

Medical/Clinical Text Annotation Guidelines (for Real-MedNLP)

Shuntaro Yada[†] Eiji Aramaki[†] Ribeka Tanaka[‡] Fei Cheng[‡] Sadao Kurohashi[‡]
[†]Nara Institute of Science and Technology
[‡]Kyoto University

Version 8
December 20, 2021

Contents

1	Medical Entity Annotation	1
1.1	Outline	1
1.2	Basic policy	1
1.2.1	No nested construction	1
1.2.2	Giving a single tag to consecutive expressions	1
1.3	Medical Entity Types	2
1.3.1	Diseases and Symptoms	2
1.3.1.1	Entity scope	2
1.3.1.2	Expressions that determine the certainty attribute)	2
1.3.1.3	Important expressions	4
1.3.2	Anatomical parts	4
1.3.2.1	Important expressions	5
1.3.3	Features and Measurements	5
1.3.3.1	Important expressions	6
1.3.4	Change	6
1.3.5	Time Expression (TIMEX3)	7
1.3.6	Test	8
1.3.7	Medicine	9
1.3.8	Remedy	10
1.3.8.1	Important expressions	11
1.3.9	Clinical Context	11
1.3.10	Pending	12
1.4	What to Be Careful during Medical Entity Annotation	12
1.4.1	Length Precedence	12
1.4.2	Priority of Diseases and Symptoms and Time Expressions	13
1.4.3	Correction Attribute	13
1.4.4	Expressions Surrounding “Post-operative”	13
1.4.4.1	“Post-operative change”	13
1.4.4.2	“post-operative xxx”	13
1.4.5	SOAP Section and Medical Entity Attribute	13
1.5	Document-level Annotation Samples	14

Chapter 1

Medical Entity Annotation

1.1 Outline

In this chapter, annotation of medical entities in clinical texts will be explained. Medical entity annotation is the process of identifying **medical entities** on clinical texts, such as electronic health records or radiology reports, by giving them tags according to the medical concepts they represent such as disease, illness, and body region.

First, the basic policy for annotation will be outlined and medical entity types listed. Then, procedure tips will be given, finally followed by document-wise examples of medical entity annotation.

1.2 Basic policy

In this section, the basic annotation policy will be formulated in a way that is generally applicable, regardless of the type of the medical entities outlined in later sections. Please refer to section 2.4 for tips that do take the types of the medical entities into account.

1.2.1 No nested construction

This annotation specification does not allow the insertion of medical entities within other medical entities. For example, the disease name “chronic sinus infection” can be divided into “chronic,” “sinus,” and “infection,” the “sinus” component referring to a location on the human body. However, the guidelines do not permit annotation such as `<d certainty="positive">chronic <a>sinus infection</d>`. It is permitted to tag either the components as one medical entity type or a sequence of various medical entities

1.2.2 Giving a single tag to consecutive expressions

Unify consecutive medical expressions of an identical entity type into one entity if they are combined with the following phrases or symbols: , “also,” “and,” “as well as,” slash (/), hyphen (-), and comma (,).

Examples

- (1) `<TIMEX3 type="DATE">Previously</TIMEX3> identified <d certainty="positive">frosted glass-like and net-like, traction bronchiectasis, honeycomb lung</d>` were recognized.

1.3 Medical Entity Types

1.3.1 Diseases and Symptoms

Diseases and Symptoms entities `<d>` are expressions given to specific diseases and symptoms. Tags are also given to findings based on observations such as appearance (e.g., "frosted glass-like," "net-like") or sound (e.g., "fine crackles," "coarse crackles").

The degree of certainty to which the disease or symptom has been recognized (**certainty**) is expressed through one of the following values:

positive: The disease or symptom has been recognized

suspicious: The patient is suspected of having the disease or symptom in question (for instance, when it is suggested by a differential diagnosis¹)

negative: The existence of the diseases or symptoms has been ruled out with regards to the patient

general: An expression used for (1) accounts of generally occurring diseases or symptoms with no relation to the patient, or for (2) diseases and symptoms that might occur in the future

XML tag

```
<d certainty="{“positive” or “suspicious” or “negative” or “general”}>...</d>
```

1.3.1.1 Entity scope

The aim is to express the entities in units of compound nouns without prepositions such as “of”. Disease or symptom names expressed by compound nouns may include feature or change expressions (see Sections 1.3.3 and 1.3.4) and body regions (see Section 1.3.2). Furthermore, in the case of phrases such as “changes of/in xxx,” prepositions “of/in” may be included, so that “changes of/in xxx” may be considered a single diseases and symptoms entity (see Section 1.3.4). If a different tag is valid, an explanatory tag will be referred to at the time of the selection of an appropriate tag.

Examples

(2) However, for the concentration in the `<a>right lower lobe of the lung` a `<c>slight decrease</c>` can be seen in `<d certainty="positive">the percentage rise</d>`

(3) The `<c>tendency to develop exacerbation</c>` in `<d certainty="positive">interstitial pneumonia</d>`

(Although “interstitial” appears to be a Feature entity [the `<f>` tag; see Section 1.3.3], “interstitial pneumonia” should be regarded here as a single disease name (entity) and thus not further segmented [see Section 1.4.1].)

