15 September 2020
Version. 1.0
EU-SRS

Data Cleansing Manual Substance Validation Group (SVG)

Guidance on EU Substance Data Cleansing as part of the EU-SRS project

External version: living document

Table of contents

Glossary

Abbreviation	Explanation
ATC	Anatomical Therapeutic Chemical Classification System
BAN	British Approved Name
CAS	Chemical Abstracts Service
EMA	European Medicines Agency
EUTCT	European Union Telematics Controlled Terms
EU-SRS	European Substance Registration System
EV	EudraVigilance
EVVET	EudraVigilance Veterinary - System for the exchange, processing and evaluation of suspected adverse reaction reports (SARs) related to veterinary medicines authorised in the EEA.
FDA	Food and Drug Administration
GSRS	Global Substance Registration System
INN	International Nonproprietary Name
ISO IDMP	ISO Identification of Medicinal Products
IUPAC	International Union of Pure and Applied Chemistry
JAN	Japanese Accepted Name
JIRA	Tracking system used at EMA to manage incidents, requests and questions
NCA	National Competent Authority
NCATS	National Center for Advancing Translational Sciences
OMS	Organisation Management Service
Ph. Eur.	European Pharmacopoeia
PMS	Product Management Service
RMS	Referentials Management Service
SIAMED	EMA system for managing product information and application tracking
SmPC	Summary of Product Characteristics
SMS	Substance Management Service
SMS-IDD	Informatica Data Director - System used in EMA for substance registration
SPOR	EMA – Substance – Product – Organisation – Referential
Sporify	Tool used to perform data cleansing
SSG	Specified Substance Group
SVG	Substance Validation Group
SVG-WG	Substance Validation Group – Work Group
USAN	United States Adopted Name
USP	United States Pharmacopeia
WHO	World Health Organization
xEVMPD	Extended EudraVigilance Medicinal Product Dictionary

1 Document Control

1.1 Document Version History

Table 1 contains an overview of the major revisions to the Data Cleansing Manual.

Table 1 Major versions of the Data Cleansing Manual Substance Validation Group

Date	Main author	Reviewer	Section	Version
26 June 2020	Bjorg Overby, Inti van Eck, SVG	SVG, EMA	Public version	Version 1.0

2 Introduction

The EU Network is currently implementing the ISO IDMP standards in a phased programme based on the four domains of master data in pharmaceutical regulatory processes: substance, product, organisation and referential (collectively referred to as "SPOR") master data. ISO IDMP compliant business services for the central management and supervision of data in each of the four SPOR areas will be established through an iterative and incremental delivery approach. Through the Substance Management Services (SMS) of the SPOR programme EMA will provide the EU network centralised substance data management services.

The EU-SRS project aims to form the scientifically rigorous back-end for the Substance Management Services of SPOR. The aim is to create an accessible EU Network wide, shared, structured database, referred to as EU-SRS, for the unambiguous identification of substances used in medicinal products based on their scientific properties in accordance with ISO IDMP standard 11238 and ISO IDMP technical specification standard 19844. These resources enable the unique identification of substances for various purposes including the enhancement of traceability of pharmacovigilance, non-clinical, clinical and quality findings with a high degree of precision to substances by their scientific identity.

One of the intentions of the EU-SRS project is to cleanse EMA substances data for all Substance Types, both for human and veterinary- specific substances by a group of European Substance Experts: the Substance Validation Group (SVG). The outcome of the data cleansing activity will be made available in SMS at regular intervals during the project.

This document has been written to serve as a reference during data cleansing activities performed by the SVG and to ensure alignment between the SVG and EMA. It aims to provide practical guidance and examples to handle different Substance Types during data cleansing activities. This manual is a living document and only provides guidance to Substance Types for which data cleansing has been initiated. Currently, the document has been written based on experiences with regards to Chemicals and Proteins and only these substances are included in the document.

The different chapters of the document describe various aspects needed to perform data cleansing:

- Data cleansing process (high level)
- General data cleansing guidance
- Substance Type specific guidance

2.1 Data cleansing purpose

Data cleansing is an activity performed in order to ensure that we will have one list of uniquely identified substances, and that that list is of good quality. During data cleansing of Chemicals and Proteins, several checks are performed with the purpose of having a list of substances that have at least:

- An EUTCT code (same as SMS ID)
- A Substance Type
- One Preferred Term that corresponds to the most appropriate name available for a given substance
- Aliases, if available, according to valid reference sources

During data cleansing of Chemicals, additionally the chemical structure is verified. During the data cleansing activities, the existing US publicly available GSRS database is used for reference purposes where possible, as the intention exists to utilize parts of GSRS data during the data load of substances in EU-SRS. Therefore, when possible, a match between EUTCT code and the US UNII code is made.

2.2 Involved parties

The SVG consists of substance experts or assessors from several European NCA's, EMA and WHO-UMC. Currently, the following NCA's are involved: MPA (Sweden), ANSES (France), SUKL (Czech Republic), MEB (Netherlands), AGES (Austria), BfArM (Germany), NoMa (Norway), AEMPS (Spain) and JAZMP (Slovenia). Additionally, there is a close cooperation with FDA/NCATS. The SVG contains experts that are focussed on both human and veterinary substances.

3 Data Cleansing Process

Data cleansing as part of the EU-SRS project is performed per Substance Type and the data cleansing process can therefore differ slightly per Substance Type. This chapter describes the high-level process for both Chemicals as well as Proteins, that are being cleansing in the Sporify application.

Sporify is an application displaying all substance records in an Excel like table and allowing users to propose changes in a traceable manner. The Sporify application was selected for data cleansing purposes as it ensures that original substance data cannot be changed, while it is able to combine data from different sources, displaying these in one overview (EMA and GSRS data).

Data Cleansing is seen as the process that starts with gathering data to be cleansed by the SVG until the processing of changes in SMS.

