SURUS Dataset Annotation Manual

General remarks

1. Rules for annotation of 26 distinct labels are described in order of label class. For each label, some correct and incorrect examples are given to clarify the context in which these should be annotated.
2. For some labels, the section is a relevant factor. At the start of the description for every annotation label, the sections where the labels should and should not be indicated are highlighted with green or red. Yellow is used when the label should only be assigned in exception cases.
3. Rules described are only relevant to **bold entities** in the examples. Other elements in the same sentence may be annotated within the same label, but these are annotated through a different rule and are therefore not bold.
4. **Blue bold** elements are examples of correct annotations, **red bold** elements are examples of incorrect annotations.
5. The current manual applies to annotation of interventional study abstracts, and with minimal adjustment can be used for annotation of observational study abstracts.

DRUG

DRUG_MOLECULE

Introduction	
Methods	
Results	
Conclusion	
No section designation	

When?

1. All references to **drugs**, including both the scientific and the commercial terms, are annotated as DRUG_MOLECULE. This is done very consistently and the DRUG_MOLECULE label has priority over most other labels.
    - "**metformin**"
    - "**fluticason** (**F**)"
    - "**tenofovir alafenamide**"
    - "The objective was to characterize safety and efficacy of **ADS-5102**"
2. References to **biologicals**.
    - "**peginterferon alfa-2 b**"
    - "**Lysozyme** (270 mg) or placebo was administered orally for 52 weeks"

When not?

1. If the drug is part of the **inclusion criteria**: patients may be selected based on the drugs they were taking before entry into the study. These instances, the drug may be part of the INCLUSION_CRITERIA class.
    - "patients inadequately controlled on **metformin** monotherapy"
    - "patients with antiphospholipid syndrome who were taking **warfarin**"
2. When the drug can be involved in a PARAMETER_EFFECT class annotation, for example when its measured as a pharmacokinetic parameter during the study.
    - "higher **adalimumab** DL predicted successful dose reduction"
    - "Antibodies to **infliximab** were detected in 26.6% of the patients."

DRUG_CLASS

Introduction	
Methods	
Results	
Conclusion	
No section designation	

When?

1. All terms referring to a **type of drug**.
    - "Combination therapy with a **long-acting bronchodilator** and an inhaled **corticosteroid** (**ICS**) is recommended in chronic obstructive pulmonary disease."

- "The **active vaccine** was safe, with mostly grade 1/2 toxicities"
- "Data on **P2Y12 inhibitor** monotherapy after short-duration dual antiplatelet therapy (DAPT) in patients undergoing percutaneous coronary intervention are limited"
- "**biosimilar**"

2. Important elements often used as **control group**.
    - "**placebo**"
    - "**saline**"

When not?

1. Any additional clarification should not be incorporated into the annotation.
    - "**inhaled** corticosteroid"

DRUG_TREATMENT_GROUP

Introduction
Methods
Results
Conclusion
No section designation

When?

1. All references to a **group studied** in the abstract.
    - "**intervention group**"
    - "**cohort 1**"
2. All references to **therapies** for patients.
    - "**maintenance therapy**"
    - "**standard of care**"
3. A **ratio** for patient assignment between groups.
    - "**1:1**"
4. The **size of a group** in a study. Only sizes of groups defined in the study methodology are annotated, not for subgroups identified during the execution of the study.
    - "**n=112**"
    - "**162** participants received vaccine"
5. Reference to a **period** or **study** within a study abstract.
    - "**period 1**"
    - "**study A**"

DRUG_DEVICE

Introduction
Methods
Results
Conclusion
No section designation

When?

1. Any reference to a **specific medical device**.
    - "**inhaler**"
    - "**needle**"
2. Any **brandname** of a medical device.
    - "**Ellipta**"

When not?

1. Within inhalers, there's distinction between the inhaler and the inhaled substance, which should be annotated as DRUG_MOLECULE.
    - "**Asmanex**"
2. When the context indicates that the device is not used.
    - "**needle**-free injection"

DRUG_FORMULATION

Introduction	
Methods	
Results	
Conclusion	
No section designation	

When?

