It must be noted that the CTR recommends adding relevant Medical Subject Headings (MeSH) terms or terms from another controlled vocabulary, such as the Systematized Nomenclature of Medicine—Clinical Terms (SNOMED CT), that have been mapped to MeSH within the Unified Medical Language System (UMLS) metathesaurus for each of the conditions. While this recommendation adds a level of standardization to the condition names present in the CTR, the MeSH terms by themselves often fall short of describing the condition name they encode. Furthermore, for a given condition name there may be multiple associated MeSH terms, thereby leaving it to the user of the data to determine the most appropriate MeSH term for that condition name. Additionally, for certain records in the CTR, there might be no associated MeSH terms provided for a condition name. The table T1 below was created by performing a full join on the files “conditions.txt” and “browse_conditions.txt”, the files were joined using the national clinical trials ID (NCT ID). The table provides examples of CTR records with their associated condition names and mesh terms and illustrates the incompatibilities between condition names and MeSH terms.

Table T1: Conditions data with MeSH Terms:

NCT ID	Condition name	MeSH term
NCT05082610	triple negative breast cancer	neoplasms,triple negative breast neoplasms,carcinoma, non-small-cell lung,breast neoplasms,neoplasms by site,breast diseases,skin diseases,carcinoma, bronchogenic,bronchial neoplasms,lung neoplasms,respiratory tract neoplasms,thoracic neoplasms,lung diseases,respiratory tract diseases
NCT04254107	triple negative breast cancer	lymphoma,carcinoma,lymphoma, t-cell, peripheral,lymphoma, large b-cell, diffuse,triple negative breast neoplasms,squamous cell carcinoma of head and neck,stomach neoplasms,neoplasms by histologic type,neoplasms,lymphoproliferative disorders,lymphatic diseases,immunoproliferative disorders,immune system diseases,neoplasms, glandular and epithelial,neoplasms by site,carcinoma, squamous cell,lymphoma, b-cell,lymphoma, non-hodgkin,lymphoma, t-cell,breast neoplasms,breast diseases,skin diseases,head and

		neck neoplasms,gastrointestinal neoplasms,digestive system neoplasms,digestive system diseases,gastrointestinal diseases,stomach diseases
NCT01590680	neuroblastoma	neuroblastoma,pheochromocytoma,paraganglioma,neuroectodermal tumors, primitive, peripheral,neuroectodermal tumors, primitive,neoplasms, neuroepithelial,neuroectodermal tumors,neoplasms, germ cell and embryonal,neoplasms by histologic type,neoplasms,neoplasms, glandular and epithelial,neoplasms, nerve tissue,neuroendocrine tumors
NCT04081701	medulloblastoma	adenoma,meningioma,medulloblastoma,paraganglioma,pituitary neoplasms,esthesioneuroblastoma, olfactory,central nervous system neoplasms,hemangioblastoma,neoplasms, glandular and epithelial,neoplasms by histologic type,neoplasms,pituitary diseases,hypothalamic diseases,brain diseases,central nervous system diseases,nervous system diseases,endocrine system diseases,neoplasms, nerve tissue,neoplasms, vascular tissue,meningeal neoplasms,nervous system neoplasms,neoplasms by site,glioma,neoplasms, neuroepithelial,neuroectodermal tumors,neoplasms, germ cell and embryonal,neuroectodermal tumors, primitive,neuroendocrine tumors,endocrine gland neoplasms,hypothalamic neoplasms,supratentorial neoplasms,brain neoplasms,neuroblastoma,neuroectodermal tumors, primitive, peripheral,olfactory nerve diseases,cranial nerve diseases,hemangioma, capillary,hemangioma

NCT04294784	gastroesophageal cancer	NA
NCT02669914	gastroesophageal cancer	lung neoplasms,carcinoma, non-small-cell lung,colorectal neoplasms,pancreatic neoplasms,ovarian neoplasms,brain neoplasms,kidney neoplasms,carcinoma, renal cell,breast neoplasms,respiratory tract neoplasms,thoracic neoplasms,neoplasms by site,neoplasms,lung diseases,respiratory tract diseases,carcinoma, bronchogenic,bronchial neoplasms,intestinal neoplasms,gastrointestinal neoplasms,digestive system neoplasms,digestive system diseases,gastrointestinal diseases,colonic diseases,intestinal diseases,rectal diseases,endocrine gland neoplasms,pancreatic diseases,endocrine system diseases,ovarian diseases,adnexal diseases,genital diseases, female,female urogenital diseases,female urogenital diseases and pregnancy complications,urogenital diseases,genital neoplasms, female,urogenital neoplasms,genital diseases,gonadal disorders,central nervous system neoplasms,nervous system neoplasms,brain diseases,central nervous system diseases,nervous system diseases,urologic neoplasms,kidney diseases,urologic diseases,male urogenital diseases,adenocarcinoma,carcinoma,neoplasms, glandular and epithelial,neoplasms by histologic type,breast diseases,skin diseases

In the table T1 above, the condition “triple negative breast cancer” is associated with two clinical trial studies with the identifiers (NCT IDs) NCT05082610 and NCT04254107. Each of these studies lists the various MeSH terms associated with “triple negative breast cancer” with the most appropriate MeSH term being “triple negative breast neoplasms”. However, there are other associated MeSH terms for each of these studies which are not appropriate: for study NCT05082610 there are MeSH terms such as “carcinoma, non-small-cell lung” and “respiratory tract diseases” while study NCT04254107 has associated MeSH terms such as “lymphoma” and “stomach neoplasms”, which do not describe the condition of “triple negative breast cancer”. Furthermore, the MeSH terms are not identical between studies where the condition names are the same, which adds to the inconsistencies between records with same condition names. For instance, NCT04254107 contains various MeSH terms associated with lymphomas such as “lymphoma, b-cell”, “large b-cell”, “lymphoma, large b-cell, diffuse”, “lymphoma, t-cell, peripheral”, etc., but these terms are not contained in the list of MeSH terms for NCT05082610. We can also see in study NCT04294784, for the condition “gastroesophageal cancer”, there are no MeSH terms, but for the same condition with a different NCT ID (NCT02669914), there are multiple associated MeSH terms.

We conclude from our analysis that the MeSH terms–even though they are internally standardized– are not suitable to be used to identify the exact term and condition name for the conditions they encode. Therefore, even though we observed that the condition names from the CTR contained syntactic and semantic inconsistencies, we decided to use the terms from the conditions file to extract tumor names and map them to their standardized nomenclature in the WHO and NCIt databases.