The data cleansing process is divided in the following high-level activities:

1. Data gathering and preparation
2. Data cleansing
3. Review
4. Upload in SMS

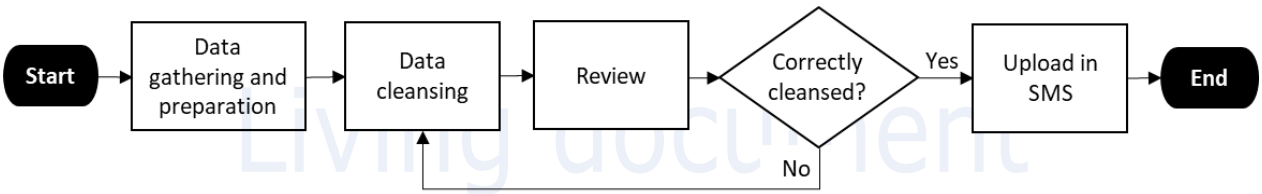

Figure 1 High level data cleansing workflow

3.1 Data gathering and preparation

Before the start of data cleansing, the SVG discusses and agrees on a framework for the cleansing of that specific Substance Type. Such discussions cover:

1. Substance data to be cleansed and gathering of that data
2. Existing sub classes within the Substance Type and respective prioritization
3. ISO IDMP guidance, official reference sources and literature to be followed when cleansing the Substance Type
4. Signature fields for the Substance Type and its sub classes (i.e. the minimum fields necessary to uniquely identify a substance)
5. Future data elements and hierarchy of the Substance Type in EU-SRS and how this influences the data cleansing
6. Training needs of SVG members
7. Data cleansing naming rules, substance class specific rules to be followed during data cleansing

With regards to human Chemicals and Proteins, the dataset to be cleansed originated from a compiled list containing xEVMPD and EUTCT substances. The file received from EMA contained approximately

60.000 substance records, of which about 50% were Chemicals and Proteins. As data cleansing for human Chemicals and Proteins was performed in the Sporify application, the full dataset was loaded in this system.

The following filters were applied to get the EMA dataset for cleansing purposes:

- Substance authorisation status = "Authorised" in EUTCT = "Approved substances" in XEVMPD (i.e. this status means that the substance has been registered by EMA and it is unrelated to product Marketing Authorisation status).
- Language = N/A or English
- Veterinary only records are excluded (from the initial data load)
- There is a 1:1 match of EUTCT codes and xEVMPD codes

During the project, on a regular basis, the dataset will be checked for updates as new substances are constantly being registered and existing substances being updated. For any future dataset extracted, the same filters will be applied.

The dataset is uploaded into Sporify using a subset of the columns available in the EMA dataset (EUTCT Code, EUTCT Substance Name and Substance Type). Once uploaded, the Sporify system tries to match the EUTCT Substance Name to the GSRS Substance Name. In case a match is found, GSRS information is shown in Sporify as well, like the UNII code and the structural formula.

3.2 Perform data cleansing (Sporify)

Cleansing of Chemical substances and protein (i.e. Monoclonal Antibodies and Fusion Proteins) substances is performed in Sporify per EUTCT code. The GSRS substance data is used as a reference. Within this process, an open line of communication exists with the FDA/NCATS team responsible for maintaining the GSRS database.

After opening Sporify, Substances are a module listed under Dashboard and the cleansing. The name of the list in Sporify used for cleansing, is 'EU-SRS Data Cleaning'.

The data cleansing process consists of 7 main steps, as listed in the figure below, in which all records belonging to a specific EUTCT code and any related records are verified for several aspects. Aspects that are being verified are:

- EUTCT code
- EUTCT Substance name type (i.e. Preferred term, alias, or English translation)
- EUTCT Substance name
- Match with GSRS substance (taking into account differences in reference sources; in GSRS the USAN name is used as the Preferred Term)
- Chemical Structure (in case of a chemical)

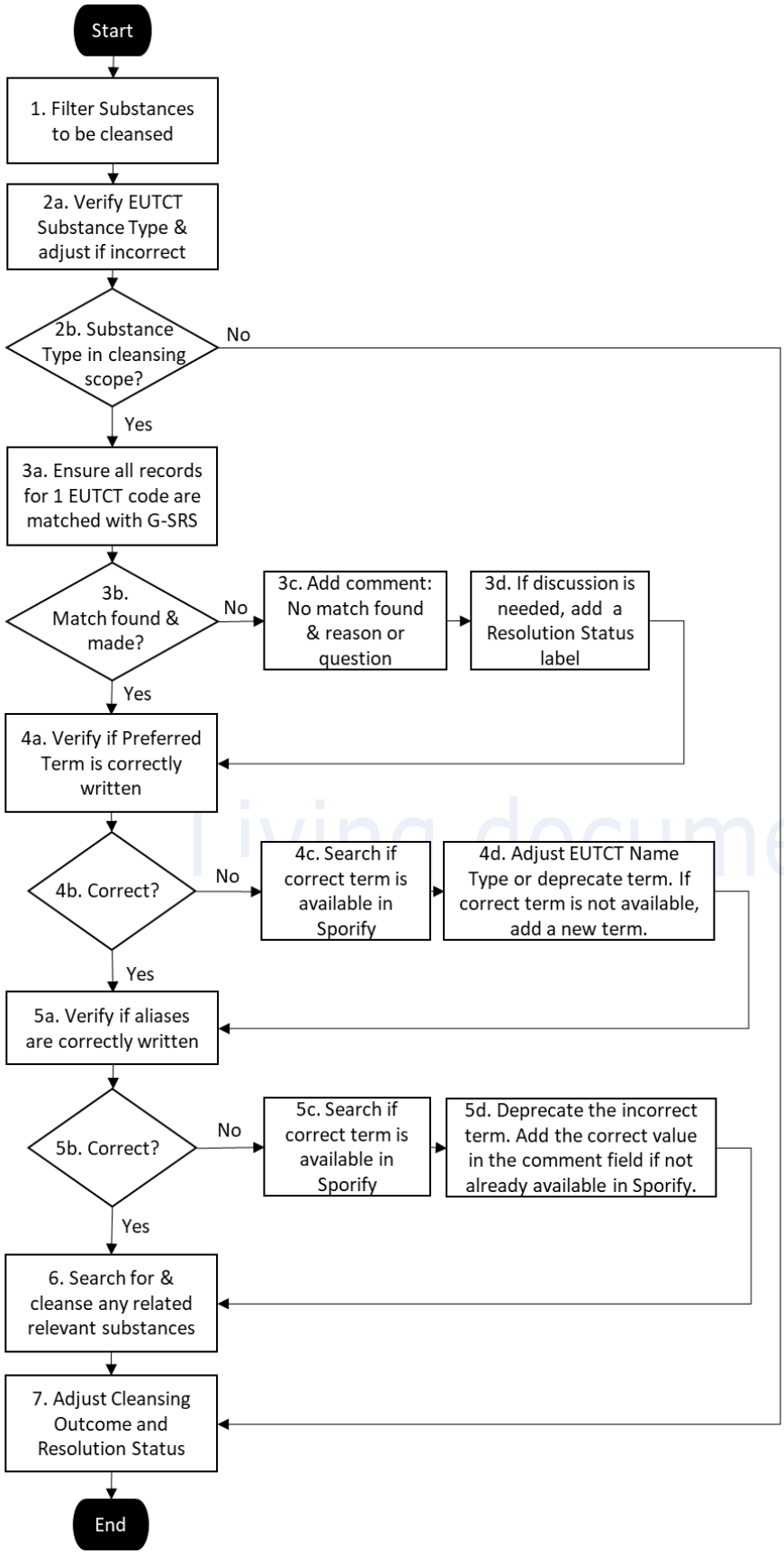

Figure 2 Data Cleansing Workflow

3.2.1 Cleansing workflow steps

The detailed Data Cleansing Workflow steps are described in the table below.