(4) `<d certainty="positive">right lung S2 nodule</d>`

(“Right lung S2” is a body region. However, because it forms a compound noun with “nodule”, they are given the `<d>` tag.)

1.3.1.2 Expressions that determine the certainty attribute)

Below, examples of key expressions that determine the value of the certainty attribute are in **bold** for each case of the attribute values.

positive Expressions that indicate existence such as “recognize,” “observe,” “scatter,” and “+”

¹Diseases with similar symptoms that need to be distinguished (differentiated). With regards to the symptoms the patient displays, several diseases are considered, and it is necessary to identify the correct one by an accurate examination. The diseases considered at such a time are referred to as a differential disease.

Examples

- (5) `<f>Uniform</f>` `<d certainty="positive">Ground-glass opacity</d>` `<f>scattered</f>` in the `<a>right lung`.
- (6) `<d certainty="positive">infectious complication</d>` **observed.**
- (7) `<d certainty="positive">Post-fracture deformation</d>` in the `<a>left first rib` was **recognized.**

suspicious “suspected,” “cannot be ruled out,” “possible,” “differentiated,” “even if,” “given that,” “appear,” and “possibly”

Examples

- (8) A `<c>tendency for exacerbation</c>` of `<d certainty="suspicious">acute pneumonia</d>` is **suspected.**
- (9) `<d certainty="suspicious">NSIP infection complications</d>` **cannot be ruled out.**
- (10) `<d certainty="positive">Calcification</d>` was recognized in the `<a>upper lobes of the bilateral lungs` that **appear to be** `<d certainty="suspicious">postinflammatory nodules</d>`.
- (11) **Even if** they are `<d certainty="suspicious">sarcoid nodules</d>`, they are `<f>atypical</f>`.

Important words and phrases “D/D” This is the abbreviation for “Differential Diagnosis,” which denotes that a disease or symptom should be distinguished (differentiated) from another diseases or symptoms following after this abbreviation. Therefore, the diseases or symptoms listed after a “D/D” sign should all have the suspicious value for the certainty attribute.

Examples

- (12) `<d certainty="suspicious">IPF/UIP</d>` is suspected
(D/D `<d certainty="suspicious">Chronic hypersensitivity pneumonitis</d>`)

negative “not recognized,” “without xxx,” “yyy vanished,” “(-),” and “after operation of zzz”

Examples

- (13) `<d certainty="negative">Small lymphadenopathy</d>` **cannot be recognized.**
- (14) The `<d certainty="negative">pneumothorax bullae</d>` in the `<a>apex of the left lung` have `<c>vanished</c>`.
- (15) **After** `<r state="executed">operation</r>` **of** `<d certainty="negative">pharyngeal cancer</d>`.

The negative certainty should also be given according to the surrounding expressions that describe changes in which symptoms disappear, because “after operation of zzz” is often used to describe “after the extraction of zzz.”

general “description of xxx,” “attention (to) yyy,” “prevention of zzz,” and “adverse reactions” (given when the disease and symptoms are described as a possible type of adverse reactions)

Examples

- (16) `<d certainty="general">IP</d>` as explained in the brochure.
- (17) **Guidance was given** to pay attention to `<d certainty="general">infection</d>`.
- (18) There are several types of `<d certainty="general">interstitial pneumonia</d>`.
`<d certainty="general">Idiopathic interstitial pneumonia</d>` includes
`<d certainty="general">hypersensitivity pneumonia</d>`, and
`<d certainty="general">secondary interstitial pneumonia</d>` caused by
`<d certainty="general">collagenosis</d>`.

1.3.1.3 Important expressions

TNM Classification When dealing with cancer, TNM Classification is an internationally recognized standard for classifying the extent of cancer progression. In clinical texts, each letter of TNM are preceded or followed by a number or lowercase alphabet and usually placed at the end of the sentence.

Disease and symptoms tags are given to TNM symbols. However, the certainty attribute is not set.

Examples

- (19) `<d certainty="positive">Primary lung cancer</d>` is considered. `<d> cT1c</d>`
 No `<f>pathologically significant</f>` `<d certainty="negative">lymphadenopathy</d>`
 was recognized in the `<a>mediastinum and pulmonary hilum`. `<d>N0</d>`

When the disorder constitutes something that does not exist or is lacking For example, the symptom “a lack of appetite” is mentioned in expressions such as “loss of appetite,” “lack of appetite,” “no appetite,” or “reduced appetite.” In such a case, where the disorder consists of the absence or lack of something essential, the entire disease and symptoms entity includes a negation part of the expression and is given a `certainty="positive"` attribute.

Examples

- (20) In `<TIMEX3 type="DATE">December that year</TIMEX3>`, treatment with
`<m-key state="negated">ofev</m-key>` was discontinued due to
`<d certainty="positive">diarrhea and a loss of appetite</d>`.

