a patient's tobacco smoking status has been removed. This means that when you are reading a patient's discharge summary, remember that **absence of information about a patient's smoking status is not an indication that they are non-smoker.**

We would like you to give us your judgment on whether a particular patient is a tobacco smoker based on any information, including general knowledge of medicine and common sense, keeping in mind that direct evidence of the patient's smoking status has been purged from the record. You should use the following labels:

1. **SMOKER:** You should classify a patient as being a smoker if you believe that the patient is a **current or former smoker**.
2. **NON-SMOKER:** You should classify a patient as being a non-smoker if you believe that the patient **has never smoked**.
3. **UNKNOWN:** If you cannot determine the smoking status of a patient from their discharge summary, you should classify a patient as unknown.

In addition, if you label the record as either a **smoker** or a **non-smoker**, you should also add a confidence score of 1, 2, or 3, as follows:

1. A score of 1 indicates that you are weakly confident in your answer and should correlate with the phrase "Possibly"
2. A score of 2 indicates that you are confident in your answer and should correlate with the phrase "Probably" or "Likely"
3. A score of 3 indicates that you are highly confident in your answer and should correlate with the phrase "Very Likely"

You may use the first letter of the label (**S**, **N**, **U**) in your annotation. The following are examples of acceptable labels:

- a. **S3:** This patient is very likely a smoker.
- b. **N1:** This patient is possibly a non-smoker.
- c. **U:** I don't have enough information to decide one way or the other.