Step	Action
1	Filter substances to be cleansed

Sporify contains *tags* with names of all SVG members. Substances to be cleansed are distributed by the SVG coordinator, by adding a name to all records belonging to 1 EUTCT code. Filtering by *tag* provides the overview of records to be cleansed.

For each EUTCT code, data cleansing is performed for the Preferred Term first, followed by its aliases.

Note: if a case is found where a record is partly assigned to someone else, the name of the other person can be removed and replaced with another name.

2	Verify EUTCT substance type and adjust if incorrect

The first check performed is to determine if the EUTCT substance type is correct.

- *EUTCT substance type* is correct: continue with step 3.
- *EUTCT substance type* is incorrect: add the correct substance type under *EUTCT adjusted substance type*. Determine if this new substance type is in scope of current cleansing activities.
    - If yes, continue with step 3.
    - If no, set the *Resolution Status* to *Adjusted substance type* for the Preferred Term and all its aliases and stop cleansing. The *Cleansing Outcome* status does not need to be adjusted and can be left to *Not set*.

3	Ensure all records for one EUTCT code are matched with GSRS

During the cleansing activities, valid reference sources, the ISO standard and rules outlined in this manual should be used to determine correct naming. One such reference source is GSRS. Sporify automatically tries to match EUTCT data to GSRS data, based on a 1-to-1 match in name. If a match is found, information from GSRS is pulled into Sporify and displayed in columns where text is coloured blue. For records where GSRS displays a chemical structure, the chemical structure is displayed in Sporify as well.

It is important that all EUTCT records in Sporify (Preferred Terms and aliases) are matched with GSRS data, as it is intended that a future data load in EU-SRS would pull additional data from GSRS for each EUTCT record. By doing so, the basis of European substance data will be all that has been cleansed, but with an enrichment in additional substance information, that were not part of data cleansing (or not even available in EU substance data).

In matching substances with GSRS or verifying if the match is correct, it is important to note that names do not always match 1-to-1. The Preferred Name in GSRS follows USAN naming, which could lead to different writing of the substance name because the EU follows INN and Ph. Eur. Therefore, it is important to always verify that the automatic match in Sporify is correct.

Note that there are cases where a substance is not found in GSRS. Usually this is the case of substances under development or used outside the EU. In such cases, EMA can investigate the substance to provide more information.

Matching the *EUTCT substance name* with the *GSRS Substance name* has the following scenarios:

- A match is already available: the match is to be verified and data cleansing can be continued once it is verified that the match is correct

Step	Action

- A match is not available and Sporify provides a suggestion: the suggestion is to be verified:
    - The suggestion is correct: the suggestion is to be added and data cleansing can be continued
    - The suggestion is incorrect, or the existing match is incorrect and no correct match can be found: see 'A match is not available' and remove the existing match by deleting the term and then clicking on a different row in the list.
- A match is not available and Sporify does not provide a suggestion: a match is to be found by searching in GSRS or other public databases.
    - In case a substance is found in GSRS: the substance can be linked to the *EUTCT substance name* by typing the UNII code or the name in Sporify
    - In case no match is found: add a comment and add the *Resolution status* label 'Ongoing'. Add a tag 'EMA', to specify that the record needs to be checked by EMA. Remove your name from the tags. Make sure the *Resolution status* is added to all records belonging to the EUTCT code. One can decide to already verify the rest of the information in the record where possible, to ease the review done by EMA.

Note: Adding or adjusting a match with GSRS, does not affect the *Cleansing Outcome* status, as this is not seen as an actual change in the record. In other words: a change in matching with GSRS should not result in the *Cleansing Outcome* of *Record Changed*.

Note: In general, a UNII code and its accompanying *GSRS name* should not be assigned to more than 1 *EUTCT id*, as this will provide issues with the future data load.

4	Verify if the Preferred Term is correctly written

For each EUTCT code, a few checks need to be performed for its Preferred Term and Aliases:

- Ensure that the record has 1 Preferred Term
- Ensure that the Preferred Term is written correctly according to reference sources, ISO and the data cleansing manual
- Ensure that the *EUTCT substance name* matches with the *Chemical structure* (if applicable)
- Verify the *EUTCT substance name type*
- Verify or add the *EU-SRS Substance name type*
- Verify the *chemical structure* (if applicable)
- Verify that there are no translations in the list
- Verify the *EUTCT ID*
- Verify that the record can be uniquely identified based on its Preferred Term and aliases. If not, an additional record will need to be added into Sporify (applicable mostly for records with only a Company Code)

In case the Preferred Term is not written correctly, another record should be made to capture the Preferred Term, and the existing record should be changed into an alias or, if deemed invalid, set to be deprecated. However, before adding a new record, perform a search by name in Sporify, to make sure the name does not already exist. Newly added substances are written with only the first letter

Step	Action

capitalized. Note: once a record is added, it cannot be removed. If a mistake was made in the naming, the record will need to be deleted and re-added.

Note that there have been many discussions with regards to INN versus Ph. Eur. In case there are questions with regards to the name to be chosen, a discussion can be started. Additionally, when a Ph. Eur Name points to multiple substances a discussion needs to be held. This is also the case for records where the pharmacopoeia name in other sources reflects a different substance and the two substances are different records in EUTCT.

For aliases that are not correctly written, it is not needed to add a new record, but the correct name can be added into the comment field.

In case changes are made to a record, it is advised to always add a comment explaining the reasoning behind it. This eases the review process later.

5 Verify if aliases are correctly written

The checks, as described under step 4, are to be repeated for the aliases belonging to an EUTCT code. This step differs from step 4 in one aspect:

- When an alias is incorrectly written and the term is deprecated, a new record does not need to be added. For aliases, we accept it when the correct value is listed in the comment field.