1. **Formulation** in which the drug can be administered.
    - "**tablet**"
    - "**injectable**"
2. Any **characteristic** of the formulation of a drug.
    - "**nebulized**"
    - "**extended release**"
3. An **abbreviation for a combination** of two drugs or a drug and other element which cannot be annotated separately.
    - "**FSC**" (for "fluticason salmeterol combination")
    - "Patients were randomized to receive **BIAsp** 30 or placebo twice daily"
4. A **vaccine for a certain medical indication** or a **specific vaccine**.
    - "1 week after receiving the last **rabies vaccine**, anti-rabies antibodies increased"
    - "To evaluate the immunogenicity and safety of a reduced antigen **diphtheria-tetanus-acellular pertussis-inactivated poliovirus (dTap-IPVB) vaccine**."

DRUG_FUNDER

Introduction	
Methods	
Results	
Conclusion	
No section designation	

When?

1. When the context suggests the funding of a research by a company or an institution. Is most commonly listed at the end of an abstract.
    - "Funding **Merck**"
    - "Funded by the **National Heart, Lung, and Blood Institute** and the **Canadian Institutes of Health Research**"

When not?

1. Any reference to a company or institution of which it is unclear that they funded the research.
    - "To evaluate safety and efficacy of adalimumab [ADA; **Humira AbbVie Inc**, North Chicago, IL, USA)]"

METHODOLOGY

METHODOLOGY_STUDY_DESIGN

Introduction	
Methods	
Results	
Conclusion	
No section designation	Annotation the sections that are contextually clear to be part of Methods

When?

1. A **direct reference** to the methods of the study set-up. Each element that fulfils the conditions should be annotated, meaning that they can be annotated multiple times in each article. Elements should be annotated separately if that is possible, except for elements that are connected.
    - "This was a **phase II**, **randomized**, **double-blind**, **placebo-controlled**, **multicenter** study"
    - "**2-period crossover**"
2. Description of the **goal** of the study

- "The primary efficacy outcome was to test the **non-inferiority** of LY IGlar to IGlar."
- "**superiority**"
- "**bioequivalence**"

3. **Location** where the study is conducted.
   - "at **eight study sites** in the **USA**"
   - "**intensive care units** of **7 medical centers**"
   - "We enrolled patients from **311 centres** in **16 countries** across **five continents**"
4. Complete description of a **period** within a study ("2-week open-label run-in period")
   - "**two 32-week treatment periods**, each with a **16-week titration** and a **16-week maintenance period**"

When not?

1. Elements that do not have any additional value.
   - "**study**"
2. If an element specifically relates to administration of a drug, rather than the entire study.
   - "patients received CZP 400 mg, **open-label**, every 4 weeks"

METHODOLOGY_INCLUSION_CRITERIA

Introduction	Only if the element is not mentioned in the Methods section.
Methods	
Results	Only if the element is not mentioned in the Methods section and/or Introduction.
Conclusion	
No section designation	Annotation the sections that are contextually clear to be part of Methods or, in accordance with the exceptions, Introduction and Results

How?

As concise as possible ("healthy infants" is divided into "healthy" and "infants". An associated duration and a drug dose that is required for participants are annotated separately as METHODOLOGY_DETERMINATION.

When?

1. A **condition** that study participants have to meet ("patients", "type 2 diabetes"). This could include:
   - A disease
   - A drug, another therapy or hospitalization, for instance.
     - "patients were on **basal insulin, ± oral antidiabetic drugs**"
   - A history of the aforementioned
   - Another characteristic, such as age, gender and body weight
2. When a participant specifically should **not** meet a condition. This annotation should include the negation.
   - "not have diabetes 2"

When not?

1. Elements that do not have any additional value
   - "**healthy**" instead of "**healthy subjects**"

METHODOLOGY_DETERMINATION

Introduction	
Methods	
Results	
Conclusion	
No section designation	Annotation the sections that are contextually clear to be part of Introduction, Methods, Results

How?