“Patterns of UIP” One of the expressions frequently used for interstitial pneumonia is “Patterns of UIP”, in which ‘UIP’ stands for Usual Interstitial Pneumonia. A `<d>` tag should be given to this entire phrase including the proposition “of.”

1.3.2 Anatomical parts

The `<a> . . . ` tag is given to expressions that describe an anatomical part, such as an organ or a body region. It include abstract expressions such as “marginal” and “internal.”

XML tag

```
<a>...</a>
```

Examples

- (21) `<d certainty="positive">Small lymph nodes</d>` are `<f>scattered</f>` in the `<a>right lung`.
- (22) An `<f>uneven</f>` `<d certainty="positive">rise in concentration</d>` has been recognized `<f>predominantly</f>` in the `<a>bilateral lower lobes of the lungs`.
- (23) `<d certainty="positive">post-fracture deformation</d>` was recognized in the `<a>left third rib`.
- (24) `<d certainty="positive">Nodules</d>` located just `<a>below the pleurae covering the bilateral lower lobes of the lungs`.

Since “along the margin” and “margin” itself indicate different regions, “along the margin” is combined as a single expression and labeled an Anatomical Part tag. Furthermore, if the “margin” is not represented by itself, but the object of the margin is also indicated, as in “xxx margin,” the object and margin are combined into one and given an Anatomical Part tag.

Examples

- (25) A `<d certainty="positive">solid part</d>` was recognized `<a>along the internal cavity margin`.

In principle, tags for anatomical parts are assumed to be given to the above (compound) nouns as a single unit, but, in the case of “xxx post-operative,” for example, where ‘xxx’ includes an anatomical part, the part is also tagged. For more details, see Section 1.4.4.

1.3.2.1 Important expressions

Abbreviations When you encounter a specific abbreviation that you are unsure of, you may find a clue by searching for it on the web by using the keywords. However, if the meaning is not immediately clear, you may give the word a pending tag (see Section 1.3.10).

Examples

- (26) No `<d certainty="negative">findings of invasion</d>` in `<a>SVC`.
(SVC stands for superior vena cava, or superior caval vein.)

1.3.3 Features and Measurements

Features and Measurements tags are given to modifying phrases or predicative adjectives (such as “scattered”) that pertain to the features, measurements, values, areas, or degrees of a given Disease or Symptom entity. They are also given to expressions that indicate degrees (such as “mild”). However, if the degree expression is connected with an expression that indicates changes, the degree expression is given a Change tag (`<c>`), (mentioned below in Section 1.3.4]), instead of features and measurements tags.

XML tag

```
<f>...</f>
```

Examples

- (27) `<f>Well-defined, smooth-margin</f>` `<d certainty="positive">nodular shadow</d>` was recognized.

- (28) `<d certainty="positive">Lymph nodes</d>` were found `<f>scattered</f>` in the `<a>mediastinum`.
- (29) `<f>Uneven</f>` `<d certainty="positive">rise in concentration</d>`
- (30) `<f>Localized</f>` `<d certainty="positive">even shadow</d>`
- (31) `<f>Diffuse</f>` `<d certainty="positive">bronchiectatic opacity</d>` occurred in `<a>both lungs`.
- (32) `<d certainty="positive">Adenopathy</d>` of `<f>less than 3 cm in diameter</f>` found `<f>scattered</f>`.
- (33) `<f>1.6 × 1.6 cm in the mediastinal window</f>`
- (34) `<d certainty="negative">Adenopathy</d>` of a `<f>significant size</f>` has not been recognized.
- (35) `<d certainty="negative">LN</d>` of a `<f>Positive size</f>` was not recognized.
- (36) Presence of a `<f>small amount</f>` of `<d certainty="positive">bilateral pleural effusion</d>` in the `<a>lungs`.
- (37) `<f>Mild</f>` `<d certainty="positive">opacity</d>` was recognized.

If a given disease is present but does **not** have any of the features, the expression indicating negation should also be included in the tag. Note that, unlike disease and symptoms entities, features and measurements tags does not have any attributes.²

Examples

- (38) Presence of `<d certainty="positive">lymph nodes</d>` in the `<a>left paratracheal mediastinum`. `<f>It is not large enough to be considered pathologically significant</f>`.

1.3.3.1 Important expressions

Prevalent xxx can frequently be found in expressions such as “prevalent on the right side” or “prevalent in the bilateral superior lobes of the lungs,” especially in radiographic findings. However, “prevalent” is always treated as a feature and measurement entity.

Examples

- (39) `<d certainty="positive">Reticulation</d>` is `<f>prevalent</f>` in `<a>the bilateral superior lobes of the lungs`.

1.3.4 Change

This tag is given to expressions that signify changes in the condition such as aggravation and decrements of symptoms, or changes related to test values and treatment status. This tag is also given to situations in which there is no change, such as “no remarkable change.” Modifiers that express the degree of change and expressions that signify change are bundled together and given one tag such as “slight augmentation.”