Example:

- Systematic name is written incorrectly, and no correct value is available in Sporify
- Resolution status: 'Review Completed'
- Cleansing outcome: 'To be deprecated'
- Comment field: correct writing of the systematic name

6 Search for and cleanse and related relevant substance

At the end of the cleansing process, SVG members make a proposal to EMA, how to adjust the substance in SMS by means of 2 statuses (*Resolution Status* & *Cleansing Outcome*). Additionally, tags can be added for filtering purposes.

Resolution statuses and *tags* are maintained by the SVG coordinator. In case an SVG member would like to make use of an additional *tag*, a request can be filed with the SVG coordinator.

For each record belonging to one EUTCT code and reviewed in steps 1-5, a *Resolution Status* and *Cleansing Outcome* should be added.

Tags usually involve the name of the SVG member that was assigned to cleanse the record. *Tags* do not need to be changed or removed and can be kept after a record has been cleansed.

Tags	When chosen
SVG member name	• To assign a record for data cleansing • Only 1 SVG member name should be assigned to any record

EMA	• In combination with the *Resolution status* 'Ongoing', when no information can be found with regards to the substance in public sources
SVG General	• In combination with the *Resolution status* 'Ongoing', for a quick consultation
SVG Coordinator	• In combination with the *Resolution status* 'Ongoing', when a discussion is needed in the broader SVG group (a quick consultation is not sufficient) or an adjustment is needed to the data cleansing manual
Ph. Eur. mismatch	• Can be added aside other tags to indicate that the INN and Ph. Eur name do not align

Cleansing outcome possibilities are:

Cleansing Outcome Status	When chosen
No Action Required	• It concerns a valid substance (Preferred term or alias) • The Substance name is correctly written • The *EUTCT ID* is correct • All data fields are verified, available and correct • An EUTCT code has 1 record with a Preferred Term
New Record Created	• The EUTCT code did not have a record with a correct Substance name and a new row was added

To be Deprecated	• The record should be removed because of:
	  • It is a translation
	  • It is an invalid substance
	  • It is a duplicate

Record changed	• Any changes made with regards to:
	  • *EUTCT ID*
	  • *EUTCT Substance Type*
	  • *EUTCT Substance Name*
	  • *EUTCT Substance Name Type*

The list of possible outcomes to choose from for the Resolution Status are:

Resolution Status	When chosen
Review Completed	• The full EUTCT code has been verified • The full EUTCT code is ready for a review

Step	Action
Adjusted Substance Type	• The EUTCT Substance Type was changed, and further cleansing is discontinued as the Substance Type is not in scope of the current cleansing activities. Note: in case the record is in scope of current activities and cleansing of a record with an adjusted substance type was finalized, the Resolution can be changed into 'Review Completed'.
Ongoing	• Data cleansing is ongoing and not finalized. This status should be combined with a *tag*.
For Discussion	• There are unclarities with regards to the cleansing of the record and further discussion is needed. A discussion has been held within the subgroup (Ongoing – SVG), but without satisfying result. A further deep dive is needed into the record by a substance expert.
Parked	• Any record type that will be dealt with in a later stage, due to agreed reasons, e.g.: ○ Records are easier to cleanse in one session, when all records have been gathered

Note: Make sure each record has a *Resolution status* and a *Cleansing outcome* at the end, or a record will not be reviewed.

7

Once all records belonging to 1 EUTCT code have been cleansed, SVG members perform an additional search in the list of substances to find any relevant or related substances with a different EUTCT code, that should be cleansed at the same time. Such checks could include:

- Salts or hydrates related to the substance reviewed
- Free base, free acid, active moiety check (especially for substances with only a company code)
- Check of expected duplicates, when common substance names are not listed under the original EUTCT code, but are expected to belong to a substance
  - Note: when a duplicate record is found, with a different EUTCT code, SVG members need to indicate this in the record by: adding a comment in Sporify, mentioning it concerns a duplicate. The *EUTCT ID* of the duplicate record is to be mentioned in the *EUTCT Adjusted ID* field. Note that the final choice which record is kept and which is to be deprecated, is made by EMA based on the impact on products and procedures linked to those IDs.

3.3 Review cleansed data

Once a full EUTCT code has been cleansed, the SVG coordinator initiates a review of all substances belonging to that EUTCT code. The review itself consists of 2 steps, after which the Sporify status is adjusted.

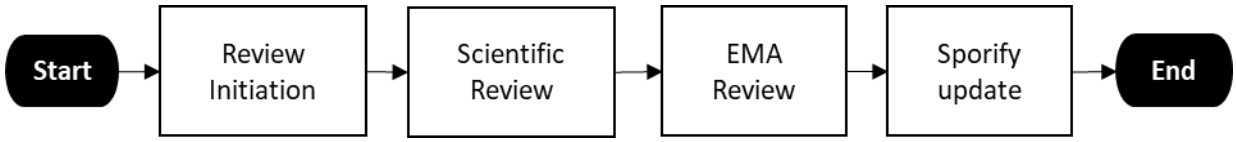

Figure 3 Workflow of the review process

3.3.1 Scientific review

The scientific review is performed by SVG members based on calculated risk:

- 100% review of records with a proposed change are reviewed
- 10% of Preferred Terms with no action required are reviewed by random sample
- 0% of aliases with no action required are reviewed

The review is performed in Sporify, guided by an Excel. SVG members receive an Excel file with 5 columns:

- EUTCT id
- Substance name
- Comment
- Originally cleansed by
- Cleansing outcome

3.3.1.1 The scientific review process

1. SVG members search for the EUTCT id in Sporify by copying the EUTCT id from the Excel
2. SVG members review each EUTCT id for its Preferred terms and aliases according to the data cleansing process and look at all fields that could have been cleansed.
3. In case a mistake is found, this is immediately updated in Sporify. For any record that was updated, a comment is made in the Excel file with the change made
4. Once finalized, the Excel file is returned to the SVG coordinator.

The SVG coordinator regularly shares an overview of all findings during the review with the SVG, so that SVG members can see what changes were made to their originally cleansed records.

3.3.2 EMA review

The EMA review concerns a second peer review and impact assessment of the proposed changes. The impact is assessed to determine what change needs to be made in SMS: changing the substance type, addition of new name, correction of a typo in a name, removal of a name, conversion of an alias to a translation, creation of a new substance or nullification. Additionally, the downstream impact is verified to see if any procedure or product record is impacted. Based on the type of change and its downstream impact, it is decided when the change could be made.