Always as an addition to a METHODOLOGY_INCLUSION_CRITERIA, not together with other labels.

When?

1. The **dose** or **number of doses** of a drug, used by participants at the start of the study, which are annotated as METHODOLOGY_INCLUSION_CRITERIA
   - "**stable dose** of sulfonylurea"

- "receiving sitagliptin **100 mg/d**"
- "treated with insulin (**bolus component <30%**)"

2. A **time period** that relates to a METHODOLOGY_INCLUSION_CRITERIA annotation.
   - ">1 COPD exacerbation **in the previous 12 months**"
   - "not during the previous month"
3. An indication that participants **do or do not** meet the inclusion criterium
   - "with or without"

METHODOLOGY_STUDY_SIZE

Introduction	
Methods	
Results	Only if the element is not mentioned in the Methods section and/or Introduction.
Conclusion	
No section designation	Annotation the sections that are contextually clear to be part of Methods or, in accordance with the exceptions, Results

When?

1. **How many participants** are included in the study. Ideally, the total number of intention-to-treat (ITT) participants is annotated, which refers to the number of participants after randomization, but before start of treatment. If the number of ITT participants is not mentioned, the number of participants at screening or administration should be annotated.
   - "**14** consecutive outpatients were randomized"
   - "**Two hundred seventeen** patients were randomized to treatment"
   - "We randomly assigned **101** patients to receive placebo (n=50) or benralizumab (n=51), of whom 88 patients completed the study"
   - "Of 718 patients randomised, **703** were in the ITT analysis"

METHODOLOGY_STUDY_DURATION

Introduction	Only if the element is not mentioned in the Methods section.
Methods	
Results	
Conclusion	
No section designation	Annotation the sections that are contextually clear to be part of Methods or Results or, in accordance with the exceptions, Introduction

When?

1. The **duration of time** that is mentioned, usually with study characteristics.
   - "randomized, **52-week**, open-label"
2. **Other time indications** that are mentioned in the methods section, which might be identical to the elements from rule 1.
   - "Responders continued on ibrutinib maintenance for up to **2 years**"
3. The **longest duration of time** that relates to the study duration in the methodology. This could be mentioned together with the outcomes.
   - "Key secondary endpoints included quality of life and partly stable COPD at **26 days**"
4. **Follow-up** duration, this is usually mentioned in the results section.
   - "After a **median follow-up of 7.6 years**, 79 men in the endocrine alone group had died of prostate cancer."

METHODOLOGY_OUTCOME

Introduction	Only if the element is not mentioned in the Methods section.
Methods	
Results	
Conclusion	
No section designation	Annotation the sections that are contextually clear to be part of Methods or, in accordance with the exceptions, Introduction

How?
A METHODOLOGY_OUTCOME is annotated without references to time, measured values rather than the parameter, or references that are not necessary for another reason (such as “maximum”, “self-measured”)

When?

1. A **clinical outcome** that is measured during the study. This could include blood values of a drug in a pharmacokinetic study.
    - “Additional analyses included **transition dyspnea index** (**TDI**), **health status** (**St George's Respiratory Questionnaire** [**SGRQ**]), and **exacerbations**”
    - “Secondary end point was the proportion of patients with **severe hypoglycemia** during the maintenance period”
    - “**viral load**”
    - “Thus, the main aim of this clinical trial study was to determine the expression of **IL-6**, **IL-8**, **tumor growth factor (TGF)-β**, **tumor necrosis factor (TNF)-α**, and **endothelin (ET)-1**)”
    - “For **PK** analysis of **IAsp** (maximum **serum IAsp** concentration) blood samples were drawn immediately before baseline.”
2. An umbrella term for a **collection of parameters** that is measured.
    - “**safety**”
    - “**pulmonary function**”
    - “**pharmacokinetics**”
3. **Elements** of a composite outcome as well as the full **composite outcome**.
    - “Secondary end points include **major coronary events** (**death from coronary heart disease** or **myocardial infarction**) and **total coronary events** (**death from coronary heart disease** or **any coronary revascularization**)”
4. A **measurement method** that is used to measure a parameter.
    - “**ELISA**”
    - “**electrocardiogram**”
    - “**laboratory analysis**”
5. A **sample** that is used to measure a parameter.
    - “**blood samples**”
    - “**nasopharyngeal sample**”
6. An indication of a **pharmacokinetic expression**
    - “The primary endpoint was FEV(1) **area under the curve** for the time period 0 to 12h (**AUC** (0-12))”
    - “**Cmax**”
7. A reference to a **statistical analysis**
    - “**ANCOVA**”
    - “**fully adjusted Wald chi (2) test**”