²Unlike diseases and symptoms, which are mainly expressed by nouns, features and measurements tend to be expressed with predicative adjectives and verbs. If the certainty attribute is also introduced in features and measurements tags in the same way as in disease and symptoms tags, 1) the handling of double negative expressions such as “not non-specific” is likely to cause confusion in determining the scope of tagging and in determining their facticity, and 2) if features and measurements co-occur with disease and symptoms entities, such as “a non-prominent adenopathy was recognized,” the interpretation of the certainty value becomes even more complicated. In such cases, the features and measurements tags are extended wordily to include all the negative expressions. Also, a change tag with the same characteristics (see below) will include the negative expressions for the same reason.

However, even when using the expression “change,” such as “post-fracture change,” the expression may be given a disease and symptom tag. Furthermore, if “change” is strongly connected with diseases and symptoms such as “contractile change” or “post-inflammatory change” to form a compound noun, the Diseases and Symptoms tag <d> is to be given to the entire compound noun (or noun phrase) instead of using a <c> tag. (Only in such cases is it acceptable to include noun phrases joined by the prepositions “of/in,” see Section 1.3.1). The expression “post-operative change” referred to in Section 1.4.4.1 is an example of this.

XML tag
<c>...</c>

Examples

- (40) <d certainty="positive">Contractile changes</d> are <c>becoming stronger</c>.
- (41) Overall <c>moderate augmentation</c> is visible.
- (42) The <d certainty="negative">pneumothorax bullae</d> of the <a>left lung apex <c>dissipated</c>.
- (43) Suggested <d certainty="suspicious">acute pneumonia</d>: <c>Slight aggravation</c> since <TIMEX3 type="DATE">last time</TIMEX3>.
- (44) <c>Slight augmentation</c> of the <d certainty="positive">rise in concentration</d> in the <a>right inferior lobe.

Note that negative statements indicating **no change** are enclosed entirely in <c> tags, including the negative part (same as feature tags, see Section 1.3.3)

Examples

- (45) <c>No remarkable changes</c> since <TIMEX3 type="DATE">last time</TIMEX3>.
- (46) <d certainty="positive">Small lymph nodes</d> were recongnized in <a>the lower part of the tracheal bifurcation and in the right and left lower paratracheal areas. However, they showed <c>little chronological change</c>.
- (47) Despite the suspected <d certainty="suspicious">benign nodules</d>, <c>no decrease is visible</c>

1.3.5 Time Expression (TIMEX3)

To the date, period, or the timing of an event such as “last time” in health records, the time expression tag <TIMEX3>³ is given. Depending on the nature of the time, one of the following values is selected for the type attribute:

DATE: Expressions which put focus on the calendar date

TIME: Expressions which focus on a moment within the day and non-specific expressions such as “now” and “currently”

DURATION: Time expressions focusing on a period instead of the start and end of the time axis

SET: Frequency set expressions focusing on multiple dates, times, and periods

AGE: Expressions related to age

MED: Time expressions specific to medicine, such as “post-operative”

MISC: Expressions that do not fall under any of the above

³With relation to time expression, these guidelines introduce TIMEX3 as defined in the previously used TimeML (version 1.2.1). However, as mentioned below, only selected values for the type attribute are adopted, with their use slightly extended for medical documents. Specifically, MED and MISC are the type attribute values extended by the guidelines.

As in the examples below, the <TIMEX3> tag is given to entire expressions such as "3 months later" and "5 months earlier." However, the tags <TIMEX3> are not given to prepositions in expressions such as "from August 3, 2023" or "since August 11, 2022."

XML tag

```
<TIMEX3 type="{DATE, TIME, DURATION, SET, AGE, MED, MISC}">...</TIMEX3>
```

Examples

- (48) The conditions of the <d certainty="positive">lymph nodes</d> in the <a>mediastinum and pulmonary hilum remain <c>almost the same</c> as <TIMEX3 type="DATE">last time</TIMEX3>.
- (49) <d certainty="positive">Difficulty opening</d> <a>right eyelid in the <TIMEX3 type="TIME">morning</TIMEX3>
- (50) <t-key>Smoking</t-key>: <t-val>20 cigarettes a day</t-val> for <TIMEX3 type="DURATION">40 years</TIMEX3> From <TIMEX3 type="AGE">59 years old</TIMEX3> <t-val>stop smoking</t-val> ("Smoking" as a medical history item is signified as a medical finding result after ":". However, a nested structure is not allowed [see Section 1.2.1], and when a time expression tag is given preference [see Section 1.4.2], the test result tag <t-val> is given to the words that remain after a time expression tag is attributed.)
- (51) <TIMEX3 type="SET">Frequent</TIMEX3> <d certainty="positive">fever</d>
- (52) <TIMEX3 type="MED">Post-operative</TIMEX3> <d certainty="positive">shortness of breath</d> after exercise at the level of climbing steps or a short run.

Furthermore, as in the example below, the time expression tag <TIMEX3> is not attributed only to the word "after" but to the entire phrase such as "post-operative" or "after resection."