In case any mistakes are found during the review, the record goes back into the data cleansing process.

3.3.3 Sporify update

After a full EUTCT code has been cleansed, the records' status in Sporify is adjusted, so that these records are excluded from any future cleansing and review activities.

3.4 Upload in SMS

Once a record has gone through the review process successfully, changes in SMS will be made based on the EMA impact assessment, on regular basis during the project, with the exception of nullification. This functionality has not been developed in SMS at the time of publication of this version of the document.

It is the intention to make the EUTCT codes that have been cleansed and are processed in SMS public, together with a summary of the changes made.

4 General Data Cleansing Guidance

In order to ensure that data cleansing is performed in a harmonised way, data cleansing rules have been established and agreed upon within the SVG and EMA. References to external documentation are made where necessary. This chapter provides general guidance, independent of Substance Type. Specific guidance per Substance Type is listed in the specific Substance Type chapter.

The concepts required for the unique identification and description of substances are described in the ISO 11238 IDMP standard on substances. Guidelines for implementing ISO 11238 are provided in the technical specification ISO/TS 19844. Although ISO 11238 does not provide any guidance on substance nomenclature, it does provide a structure for the capture of names and codes that are used to refer to a substance. This section aims to provide supplementary guidance and should be read in conjunction with the standard and technical specification.

4.1 Substance Type

Substance Types are aligned with the Substance Types available in SMS[1], which is based on ISO IDMP. Substances shall be defined using one of the following terms:

- Chemical
- Mixture
- Nucleic acid
- Polymer
- Protein - Other
- Protein - Vaccine
- Specified Substance Group 1
- Specified Substance Group 2
- Specified Substance Group 3
- Specified Substance Group 4
- Structurally Diverse - Allergen
- Structurally Diverse - Cell therapy
- Structurally Diverse - Herbal
- Structurally Diverse - Other
- Structurally Diverse - Plasma derived
- Structurally Diverse - Polyclonal Immunoglobulin
- Structurally Diverse – Vaccine

[1] https://spor.ema.europa.eu/rmswi/#/lists/100000075826/terms

4.2 Name types

The EU-SRS system will contain name type information. During data cleansing, EMA data can therefore be enriched with a name type. ISO 11238 list several name types. Within the EU-SRS project, examples of name types used are:

- Common name, Company name, Multisubstance material, Official name, Scientific name, Specified substance group 1, Specified substance group 2, Specified substance group 3 and systematic name.

4.3 Naming convention

In SMS, each unique substance receives an SMS ID (EUTCT ID) and each EUTCT ID has one substance Preferred Term. The Preferred Term is the best substance name available at a given time and could change.

4.3.1 Hierarchy for Preferred Terms

The Preferred Term of a substance should be selected according to the priority ranking of the following reference sources and name types:

1. European Pharmacopoeia (Ph. Eur.) (Official Name Type)
   NOTE: There are cases were the Ph. Eur. name and the INN name are not aligned or where the monograph definition does not give sufficient level of depth. For these cases, we look at a case by case basis what name would reflect the substance best. Slowly, a list will be compiled in Annex I of this document displaying these cases. The Annex I with the first examples will be published in the next version of this document.

2. Recommended International Non-Proprietary Name (rINN) (Official Name Type)

   NOTE 1: An INN for a new chemical entity does not routinely specify the stereoisomeric state of the molecule in the non-proprietary name. If stereochemistry has been determined, then this information is presented in the chemical name(s) to identify the substance. An INN can, therefore, represent the racemic mixture (e.g. ibuprofen), the levo-isomer (e.g. amifostine), or the dextro-isomer (e.g. butopamine).

   NOTE 2: Details on how the INN names are established can be found here: http://origin.who.int/medicines/services/inn/stembook/en/

   NOTE 3: This includes Modified INN

3. Other official name type with EU jurisdiction (INCI, BAN, etc.)
4. Common name mentioned in the SmPC or PiL (Common Name Type)
5. International Union of Pure and Applied Chemistry (IUPAC) name (Systematic Name Type)
6. Other systematic name (Systematic Name Type)
7. Company code

   NOTE: A company code can be temporarily used as a Preferred Term only if no other name is available in the public domain, e.g. for substances under development. Once another name becomes available, the company code should be changed into an alias and another term should become the Preferred Term.

4.3.2 Aliases

Aliases are valid alternative names for a Preferred Term, according to valid reference sources. SMS provides aliases when available.

In addition to the sources/name types used for preferred terms, the following sources can also be used as an alias:

- Proposed INN (pINN) (Official Name Type)
- United States Approved Name (USAN) (Official Name Type)
- United States Pharmacopoeia (USP) (Official Name Type)
- Japanese Approved Name (JAN) (Official Name Type)
- Official name in other jurisdiction, e.g. AAN (Official Name Type)

In addition, other common names can be used as aliases when:

- It is cited as such in a valid reference source;
  Example: *Ascorbic acid = Vitamin C*
- The name is presented differently based on order of the words or when there is a comma or hyphen or brackets in the substance name:

  Example: *Fluoxetine hydrochloride = Hydrochloride fluoxetine*

  Example: *Calcitonin (Human) = Calcitonin, Human*

- The name contains an E-number.
  E-Numbers are acceptable as alias of an approved substance name and shall be written according to the Commission Regulation (EU) No 231/2012. In case there are multiple aliases with different writings of the E-number, one shall be kept with correct writing.

  Example:
  Preferred Term: Calcium hydroxide;
  Alias: Calcium hydroxide (E 526);
- The name is a Latin translation.
- The name is an American English writing.

EU English	US English	Comment
-oxide	-oxyde	Use of **"i"** in UK, and **"y"** in US
-ilate	-ylate	Standard ending see examples of use below
-f-	-ph-	Pronounce **"F"** see examples below
aluminium	aluminum	Translation from latin can be different
besilate	besylate	
camsilate	camsylate	
colour	color	

mesilate	mesylate	
sulfuric	sulphuric	
tosilate	tosylate	

Table 2 Examples showing different spellings between EU English and US English

Note: Translations in all EU languages are valid substance names and are registered in SMS. However, they are not in scope of this data cleansing exercise.

4.3.3 Invalid substance names

Any substance name that is not an alias as described in paragraph 4.3.2 and that is not available in any valid reference source is considered invalid and should be deprecated.