When not?

1. A **time indication** of an outcome.
    - “The primary endpoint was sustained virologic response at **12 weeks**.”

PARAMETER

PARAMETER_EFFECT

Introduction	
Methods	
Results	
Conclusion	
No section designation	Annotation the sections that are contextually clear to be part of Results

How?

- Always together with one or more annotations of RESULT_VALUE or RESULT_SIGNIFICANCE.

When?

1. A **parameter** that is measured during the study, as complete as possible.
    - “**FEV1**”
    - “**progression - free survival**”
    - “More patients in the mFOLFOX6 arm **withdrew** from the study”

- "78.5% of the patients were **women**"

When not?

1. See PARAMETER DETERMINATION.

PARAMETER_BASELINE

Introduction	
Methods	
Results	
Conclusion	
No section designation	Annotation the sections that are contextually clear not to be part of Conclusion

How?

- The PARAMETER_BASELINE label is comparable to PARAMETER_EFFECT, except that results of elements of PARAMETER_BASELINE were collected at the start of the study. Therefore, every annotation of PARAMETER_BASELINE should be accompanied by an element of RESULT_VALUE or RESULT_SIGNIFICANCE.

When?

1. A complete description of a property of the study group at the **start of the study.**
    - "Patients had mild to severe TBI (n = 33) reporting sleep disturbances post-injury (mean **age** 37 years, standard deviation 11 years; 67% **men**)"
    - "A majority (80%) were **HCV treatment-naive**, and 84% were **infected** through perinatal transmission"

PARAMETER_DETERMINATION

Introduction	
Methods	
Results	
Conclusion	
No section designation	Annotation the sections that are contextually clear to be part of Results

How?

- Always together with a PARAMETER_EFFECT or PARAMETER_BASELINE annotation.

When?

1. **Small additions** to the PARAMETER_EFFECT ("achieved", "reached", "occured", "attaining").
2. **Time indications** related to a parameter
    - "**After 24 weeks**, 61.2% of patients were ASAS20 responders."
    - "**during Ramadan**"
    - "**until day 13**"
    - "**24-h** UGE"
    - "**Baseline** QTc interval was 406–418 milliseconds"
3. Determination of **causality** of a parameter.
    - "Seven patients dropped out **for administrative reasons**"
    - "two patients died (**unrelated to implant, system, or therapy**)"
    - "1 patient died **due to herpes simplex infection**"
    - "the rate of **genital infections that led to** discontinuation were 0.9% vs 0.1%"
4. A **subgroup** within the study
    - "**patients with genotype 4 HCV**"
    - "**naive patients**"
    - "**Caucasian subjects**"
5. Additions to the PARAMETER_EFFECT that are **mentioned separately** in the text.
    - "12 adverse events were reported, including **neutropenia** and **anaemia**".
6. Umbrella term for a number of parameters.
    - "**Adverse events** included diarrhoea (5 patients) and syncope (3 patients)"
7. Element that makes a distinction between different results.
    - "5 and 10 patients had **moderate** and **severe** COPD, respectively"
    - "Risk of **nocturnal** and **overall** hypoglycaemia was, respectively, 50% and 18% lower"
8. **Pharmacokinetic expression** related to a parameter
    - "Canagliflozin reduced **AUC** RaO by 31%"
    - "Co-administration of peficitinib increased rosuvastatin **maximum plasma concentration**"

9. **Method** used for measurement of the parameter
    - "Most patients had high disease activity state by the **ASDAS** (25 out of 31)"
    - "Mean titers **measured by ELISA** were 804.1"

When not?