Examples

- (53) <m-key state="executed">Nasal</m-key>: <m-val>1L/min</m-val> <t-key>SpO2</t-key>: <t-val>94-95%</t-val> was initiated, and combined with <TIMEX3 type="MED">post-sedation</TIMEX3> <m-val>6L/min</m-val> due to a <p>decline in oxidation capability</p>.

1.3.6 Test

The <t-test> tag is given to a set of tests or interviews, the <t-key> tag is given to a test or interview item (or medical indicator), and the <t-val> tag is given to the value of a test or interview result (or a medical indicator result). However, <t-val> is also used with non-numerical values such as "positive" or "negative."

To the <t-test> test name, the state attribute can be given to identify an applicable state, which has the following values:

- scheduled:** events scheduled for the future (does not apply to the time of scheduling)
- executed:** events that have already been carried out
- negated:** events that have not been carried out
- other:** situations other than those stated above

Note that the state attribute is, in principle, not given to the <t-key> and <t-val> tags. However, when applicable to a specific examination item that is listed, the state attribute may be added to the <t-key> (see Section (58)).

XML tag

`<t-test state="{scheduled, executed, negated, other}">...</t-test>` — test or interview name

`<t-key (state="{scheduled, executed, negated, other}")>...</t-key>` — test or interview item, or medical indicator

`<t-val>...</t-val>` — value of test or interview result, or medical indicator result

Examples

(54) `<t-test state="executed">Chest CT</t-test>`

(55) `<t-key>FEV1</t-key>: <t-val>1.97L(80.0%)</t-val>`

When giving tags to the values in the test results, do not give tags to the items such as “before,” “L,” or “H,” which are listed with a space to the left of the value.

Examples

(56) `<t-key>KL-6</t-key>_... <t-val>559</t-val>_H`

(Spaces are represented as “_” in this example.)

(57) `<t-key>cons</t-key>: <t-val>alert</t-val>`

Test items can essentially be expected to be followed by test values. However, for reasons such as the test not yet having been carried out, there are situations where the test result is not given. When the sentence makes it possible to conclude that a test will be performed at a later time, a `state="scheduled"` is given.

Examples

(58) `<t-key state="scheduled">6-minute walk</t-key>`

(In the P column of SOAP, various test items are stated. Therefore, it is possible to conclude that a test is planned for next time [see Section 1.4.5]. Also, because, in this item, it is neither 5 nor 10 minutes but a 6-minute walk, it is a fixed test. Therefore, the whole phrase is given a test tag without the inclusion of a time expression tag.)

Also, the diseases and symptoms tag or time expression tag takes precedence over the test tag for the phrases to which the test item or test result is also applicable (see Section 1.4.2).

Examples

(59) `<d certainty="positive">phlegm</d> (+)`

(This kind of expressions could be listed in a report of test/interview results.)

(60) `<t-key>Started smoking</t-key> at: <TIMEX3 type="AGE">20 yo</TIMEX3>`

`<t-key>Stopped at</t-key>: <TIMEX3 type="AGE">75 yo</TIMEX3>`

1.3.7 Medicine

The name of medication is given the `<m-key>` tag, while the value of the prescribed amount is given the `<m-val>` tag. In the same way as with test tags, the `<m-key>` medication name is also given a state attribute depending on the applicable administration state.

scheduled: when the medication is scheduled to be administered in the future (and has not yet been administered at the time of scheduling)

executed: when the medication is being administered

negated: when administration stopped due to steps such as discontinuation

other: situations other than those stated above

Furthermore, there are situations in which the `<m-val>` may also be given a state attribute (see Section (63)).

XML tag

```
<m-key state="{scheduled, executed, negated, other}">...</m-key> – medication name
<m-val (state="{scheduled, executed, negated, other}")>...</m-val> – medication value
```

Examples

(61) Administration of `<m-val>100mL/1hr</m-val>` of `<m-key state="scheduled">pulmocare</m-key>` scheduled to start `<TIMEX3 type="DATE">tomorrow</TIMEX3>`.

If there are multiple expressions related to medication, the condition mentioned last takes precedence, and a tag is given after checking the sentence ending.

Examples

(62) In `<TIMEX3 type="DATE">2027/7</TIMEX3>`, `<m-key state="negated">perispa</m-key>` `<m-val>1,200mg</m-val>` induced `<d certainty="positive">anorexia</d>` and was **abandoned**.

In principle, the state attribute is given only to the medication name (`<m-key>`) tags. If there is a description to change the value of a medication that is being (or scheduled to be) administered or to discontinue it, the state attribute can be given to the medication value tag (`<m-val>`).

Examples

(63) `<c>Reduction</c>` of `<m-key>predonine</m-key>` from `<m-val state="executed">10mg</m-val>` to `<m-val state="scheduled">7.5mg</m-val>`

Also, in the same way as with test entities (`<t-val>`), if there is no disease or symptom and if a time expression tag is applicable as a medication entity (`<m-key>`) value, it takes precedence over the medication value tag (`<m-val>`).