Not acceptable names include:

- Product names: Product names should not be inserted as substance names. This applies also in cases where in official reference sources they are reported as a synonym of the substance.
- Pharmaceutical product characteristics as part of the substance name; Pharmaceutical product characteristics reported as part of the substance name e.g. 'For Injection', 'For Solution for Infusion' are acceptable in the dictionary only if these are referring in a specific Pharmacopoeia monograph. Otherwise, the term is not considered to be valid.
  Example 2:
  Calcium gluconate 50 mg/ml is not a valid substance name. The strength should be expressed in the context of the product submission, as part of the pharmaceutical product information in the field relevant to the active ingredient or excipients.

  Note: The expression of the strength is different from the concentration of a substance; in this case the information can be included in the name according to the definition of Specified Substance Group 1. e.g.: Hydrochloride1N
- Substance names in in the form 'SUBSTANCE NAME (AS SOLVATED/SALT/PRODRUG)';
  Example 1:
  Clopidogrel (as hydrochloride) or *'Abacavir (as abacavir sulfate)* are not valid substance names. Instead, for *Clopidogrel (as hydrochloride)* the name *Clopidogrel* or *Clopidogrel hydrochloride* should be used. This applies to terms in English and translations.
  Example 2:
  Macrogol (PEG 400) or *Macrogol 400 (PEG 400)* are not valid names. The substance names should be *Macrogol 400, Polyethylene glycol 400* and *PEG 400.*
- Multiple substance names or Substance Type; A substance is considered not valid when the name refers to a class of substances or when more substances are listed (e.g. separated with commas, pluses).
  Example 1:
  Antihypertensives is referring to a therapeutic class and is not a valid substance name; *Antihistaminics* is referring to a therapeutic class and is not a valid substance name; *Herbals+ Vitamins + Minerals* is a multiple name also referring to a groups of substances and is not a valid substance name; *Vitamin C, Acerola, Propolis* is referring to a list of substances, which should be registered individually.
  Example 2:
  Substance names like *Vitamins NOS* and *Lipids NOS* (NOS = not otherwise specified) are not valid for the reason explained above.

Example 3:
Caramel 150 is considered not valid. This excipient should be further specified as *Caramel 150A*, *Caramel 150B* or *Caramel 150C* in accordance with Commission Regulation (EU) No 231/2012".
Example 4:
Codes that refers to more than one substance like acronym describing chemotherapy (*All BMF-86*). Exceptions: names referring to Substance Type but that are reported in individual case safety reports (ICSRs) must be retained in XEVMPD due to safety monitoring and public health purposes.
Example: *Immunoglobulins* and *Pancreatic Enzymes*. The Substance Type to be chosen for these records is to be defined but for the purposes of the cleansing 'Concept' should be used.

- Molecular formulas used as name
- e.g. HCl instead of Hydrochloride

4.4 General Data Cleansing Principles

- No records should be deleted from Sporify
- Substance names can be added to the existing substance list when the substance cannot be uniquely identified without the addition. If a systematic name is missing, the existing record can be extended with systematic names as an alias, where needed. The Systematic name may be a mandatory name in certain cases, e.g. when only a Company Code is available.
- EMA-original data will stay unchanged during data cleansing activities and any changes are made in a separate column
- In case changes to the record are proposed, or questions are raised, this should be clearly described in the Sporify comment field, to ease answering the question or reviewing the proposed change by other SVG members or EMA.
- When adding a new Substance in Sporify, only the first letter should be capitalized and no dots are used within names, e.g. "*Beta-damascone*" is found in GSRS as in ".*BETA.-DAMASCONE*"
- In case duplicate substances are found in Sporify, a comment needs to be added mentioning the EUTCT code of the duplicate in the record that is chosen to be retained. The final choice which EUTCT code is kept, is made by EMA.
- The EU-SRS Preferred Term should be written in European English. Any US English term is however to be kept as an alias.
- The Preferred Term at substance level should not contain a comma; commas are used to go to a different substance level. An exception to the rule is seen with vaccines.
  Example: *Codeine phosphate* is preferred over *Codeine, phosphate*

4.5 List of databases

When additional information concerning a substance is needed, the following databases can be used as a reference.

4.5.1 General

Database Name	Description
INN	The INN Programme assigns International Nonproprietary Names to medicinal substances through a broad consultative process. WHO is responsible for the INNs.
European Pharmacopoeia	The purpose of the European Pharmacopoeia is to promote public health by the provision of recognised common standards for the quality of medicines and their components. As these standards ensure that medicines reaching the market are safe for use by patients, it is essential that they are appropriate. Their existence also facilitates the free movement of medicinal products in Europe and beyond.
Medicines Complete	A site which guides on to several different publications, databases containing information about medicines.
Inxight: Drugs	Site provided by NIH, National Center for Advancing Translational Sciences. Information about e.g. treatment and pharmacology.
FDA Substance Registration System	Registration system in the U.S. by FDA and the U.S. National Library of Medicine (NIH), provides UNII-codes, Unique Ingredient Identifier.
G-SRS	Database built by GiNAS, NIH. This is the basis for the EU-SRS.
United States Approved Names	This is a site for USAN, where to find the approved names, provided by American Medical Association, AMA.
Japanese Accepted Names	This is a site for JAN, where to find the approved names, as part of the Japanese Pharmacopoeia.
PubChem	Chemical information from authoritative sources provided by U.S. National Library of medicine, NIH
European Union Food Additives	This database can serve as a tool to inform about the food additives approved for use in food in the EU and their conditions of use. It is based on the Union list of food.
EU CosIng	CosIng is the European Commission database for information on cosmetic substances and ingredients.
European Chemicals Agency	ECHA is an agency of the European Union and the site provides data from registration dossiers.
EC Active substance database	Site from the European Commission and it provides General index of products by active substance.
Merck Index	Online version of the Merck index, regarded as the most authoritative and reliable source of information on chemicals, drugs and biologicals. Now this trusted resource is available online from the Royal Society of Chemistry.
EU Orphan Database	Site from the European Commission and it provides The Community Register of orphan medicinal products.

FDA Orphan substance database	Site from FDA and it provides The Community Register of orphan medicinal products.
Index Nominum	This is an International Database of Pharmaceutical Substances and Preparations, provided by Wissenschaftliche Verlagsgesellschaft Stuttgart
International Pharmacopoeia	International Pharmacopoeia provided by WHO.
Scifinder	Research discovery application that provides integrated access to the world's most comprehensive and authoritative source of references, substances and reactions in chemistry and related sciences.

4.5.2 Proteins

Database Name	Description
UniProt	The mission of UniProt is to provide the scientific community with a comprehensive, high-quality and freely accessible resource of protein sequence and functional information.

