General Rules	**Example**
Naming remained faithful to the description of the excipient in the Summary of Product Characteristics, rather than adhering strictly to chemical names.	"Brilliant blue". Note: this results in a small number of inconsistencies between British and American spellings.
Where multiple names refer to the same excipient, the more commonly used term in the SmPCs was used.	"Copovidone" rather than PVP/VA6:4
The abbreviations of excipients which are very infrequently used or have a long name were not given in full.	"DPPA" was used instead of 1,2-Dipalmitoyl-sn-glycero-3-phosphat**e**
When automated matching was applied to assign chemical identifiers to excipients, some polymers were assigned identifiers. These were not removed, as the PubChem and ChEMBL websites emphasize that these are repeating units.	"Methyl cellulose" is assigned CID 51063134
Discrete Structures:	**Example**
Different particle sizes of excipients with discrete structures were combined under one heading.	Micronized and coarse anhydrous lactose were described as "lactose"
Silica in oral formulations was described using limited terminology.	"hydrophobic colloidal silica," "silicon dioxide, hydrated," and "colloidal hydrophobic silica."
Hydrates and anhydrous forms were generally not combined	"lactose" and "lactose monohydrate"
Forms which differed due to the presence or absence of a counterion were generally not combined	"ascorbic acid" and "sodium ascorbate
All colors and oxidative states of iron oxide were grouped.	Iron oxide red was assigned "iron oxide"
Enantiomers were separated where specified	"tartaric acid" and "L-tartaric acid"
Polymers	**Example**
Different molecular weights or grades of a polymer were evaluated on a case-by-case basis but were largely combined into one commonly used term which identifies the monomer unit.	**"**Polysorbate"
Copolymers were generally grouped under one class for each monomer combination.	"poly-lactide-co-glycolide"
More detail regarding macroscopic structure was provided for very common polymers.	"cellulose", "cellulose powdered", "microcrystalline cellulose", and "dispersible cellulose" were separated.
All Opadry excipient systems were grouped under "opadry"	
Mixtures	
Common names were used in most cases	"heavy kaolin", "white wax"
Trade names of lipid formulation excipients were generally assigned to a single pharmacopoeial name where available.	"Medium chain triglycerides"
The source of the ingredient was retained where practical.	"Egg phosphatidylcholine" and "phosphatidylcholine" were separated.
Transdermal Formulations	**Example**
Given the nature in which these excipients are reported in the SmPCs, there was generally not great scope for grouping. These excipients were kept as terms which describe each layer	"low-density pigmented polyethylene outer layer", "non-woven polyester fabric", and "polyisobutylene/ polybutene adhesive"
Miscellaneous	**Example.**

Printing Inks and coloring agents: “Printing Ink” assigned to any regex match which referenced printing ink. This included matches which *only* contained the term printing ink, but also matches where substances were listed. Given the scope of this study, manually extracting these excipients from regex matches was not deemed beneficial for modeling purposes.	The match “printing ink brilliant blue iron oxide” would be assigned “Printing Ink,” but multiple matches of [“printing ink”, “iron oxide”] would be assigned to “Printing Ink” and also “Iron Oxide.” The term “colour mixture” also appears in some SmPCs.
The same approach as for printing inks was adopted for flavoring agents. where The term “Juice” was used to refer to any flavored juice used in oral products. If the language of the SmPC specified a common flavour (e.g. “peppermint flavour”), this was retained.	Longer regex matches result in the term “Flavour,” but shorter matches results in “Flavour,” “vanillin,” “ethyl vanillin” and “vanillic acid” being separated.