Examples

(64) `<m-key state="executed">Enteral nutrition</m-key>` was scheduled from `<TIMEX3 type="DATE">today</TIMEX3>` `<TIMEX3 type="SET">three times a day</TIMEX3>`. However, that has been changed to the same pace as until `<TIMEX3 type="DATE">yesterday</TIMEX3>`, `<TIMEX3 type="SET">once every morning</TIMEX3>`

1.3.8 Remedy

Therapy, surgery, treatment and other treatment methods are given the `<r>` tag. In the same way as with test entities (`<t-test>`) and medication name entities (`<m-key>`), a state attribute is given according to the applicable state of the treatment.

scheduled: treatment that is scheduled for a point in the future (not carried out at time of record)

executed: treatment that is scheduled for a point in the future (not carried out at time of record)

negated: treatment is no longer carried out due to reasons such as discontinuation

other: situations other than those stated above

XML tag

```
<r state="{scheduled, executed, negated, other}">...</r>
```

Examples

- (65) Not `<r state="negated">PCI</r>`, but `<t-test state="executed">CAG</t-test>` was carried out.
- (66) Due to `<r state="executed">defibrillation</r>`, the patient was `<cc state="executed">admitted</cc>` to the Cardiovascular Department on `<TIMEX3 type="DATE">March 6-7, 2026</TIMEX3>`.

Many of the expressions that are considered remedy entities (`<r>`) are similar to medication names. If the expression is not accompanied by an applicable value or frequency, the `<r>` is given to signify an implementation status. Conversely, as in the example below, if it is accompanied by an applicable value (even if expressed as a treatment), it is considered a medical entity (`<m-key>`, `<m-val>`).

Examples

- (67) Perhaps due to `<m-val>2.5mg</m-val>` `<m-key state="executed">steroid</m-key>` taken `<TIMEX3 type="cc">during chemotherapy</TIMEX3>` for `<d certainty="positive">NHL</d>` mentioned above, absolutely no `<m-key state="negated">cough medicine</m-key>` was taken orally, the patient did not `<d certainty="negative">cough</d>` and there was no `<d certainty="negative">DOE</d>`.

1.3.8.1 Important expressions

Medicine name + Remedy If a “remedy” expression follows a medicine name, the whole phrase is treated as a remedy entity. For example, “steroids” are usually given an `<m-key>` medical name tag. However, if it is mentioned as a part of “steroid treatment,” the whole expression is given an `<r>` remedy tag.

Pacemakers A pacemaker itself is a piece of medical equipment. However, in a clinical test, it indicates pacemaker treatment and is therefore given an `<r>` tag.

1.3.9 Clinical Context

A clinical context tag `<cc>` is given to expressions that signify situations related to the patient and a medical facility such as hospital admission, discharge, readmission, transfer, visit, initial diagnosis, and follow-up (f/u). Diagnosis and referral, i.e., actions taken by the doctor, are not given this tag. The implementation status is distinguished by the state attribute.

scheduled: treatment scheduled for the future (not carried out at time of record)

executed: treatment that has already taken place

negated: treatment that is no longer carried out due to reasons such as discontinuation

other: situations other than those stated above

XML tag

```
<cc state="{scheduled, executed, negated, other}">...</cc>
```

Examples

- (68) The patient was an `<cc state="executed">outpatient</cc>` at the Cardiovascular Department.
- (69) However, the patient `<cc state="executed">consulted</cc>` a local doctor at the `<TIMEX3 type="DATE">end of March</TIMEX3>` and came to a `<cc state="executed">referral visit</cc>` at our Immunology Department.
- (70) The patient is scheduled to be `<cc state="scheduled">discharged</cc>` at the `<TIMEX3 type="DATE">end of this week (August 27)</TIMEX3>`.

1.3.10 Pending

To a medical term that may correspond to a medical entity for which a decision cannot be made, give a pending tag proactively.

XML tag

```
<p>...</p>
```

Examples

- (71) A `<p>spare</p>` was done `<a>directly under the pleura`.
- (72) `<p>LNs</p>` with a `<f>breadth of up to 4 cm</f>` inside the `<a>mediastinum`.
- (73) `<d certainty="positive">Reticular shadow</d>` in the `<p>Coronary</p>`

1.4 What to Be Careful during Medical Entity Annotation**1.4.1 Length Precedence**

In principle, the length of one medical entity is compound words in the case of Diseases and Symptoms entities and anatomical parts. In the case of features and measurement, it is the modifier or structural stem of a disease and symptoms entity. For change entities, the unit is a verb.

Particularly in the case of compound words, rather than subdividing them into smaller segments to be tagged, one type of tag should be given to the whole span, taking into account its primary medical meaning. For example, "brain metastasis" can be divided into "`<a>brain <d certainty="positive">metastatis</d>`." However, based on this annotation guide, the tag is given to the entire span, `<d certainty="positive">brain metastasis</d>`.

Since cases vary depending on the context, the policy is for **the annotator to give a tag to the longer expression when in doubt**. For example, as seen in the example below, it is desirable to treat all of the phrase, "combination of infectious diseases caused by bacteria, viruses, and pneumocystes" as a single diseases and symptoms entity (see also Section 1.2.2). As an exception, there are certain medical entities that always need to be given a tag; for example, the word "prevalent" in phrases like "prevalent xxx" must be tagged as Features and Measurements entities [see sections such as 1.3.3.1).