4.5.3 Vaccines

Database Name	Description
International Committee on Taxonomy of Viruses	A site for information about viruses, it is also possible to send a question for help on this site.
WHO Influenza Vaccines	A site where information about Influenza Vaccines are published by WHO.
Influenza Research Database	A database updated by a project funded by The National Institute of Allergy and Infectious Diseases (NIH/DHHS), it provides a resource for the influenza virus research community that will facilitate an understanding of the influenza virus and how it interacts with the host organism, leading to new treatments and preventive actions.

4.5.4 Excipients

Database Name	Description
FDA Inactive Database	Site provided by FDA with a database of Inactive ingredients.
Colorcon	Site with information about excipients used for coating, colouring and solid dose design.

5 Chemicals

This chapter provides specific rules followed when cleansing Chemical substances. General Data Cleansing Rules, overarching all Substance Types are outlined in chapter 4.

5.1 Definition

The definition of Chemical substances is found in Annex B in the ISO-standard ISO/TS 19844.

5.2 Data cleansing rules

- The Preferred Term is written according to INN. If there is none, the Ph. Eur. is to be followed according to the Naming convention.
- The IUPAC name should not be selected as the substance Preferred Term, unless there is no other official name available.
- Company code should not be select as substance Preferred Term, unless there is no other other public name available.
- The order of the information on chemical substances is to state first the name of the active molecule followed by any additional information (hydration, salt ester).
  Example: *Pheneticillin potassium* is preferred to *Potassium pheneticillin*.
- If the substance does not exist as any hydrate, the addition "anhydrous" is superfluous. However, when a substance is a hydrate, then monohydrate or dihydrate is added.
  Example 1: *Naproxen Sodium* is preferred to *Naproxen sodium anhydrous*
- Molecular formulas are not acceptable to be provided as such or as part of the name and the full English name should be retained as preferred name.
  Example: *Tolycaine hydrochloride* is preferred to *Tolycaine HCl*.
- The active moiety corresponds to a different substance (i.e. different EUTCT ID) than the respective salts, esters, or hydration forms.
  Example:
  Iron sulfate; EUTCT Code 1
  Iron monohydrate; EUTCT Code 2
  Iron tetrahydrate; EUTCT Code 3
  Iron; EUTCT Code 4
- In the Preferred Term, ferrous/ ferric should be used. Within an alias, Iron (II) or Iron (III) is accepted, and the name will not be adjusted
- Enantiomer molecules should be entered as separate substances.
  Example:
  Verbenone; EUTCT Code 1
  D - Verbenone; EUTCT Code 2
  L – Verbenone; EUTCT Code 3
  DL- Verbenone; EUTCT Code 4
- According to the ISO 11238, irreversible changes in the underlying molecular structure of a substance are described as a modification of the antecedent material and the modification will typically result in a new chemical substance

- When two options are correct, we follow the approach of the European Pharmacopoeia (e.g. both cetyl alcohol and hexadecane-1-ol are correct, however the preferred term is cetyl alcohol)
- In case there are 2 or more moles of the base/ acid compared to the molecule of the salt, this needs to reflect the stoichiometric ratio's and the molecular weight. Both names are used, but the Di-active moiety form will be the name for the alias.
  Example: *Atorvastatin hemicalcium* is preferred to *Di-atorvastatin calcium*
- Ceramides used a nomenclature defined by INCI, whichh can be checked in the EC Cosmetic substances database.
  The nomenclature with numbers was decommissioned in 2014 and was replaced by letters.
  Example: *Ceramide NP* is preferred to *Ceramide 3.*
- Gangliosides are chemicals.
  Example: *Monosialoganglioside sodium*
- Naming of Isotopic inorganic salts should follow ISO.

5.2.1 Radiopharmaceuticals naming convention

The following naming convention should be used for radiopharmaceuticals, based on the INN:

Radionuclide being the Isotope number - the Element symbol - Carrier agent name

- In the absence of an INN, a Ph. Eur. monograph title should be specified.
- When the Ph. Eur. monograph title contains additional characteristics (e.g. Technetium (99mTc) bicisate injection) the full monograph title should be provided as the official name of the substance.
- The USAN name cannot be specified as a substance preferred name.
- A systematic name can be the Preferred Term, when there is no better name available according to the Naming convention.
- The radionuclide applies to the full name of the radioactive isotope whereas the isotope number and element symbol may vary from one isotope to another (e.g. Cobalt (56Co) or Cobalt (60Co)).
- The carrier agent name relates to any additional element linked to the radionuclide.
- The Preferred Term for a di-substituted benzene ring is according to this example: *4-toluenesulfonic acid* is preferred to "*ortho, meta, para" or "2-, 3-, 4-*"

5.3 Examples of correct naming

Examples of correct naming of Chemicals, based on the rules described above, are listed below.

- Hydrochloride: it contains the HCl-salt of a parent substance in which the amount of the salt moiety is not reflected in the name.

  -dihydrochloride means two HCl-molecules and so on

  - 'hydrate' in a substance name should be an 'alias'. The Preferred Term should specify hydrate with mono-, di- or x- hydrate. Only the term of 'hydrate' is allowed in case of a Non-stoichiometric substance or in case that the amount of water is variable. When this is the case

the record should have a property 'Water content' which is defining the amount of water in the substance. -monohydrate means one H2O-molecule in the crystal lattice.

Example: *Halometasone monohydrate* is preferred to *Halometasone hydrate*

- -di, -tri and so on describes more than one H2O-molecule; -sesqui describes 1.5 H2O; hemipenta describes 2.5 H2O
- For Organic Substances the syntax is the following:
  <Active moiety (Base or acid)>, <Salt>, <Hydrate>
  Only stoichiometric moieties are acceptable
  Exceptions are made for protein substances. Variable counter-ions are accepted in the name.
  In case of a non-stoichiometric relationship the percentage of the salt/ salt-hydrate must be provided.
- For salts of inorganic acids, the syntax is the following:
  the metal precedes the hydrogen (e.g. NaH2PO4). Molecules of water of crystallisation or of substances of solvation follow the formula of the salt. (e.g. H3PO4.5H2O).
  -If metal salts of inorganic acids include several metals, the symbols for the metals are shown in alphabetic order (e.g. K2NaPO4).
- In non-cyclic linear structures such as Sodium nitroprusside: Na2[Fe(CN)5(NO)].2H2O, a non-cyclic structure is constructed in the following order:
  - Symbol of the central atom is placed on the left
  - Ionic ligands with cations are placed before anions