1. If the time indication is part of an abbreviation.
    - "SVR**12**"

RESULT

RESULT_VALUE

Introduction	
Methods	
Results	
Conclusion	
No section designation	Annotation the sections that are contextually clear to be part of Results

How?

- Always together with a PARAMETER_EFFECT.

When?

1. **Result** of a measurement
    - "**5** exacerbations"
    - "HbA1c decreased by **2**%"
    - "**three** patients"
2. If a precise result is not present, the **range** of values suffices.
    - "ranging from **87 to 103** ml"
    - "few genital infections were reported **8.6-9.2**%
3. **Difference** between results of different groups.
    - "mean difference was **3.13**"
4. A **ratio** used to indicate a difference.
    - "estimated hazard ratio of **0.77**"
5. **Total group size** that refers to the group in which the result value is measured.
    - "5 out of **15** patients died"

RESULT_BASELINE

Introduction	
Methods	
Results	
Conclusion	
No section designation	Annotation the sections that are contextually clear not to be part of Conclusion

How?

- Always together with a PARAMETER_BASELINE.

When?

1. **Values** of characteristics of the population at the start of the study.
    - "baseline median plasma HIV-1 RNA was **4.89** log 10 copies/mL"
    - "**61.2**% had prediabetes"

RESULT_VARIABILITY

Introduction	
Methods	
Results	
Conclusion	
No section designation	Annotation the sections that are contextually clear not to be part of Conclusion

How?

- Always together with a RESULT_VALUE or RESULT_BASELINE.

When?

1. A **variation (deviation)** of a result or of a difference between results.

- "decrease by -0.66 ± **0.05** %"
- "hazard ratio, 1.03; 95% CI, **0.91 to 1.17**",
- "LSM (± SEM) of 190 ± **28**"
- "Patients maintained suppression for **between 15 and more than 30** weeks (median of 21 weeks)."

2. A **range** or **interval** in which the results fall
    - "median (range) exposure was 8 (**5, 13**) days"

When not?

1. When a variability value is given without a result, in this case the RESULT_VALUE/RESULT_BASELINE label should be used.
    - "ranging from **87 to 103** ml"

RESULT_SIGNIFICANCE

Introduction	
Methods	
Results	
Conclusion	
No section designation	Annotation the sections that are contextually clear not to be part of Conclusion

How?

- Always together with a PARAMETER_EFFECT or PARAMETER_BASELINE.

When?

1. **Significance** of a difference between groups or timepoints
    - "hazard ratio, 0.78; P **= 0.007**"
    - "Symptomatic hypoglycaemia was more common with glargine (p **< 0.001**)"
2. **Indication of significance** of a difference between groups or timepoints
    - "there was **statistically significant** difference between the groups"
    - "blood glucose levels did **not** differ **significantly**"

RESULT_UNIT

Introduction	
Methods	
Results	
Conclusion	
No section designation	

How?

- Always together with a RESULT_BASELINE, RESULT_VALUE, RESULT_VARIABILITY or THERAPY_DOSE_REGIME.

When?

1. A unit in which a **result** is expressed.
    - "rate of exacerbations was 8.4**%**"
    - "488 (8**%**) of **patients** experienced the primary endpoint"
    - "plasma HIV-1 RNA ranged from 1.21 to 1.73 **log (10) copies/mL**"
2. Indication of a **ratio**
    - "**hazard ratio** for disease progression or death, 0.58"
    - "SRI-4 response rate was 53.8%; **adjusted OR** 1.56"
3. A unit in which the **variation** of a result is expressed.
    - "HbA1c was -2.3% ± 1.5**%**"
4. The unit of a dose.
    - "randomly assigned to perindopril 8 **mg**"

RESULT_DETERMINATION

Introduction	
Methods	
Results	
Conclusion	
No section designation	Annotation the sections that are contextually clear not to be part of Conclusion

How?