Examples

- (74) There were no findings on `<d certainty="negative">bone metastasis</d>`.
- (75) There is a `<d certainty="positive">lymphatic node</d>` in the `<a>#4L mediastinum` with `<f>a short diameter of 1.9 cm</f>`.
(In this case, "lymphatic node" is not a location but a short version of "lymphadenopathy" and is therefore given the `<d>` tag.)

(76) The `<c>aggravation</c>` of `<d certainty="suspicious">interstitial pneumonia</d>` is suspected to be a `<d certainty="suspicious">combination of infectious diseases caused by bacteria, viruses, and pneumocystes</d>`.

1.4.2 Priority of Diseases and Symptoms and Time Expressions

If an expression can be given different various tags, Diseases and Symptoms `<d>` and Time Expression tags `<TIMEX3>` are prioritized.

1.4.3 Correction Attribute

If a typographical error is found in an expression that has been given a tag, a correction attribute is given to the tag, and the expression that is thought correct is added as the value. It is not necessary to give this attribute to the medical entities that do not have any tag. Also, there is no need to correct such expressions that require advanced medical knowledge to correct properly.

Examples

(77) `<d certainty="positive" correction="ground glass opacity">ground grass opacity</d>`

1.4.4 Expressions Surrounding “Post-operative”

The expression “post-operative” appears frequently. However, the tag to be given depends on the context.

1.4.4.1 “Post-operative change”

Prioritized examples shown henceforth, the `<d>` tag is to be given to the phrase “post-operative changes.” The expressions preceding “post-operative changes,” such as “**Cancer stomach** post-operative changes,” “**left nephrectomy** post-operative changes,” should be annotated separately while the part “post-operative changes” is labeled `<d>`.

Examples

(78) `<d certainty="positive">Post-operative changes</d>` were found in the `<a>rectum`.

1.4.4.2 “post-operative xxx”

In cases in which “post-operative” stands at the beginning of a compound noun, it becomes a time expression entity and an appropriate tag is given to its succeeding part. However, as mentioned above, be careful to make “post-operative **change**” a Diseases and Symptoms entity (see Section 1.4.4.1).

Examples

(79) `<TIMEX3 type="cc">Post-operative</TIMEX3>` `<d certainty="positive">left-lung pleural effusion</d>`

1.4.5 SOAP Section and Medical Entity Attribute

If the target text is a medical record, it is frequently put together using a format called SOAP. SOAP denotes:

- S (Subject):** subjective data, such as the patient’s account;
O (Object): objective data, such as physical examination and test data;
A (Assessment): data evaluation; and
P (Plan): treatment plan based on data and its evaluation.

This SOAP format enables organizing the information about a medical treatment into these four sections. An **I (Informed consent; direct explanation to the patient)** may also be included into this SOAP template.

For example, since the P section usually contains a future treatment plan, the scheduled value is basically assigned to the state attribute. In the case of Disease or Symptom entities, since they can be mentioned generally within a sentence that describes a certain medical plan in the P section, certainty="general" attributes should be assigned. In the case where a time expression tag is given, the document creation time as metadata should be carefully considered.

Examples of the annotation according to the SOAP sections are listed below.

Examples

- (80) `<d certainty="positive">IP (IPF, NSIP, others)</d>`
 (If this phrase appears in the A section, it means, “synthesizing various types of data show that IP = interstitial pneumonia [however, IFP = idiopathic pulmonary fibrosis and NSIP = non-specific interstitial pneumonia, or ‘other’ options are possible].” It is, therefore, given a certainty="positive" attribute.)
- (81) `<TIMEX3 type="DATE">September 27</TIMEX3> <t-test state ="scheduled">CT</t-test>`
 (If this phrase appears in a P section, it means, “CT is planned on September 27, a date after the medical record was made.” Therefore, it is given a state="scheduled" attribute.)
- (82) `<d certainty="general">IP</d>` is explained in the brochure.
 (This expression is often see in the I section. As it means explaining IP (interstitial pneumonia) to the patient or their family, it is given a certainty="general" attribute.)

1.5 Document-level Annotation Samples

These are examples of whole documents with the annotations as specified in these guidelines.