Preferred Term	Alias	EUTCT Code
Lufenuron	Anhydrous lufenuron	EUTCT Code 1
Sulfosalicylate disodium	Disodium sulfosalicylate	EUTCT Code 2
Alpha-cypermethrin	Cypermethrin, (alpha-)	EUTCT Code 3
Trans-cinnamic acid	Cinnamic acid, (e-)	EUTCT Code 4
Menthyl acetate	Menthyl acetate, (+/-)-	EUTCT Code 5
Yttrium (90Y) edotreotide	Edotreotide [90y]yttrium	EUTCT Code 6
Tetracalcium dicitrate malate	Calcium citrate malate	EUTCT Code 7
Doxorubicin dihydrogen citrate	Doxorubicin citrate	EUTCT Code 8
Chloride Ion	Chloride anion Cloride (Cl-)	EUTCT Code 9
Naproxen Sodium	Naproxen sodium anhydrous	EUTCT Code 10
Platinum Dichloride	Platinous chloride platinum(II) chloride	EUTCT Code 11
Cetylphosphate potassium	Potassium cetylphosphate	EUTCT Code 12

Table 3 Examples of Correct naming

6 Proteins

General data cleansing principles and rules, relevant for all Substance Types, are outlined in Chapter 4. Specific Protein rules and decisions made are included in this chapter. Note that this chapter is currently written based on Monoclonal Antibodies and Fusion Proteins. Currently only the Substance type is verified for the protein vaccines and allergen proteins.

6.1 Definition

The definition of Protein substances is found in Annex C in the ISO-standard ISO/TS 19844.

Exceptions are made for protein substances. Variable counter-ions are accepted in the name. In case of a non-stoichiometric relationship the percentage of the salt/ salt-hydrate must be provided.

A protein is defined as a single unit of a linear amino acid sequence, or a combination of subunits that are either covalently linked or have a defined invariant stoichiometric relationship. This includes all synthetic, recombinant, and purified proteins of defined sequence, whether the use is therapeutic or prophylactic. This set of elements will be used to describe albumins, coagulation factors, cytokines, growth factors, peptide/protein hormones, enzymes, toxins, toxoids, recombinant vaccines, and immunomodulators.

Proteins and peptides are defined by their molecular structure based on the amino acid sequence, disulfide linkages, sites and a general type of glycosylation, based on the cell or organism type from which the protein was isolated from or produced (e.g. yeast, plant, mammalian, human). The method of production is generally not a defining element for proteins and peptides at Substance Information level. For a given non-glycosylated peptide, whether naturally isolated, produced by recombinant technology, or chemically synthesised, it will be defined as the same substance when there are no resultant differences in the amino acid sequence and disulfide linkages.

Amino acids are represented with upper case Letter Codes (also known as 'Dayhoff Codes') in accordance with the IUPAC 'A one-letter notation for amino acid sequences (Definitive rules)'.
Example:
Asparagine is represented by 'N' for the alpha-Amino acid in the L-configuration.
Asparagine is represented by 'n' for the alpha-Amino acid in the D-Configuration

6.2 Protein sub types

Protein sub types currently being recognized are (Note: this list is not exhaustive):

- Monoclonal Antibodies
- Fusion Proteins
- Insulins
- Allergen Proteins
- Protein-vaccines

Other possible sub types could be:

- Enzyme, receptor, peptide, monoclonal antibody conjugate, transporter, cytokine, growth factor, hormone, regulator protein, bispecific antibody, structural protein, cell adhesion protein,

toxin, coagulation factor, monoclonal antibody fusion protein, enzyme inhibitor, signal transducer (GTPase)

6.3 Data cleansing rules

- Names given to substances during an Orphan Designation procedure might not follow rules as highlighted in this manual or other official reference sources, like ISO. However, these names are to be kept in Sporify as an alias. Orphan Designation named substances can be recognized by searching for a substance in the Commission database, or via an Excel list shared by EMA.
- Peptides are described as chemicals up to 3 amino acids. A sequence of 3 or more amino acids are considered as a protein. This is in alignment with GSRS.
- Proteins that differ in protein sequence, type of glycosylation, disulphide linkages or glycosylation site shall be defined as two separate substances. Single protein substances are further classified as 'Protein – Vaccine' or 'Protein – Other'. Vaccines that contain protein subunits or recombinant proteins can be classified as 'protein – Vaccine'.
  Example: *Diphtheria toxoid*
  Note: For most Proteins the signal peptide is an integral portion of the final sequence. In a lot of cases the complete sequence is provided without the modifications. In a lot of cases the 'Final Expressed sequence' is not known or circulates in the blood in several stages, e.g. this is the case for Factor VIII/VWF complex.
- For Monoclonal Antibodies under development that do not have an INN yet, should have a common name or company code as preferred term.
- When information about the manufacturing process (e.g. recombinant, synthetic) is included in the substance name, this will have a distinct SSG1 Code EUTCT CODE, different from the single protein substance, and will be classified as Specified substance Group 1.
  EXAMPLE 1:
  Calcitonin salmon; EUTCT code 1
  Calcitonin bovine; EUTCT code 2
  EXAMPLE 2: In vaccine substances the nearly same approach applies. However, *Cholera toxin b subunit* should not be a synonym of *Cholera toxin b subunit, recombinant (rctb).*
- In the substance naming conventions of the Japanese Pharmacopeia the term 'GENETICAL RECOMBINATION' is the common part of the substance name for all recombinant substances.
  Example 1: *Pamiteplase (genetical recombination) (JAN)* is also known as *Palmiteplase (INN*). Palmiteplase is defined as a recombinant modified human tissue plasminogen activator; Therefore, the recombination is an integral part of the substance name. In this case *Pamiteplase (genetical recombination*) should be considered as a synonym of *Palmiteplase* sharing the same EUTCT Code.
  Example 2: The INN *Nonacog alfa* is defined as *Recombinant human coagulation Factor IX* therefore *Nonacog alfa (genetical recombination) (JAN)* is a synonym of *Nonacog alfa*.
- Monoclonal Immunoglobulins are described as proteins, polyclonal immunoglobulins shall be described as structurally diverse materials.

General rules applying for protein naming convention are listed below:

- The most recommended name is a word that ends with '-in';
  EXAMPLE: *zyxin*, *insulin*, *hemoglobin*, *caveolin*, *desmoglein*, *secretin*, etc.

- Names ending in '-ine' should be treated as synonyms;
  EXAMPLE: *maurocalcine* alias of *maurocalcin*.