BIO Annotation Guidelines for Sleep Apnea

Generated with the help of ChatGPT, and modified.

1. Scope of Annotation

1. You have two labels:

- **B** ("Beginning"): Used for the first token belonging to an entity mention.
- **I** ("Inside"): Used for all tokens except the first belonging to an entity mention.

  No label, blank ("Outside") denotes tokens that are **not** part of an entity.

2. These guidelines focus on **sleep apnea** texts and the relevant medical concepts.
3. **Goal**: Produce a **gold standard** dataset where any important medical or domain-specific concept is labeled as the entity type, followed by "B", or "I", and all other text is not labelled.

2. What Counts as an Entity?

The following are defined as entities:

1. **Symptom**

A characteristic or manifestation of a condition. In this annotation, an entity considered a symptom must be a term a person without medical knowledge can say that they have observed. E.g. snoring, daytime fatigue. Examples of entities NOT considered a symptom: apnea (since this is the condition itself, the symptoms are what points to the person having apnea), closed upper airway (since this is not something observable to a regular person).

2. **Condition**

A medical condition or diagnosis, and abbreviations of these. E.g., "Obstructive Sleep Apnea,", "OSA", "Central Sleep Apnea", "Insomnia", "Hypoxemia".

3. **Risk Factor**

Factors that predispose to or precipitate a condition. E.g. obesity, sex, smoking, strokes, heart disease.

4. **Test**

Medical or diagnostic tests. E.g. polysomnography.

5. **Treatment**

Interventions or therapies for a condition. E.g. CPAP, sleep study.

6. **Outcome**

Potential results of a condition or treatment. E.g. improved oxygen saturation, reduced fatigue.

7. **Concept**

General medical or scientific concepts. E.g., positive airway pressure.

8. **Document**

Guidelines, classifications, or references. E.g., ICD-10, DSM-5.

3. What Is Not an Entity?

1. **Common nouns**

- E.g., "patient," "doctor," "study," "breathing," unless part of an official name.

2. **Adjectives / Descriptors** that are *not* part of the formal name

- E.g., “mild,” “severe,” “new,” “chronic.”

3. **Function words**

- Articles (e.g., “the,” “an”), prepositions (e.g., “in,” “of,” “for”), conjunctions (e.g., “and,” “or”), unless explicitly part of the entity’s official name (“The Ohio State University” scenario).

4. **Verbs, pronouns, adverbs**

- E.g., “is,” “he,” “often,” “quickly,” etc.

5. **Numbers / Measurements**

6. **Any text that is not directly recognized as an entity from the categories above.**

4. Labeling Rules

1. **Assign "B", then "I" to All Tokens of an Entity Mention**

- When you identify a span of text as an entity (e.g., "Obstructive Sleep Apnea"), the first token in the span is labeled **B**, and every following token in that span is labeled **I**.
- Example: "Obstructive" → I, "Sleep" → I, "Apnea" → I.

2. **Everything Else Is Not Labelled**

- If a token does not belong to an entity (it is "O"), do not label it. Leave it blank.

3. **Articles, Adjectives, or Prepositions in the *Official* Name**

- If it’s **truly** part of the name (e.g., “The National Sleep Foundation”), label it as **B** for the first token, and **I** for the rest. Otherwise, do not label it, but leave the field blank.
- Example: "The National Sleep Foundation" might be “The” → B, “National” → I, “Sleep” → I, “Foundation” → I, depending on how official you deem “The” to be.

4. **Acronym Mentions**

- Label acronyms (e.g., “OSA,” “CPAP”) as **B** if they stand for an entity in your defined scope.
- Example: “Obstructive Sleep Apnea (OSA)” → “Obstructive” → B, “Sleep” → I, “Apnea” → I, “(” → _, “OSA” → B, “)” → _.

5. **Entity Boundaries**

- Keep your annotation spans as **tight** as possible. Label only the words that directly constitute the entity name or recognized concept.
- Descriptive words or qualifiers outside the official name remain blank.

5. Consistency Guidelines

1. **Use the Same Rules Across All Documents**

- E.g., if “Obstructive Sleep Apnea” is labeled as “Obstructive” → B, “Sleep” → I, “Apnea” in one sentence, it should be labeled the same way in every occurrence if the term is used in the same context.

2. **Review Common Mistakes**

- Accidentally labeling a partial span (e.g., “Obstructive” but not “Sleep Apnea”).
- Forgetting to label abbreviations.
- Inconsistency with articles or brand names (e.g., sometimes labeling “The” as “I,” sometimes not).

If guidelines are unclear, pick one approach and **document** it so you can remain consistent.

6. Example Annotations

Example 1

Text:

“Patients with obstructive sleep apnea (OSA) often receive CPAP therapy.”

Annotation:

```
Patients →
with →
obstructive → condition B
sleep → condition I
apnea → condition I
( →
OSA → condition B
) →
often →
receive →
CPAP → treatment B
therapy → treatment I
. →
```

Here, “obstructive sleep apnea,” “OSA,” and “CPAP therapy” are considered entities and get labeled "B" and “I.” Everything else is blank, denoting "O"

7. Practical Steps for Annotators

1. **Identify Potential Entities**: Read each sentence to detect relevant concepts (diseases, devices, procedures, organizations, etc.).
2. **Check if It’s Official**: Is that word/phrase recognized as a name (e.g., brand, organization, medical condition)? If yes, label **all tokens** in that phrase as “I.”
3. **Everything Else**: Assign “O.”
4. **Keep Notes**: If you see repeated or ambiguous mentions, make sure you handle them consistently.

8. Final Check

- **Are all known entities in the text labeled?**
- **Did you avoid partial labeling?** (i.e., label the entire entity, not just part of it)
- **Are common words and general descriptors left without a label / blank?**

Following these steps ensures a **clear and consistent** BIO annotation process for your text about **sleep apnea**. Once done, you’ll have a gold standard that can train or evaluate simple Named Entity Recognition models with only two labels.