Examples

- (83) At the `<a>base of the lung` `<a>straight under both sides of the pleura`, a `<f>prevalently</f>` `<a>right` `<d certainty="positive">reticular shadow and traction bronchiectasis</d>` was found, no `<d certainty="negative">honeycomb lung</d>` was found. Suspected `<p>Possible UIP pattern</p>`. `<f>Scattered</f>` `<d certainty="positive">patchy ground-glass opacity</d>` in the `<a>right lung`. No inconsistency in single-string `<d certainty="positive">interstitial pneumonia changes</d>`. No indication of `<f>significant</f>` `<d certainty="negative">adenopathy</d>`. No findings of `<d certainty="negative">pleural effusions</d>`. `<a>Aorta` `<d certainty="positive">calcification</d>` was found. `<d certainty="positive">Small LDAs</d>` were found in `<a>both thyroid glands`. `<d certainty="positive">Cysts in both kidneys</d>`. Suspected `<d certainty="suspicious">interstitial pneumonia</d>` diagnosis.
- (84) A `<a>right` `<f>prevalent</f>` `<d certainty="positive">reticular shadow and traction bronchiectasis</d>` were found on the `<a>right side` of the `<a>base of the lung`. `<a>Right under the pleura`, a `<p>spare</p>` was made, and `<d certainty="suspicious">NSIP pattern</d>` is suspected. `<f>Scattered</f>`

<d certainty="positive">patchy ground-glass opacity</d> in the <a>right lung. No inconsistency in single-string <d certainty="positive">interstitial pneumonia changes</d>. <f>Slightly</f> <d certainty="positive">strengthened density</d> compared to <TIMEX3 type="DATE">last time</TIMEX3> was found. <f>Small</f> <d certainty="positive">lymph nodes</d> were <f>scattered</f> <a>in the mediastinum. However, there are no indications of <f>significantly large</f> <d certainty="negative">lymph nodes</d>. No findings of <d certainty="negative">pleural effusions</d>. <a>Aorta <d certainty="positive">calcification</d> was found. <f>Small</f> <d certainty="positive">LDAs</d> were found in <a>both thyroid glands. <c>No remarkable changes</c> since <TIMEX3 type="DATE">last time</TIMEX3> Suspected <d certainty="suspicious">interstitial pneumonia</d> diagnosis. <c>Slight aggravation</c> since <TIMEX3 type="DATE">last time</TIMEX3>.

(85) S) <t-key>MRC</t-key> <t-val>grade 3</t-val>

O) <t-key>SpO2</t-key> <t-val>90% (RA)</t-val> <t-test state="executed">Home pulse oxymeter</t-test> (<TIMEX3 type="CC">before bed</TIMEX3>) <t-key> Average SpO2</t-key> <t-val>89%</t-val> <t-key>SpO2 frequency distribution (& <90%></t-key> <t-val>62.11%</t-val>

A) # <d certainty="positive">Interstitial pneumonia (UIP-like)</d> <TIMEX3 type="DATE">Around two years earlier</TIMEX3> <d certainty="positive">exertional dyspnea</d> <c>appeared</c>, <t-key>Hugh-Jones</t-key> <t-val>II</t-val>. <f>Gradual</f> <c>aggravation</c> <TIMEX3 type="DATE">since then</TIMEX3>, <TIMEX3 type="TIME">Currently</TIMEX3>, <t-key>MRC</t-key> <t-val>grade 3</t-val> <t-test state="executed">stable breathing function</t-test> <TIMEX3 type="DATE">May 30, 2013</TIMEX3> <t-key>VC</t-key> <t-val>2.27 (120.3%)</t-val> <t-key>FEV1.0</t-key> <t-val>1.58 (120.2%)</t-val> <t-key>FRC</t-key> <t-val>2.35 (159.8%)</t-val> <t-key>DLCO</t-key> <t-val>3.53 (30.2%)</t-val> <TIMEX3 type="DATE">August 3, 2010</TIMEX3> <a>thorax <t-test state="executed">CT</t-test> <d certainty="positive">pneumonectasia</d> (<a>superior lobe predominant) + <d certainty="positive">UIP pattern</d> (<a>inferior lobe predominant) <TIMEX3 type="DATE">August 16, 2010</TIMEX3> - <r state="executed">pirespa</r> (stopped due to <d certainty="positive">eyelid dermatitis</d>) <r state="executed">pirespa</r> reintroduced on <TIMEX3 type="DATE">December 27, 2010</TIMEX3>, stopped on <TIMEX3 type="DATE">May 7, 2012</TIMEX3> (<d certainty="positive">cutaneous pruritus</d>) <TIMEX3 type="DATE">August 29, 2011</TIMEX3>: blood serum # <d certainty="positive">hyperlipidemia</d> (<t-test state="executed">medical history</t-test>) <d certainty="positive">Acute hepatitis</d> B (<TIMEX3 type="AGE">32 years old</TIMEX3>) <t-key>Smoker</t-key>: <t-val>20 cigarettes</t-val>, <t-val>quit</t-val> <TIMEX3 type="DATE">5 years ago</TIMEX3>

P) <r state="scheduled">HOT</r> induced application for <cc state="scheduled">hospitalization</cc>

Acknowledgment

These guidelines are supported by the Public/Private R&D Investment Strategic Expansion Program (PRISM), Japan.

We thank Joh Ayami and Kaoru Ito for the cooperation in the draft production of the entity-annotation scheme. Valuable advice is provided by the following medical specialists: Mari Ito and Masataka Kuroda at National Institutes of Biomedical Innovation, Health and Nutrition Japan; Kazuma Kobayashi at National Cancer Center Japan; and Yuta Nakamura at The University of Tokyo Hospital. We highly appreciate their support.