- Always together with a RESULT_VALUE, RESULT_BASELINE or RESULT_SIGNIFICANCE

When?

1. As a determination of a **result** or **baseline measurement**.
   - "**Mean** age of patients was 60 years"
   - "primary outcome occurred in 5.7% of **patients**"
   - "**Mean** percent **change** in LDL-C was 16.7%"
2. As a determination of a **difference** between groups.
   - "between-group **difference** in FEV1 was 120 ml"
3. As an indication of the **direction** of difference
   - "HbA1c **decreased** in sitagliptin groups vs placebo (-0.7%)"
   - "ACR20 response was **lower** with sirukumab vs placebo (69% vs 21%)
4. As a determination of **significance**.
   - "There was a significant between-group **difference** in SF-36 scores"
   - "The **mean** 24-h UGE significantly **decreased**"
   - "Myocardial infarction occurred **less** in patients receiving nifedipine ($p < 0.05$)"
5. As a determination of **variation**.
   - "median duration of response was 17.2 months (**95% CI**, 78 to 87)"
   - "onset of joint swelling was 12 days [**IQR** 10-14]"
   - "HbA(1c) decreased on glargine treatment (-1.72%, **SE** 0.06)

THERAPY

THERAPY_DOSE_REGIME

Introduction	
Methods	
Results	
Conclusion	
No section designation	

When?

1. **Dose** of a drug
   - "317 patients were assigned to the **70** – mg erenumab group"
   - "Patients received MXR 1000 mg **QD** or continued MIR 500 mg **BID**"
2. **Frequence** of administration.
   - "Patients received 1000 mg/m^2 **weekly** of gemcitabine alone"
3. **Moment** of administration
   - "To assess the safety and efficacy of **perimenstrual** telcagepant"
4. **Duration** of treatment
   - "Group 1 received velpatasvir for **14 weeks**"
   - "Patients received treatment for **5 consecutive days**"
   - "87% of patients treated for **12 weeks** achieved SVR12"
5. **Intensity** of a treatment
   - "133 underwent chemotherapy (**54** Gy)"

When not?

1. If the duration of treatment also refers to the **study duration**. In that case, the element is annotated as METHODOLOGY_STUDY_DURATION

THERAPY_METHOD_OF_ADMINISTRATION

Introduction	
Methods	
Results	
Conclusion	
No section designation	

When?

1. **Route** that is used to administer a drug
    - "Patients were randomized to AVP-825 plus **oral** placebo tablet"
    - "**Inhaled** corticosteroids are standard treatment in patients with COPD."
    - "160 patients received ALA – PDT by means of **needle-free injection**"
2. **Actions** that are performed with a formulation before administration.
    - "A tablet containing GLE and PIB was **crushed** before administration"

DISEASE

DISEASE_INDICATION

Introduction	
Methods	
Results	
Conclusion	
No section designation	Annotation the sections that are contextually clear to be part of Introduction or Conclusion

When?

1. Every form of a **disease** or **symptoms**, including severity
    - "**moderate to severe COPD**"
    - "**advanced or recurrent endometrial carcinoma**"
    - "**major adverse cardiovascular events**"

When not?

1. If the disease is formulated as an adjective to the patient.
    - "**type 2 diabetic** patient"

ID

ID_TRIAL

Introduction	
Methods	
Results	
Conclusion	
No section designation	

When?

1. A clinical trial **identifier**, including references to the database.
    - "**NCT 00615030**"
    - "**ISRCTN 47533126**"
    - "**EudraCT 2011 - 004804 – 38**"
2. **Name** of a clinical trial, usually an acronym.
    - "**FULFIL**"
    - "**KEYNOTE – 052**"