3) CSO 分類別公的がん研究費の国際比較（図表3）

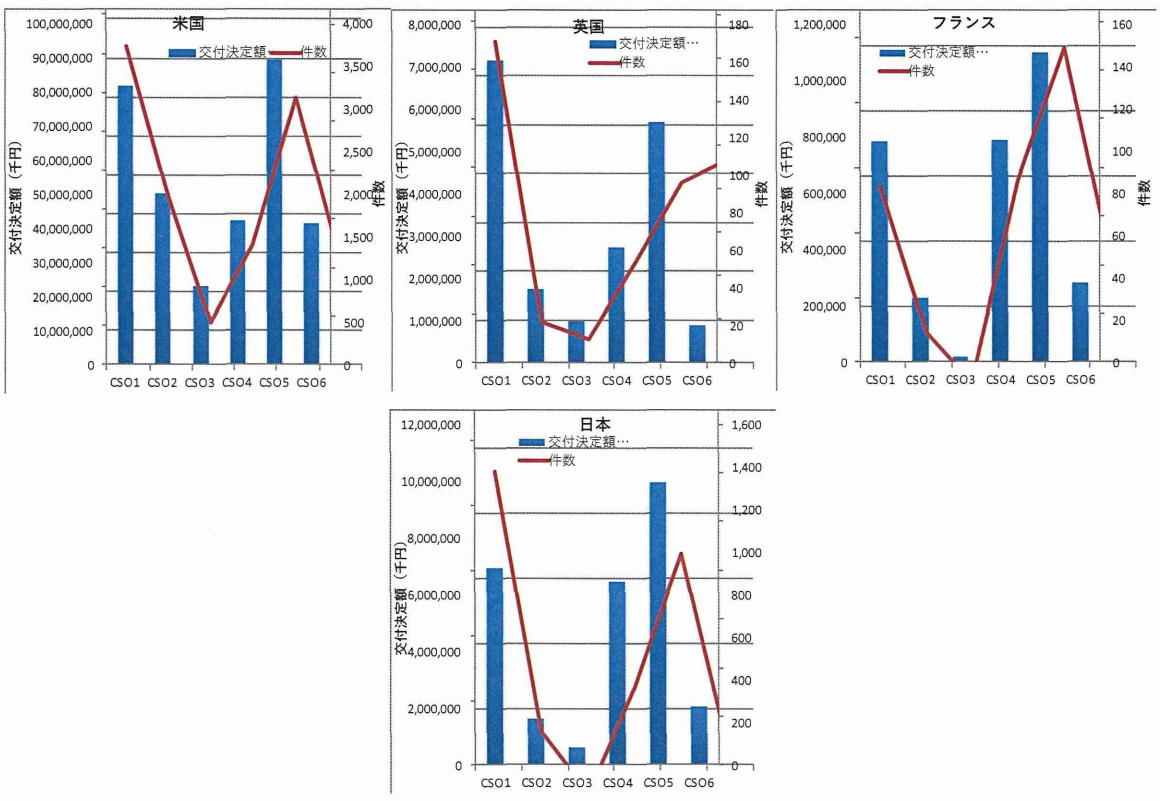
米国、英国、フランス及びわが国の公的がん研究費を CSO 分類別で分析した。米国とフランス、わが国では最も配分が多かったのが「CSO5 治療」であったが、英国では「CSO1 生物学」への配分が最も多いことが示唆された。また、米国では「CSO6 がんコントロール、サバイバーシップ、アウトカム研究」への配分がわが国を含む他の 3 か国よりも多い傾向が示唆された。わが国では最も配分の少ない「CSO3 予防」については、米国とフランスでも最も少なく、また英国でも「CSO6 がんコントロール、サバイバーシップ、アウトカム研究」についで少なく、わが国だけの傾向ではないことが示唆された。

4) 臓器分類別公的がん研究費の国際比較（図表4）

米国、英国、フランス及びわが国の臓器分類別がん研究費を、研究費総額順に上位 10 疾患を比較した。日本以外の米国、英国、フランスでは「部位が特定できない研究」と「乳がん」ががん研究費の総額として第 1 位、第 2 位であった。日本でも「部位が

図表 3　CSO 分類別公的がん研究費の国際比較

	米国		英国		フランス		日本	
	交付決定額（千円）	件数	交付決定額（千円）	件数	交付決定額（千円）	件数	交付決定額（千円）	件数
CSO1	81,760,134	3,725	7,153,361	171	782,483	84	6,932,328	1,379
CSO2	50,083,075	2,179	1,732,622	41	230,242	23	1,612,133	308
CSO3	23,025,652	885	978,114	33	18,451	2	602,157	57
CSO4	42,188,071	1,681	2,730,730	68	785,829	86	6,459,696	490
CSO5	88,979,242	3,208	5,665,036	105	1,096,821	140	9,925,140	1,041
CSO6	41,549,606	1,617	892,920	116	281,284	54	2,045,026	242
合計	327,585,779	13,293	19,152,783	536	3,195,111	388	27,576,480	3,516



特定できない研究」は第 1 位であったが、第 2 位は「肺がん」で、「乳がん」は第 5 位であった。

この 4 か国でがん研究費の上位 10 疾患に入っているのは、「部位が特定できない研究」「乳がん」「結腸／直腸がん、大腸がん」「白血病」の 4 疾病であった。一方で、「胃がん」「膵臓がん」「頭頚部がん」は上位 10 疾患のうち日本にのみ見られた。

(2)CSO 自動コーディング導入の検討

現在 ICRP では UberResearch 社が開発した CSO 自動コーディングの導入を検討しており、2015 年 4 月にカナダ・トロントで開催された ICRP 年次会議に UberResearch 社代表の Mr. Ashlea Higgs が招待され、同社の CSO 自動コーディングやデータベース構築への取り組みについて紹介された。

わが国の公的がん研究費データベース構築と維持・運用のためには、年間 3,000 件程度の公的がん研究を毎年抽出し、その全ての抄録に CSO 及び臓器コードを付加する必要がある。本研究班では、このコード付加は人力で行っている。すなわち、2 名が抽出した抄録やタイトル、キーワードなどから CSO 及び臓器コードを決定している。この作業には時間が膨大にかかるほか、本研究班での作業から、コーディングの質が、がん研究に関する知識とコーディング経験に左右され、その結果にはかなりばらつきがあることが見受けられた。この問題に対処するため、UberResearch 社の開発した自動コーディングの活用と、自動コーディングと人力によるコーディングの 2 本立てによる Double-blind のコーディングについて、検討を行った。

この自動コーディング導入の検討のため、2016 年 1 月 31 日から 2 月 3 日の日程で研究分担者の小川が渡米し、ワシントン DC にて UberResearch 社代表の Mr. Ashlea Higgs との面談を行った。また、面談に先立ち本研究班で構築した公的がん研究費データベースより、2011 年度の厚労科研費、文科科研費よりサンプルデータを抽出して Mr. Higgs に送付した。

Mr. Higgs との面談は、彼らが受託業務を行っている米国 NIH（National Institute of Health）本部・Bethesda キャンパス 2 号館の会議室にて実施した。面談内容は、1) UberResearch 社の自動コーディングシステムの概要説明、2) サンプルデータのコーディング結果の検討、3) わが国への自動コーディングシステムの導入可能性に関する討議、についてであった。

1) UberResearch 社の自動コーディングシステムの概要説明

本研究班の実施している CSO 及び臓器コードの付加方法と分析について概説を行ったのち、UberResearch 社の自動コーディング技術について解説を受けた。UberResearch 社の開発した自動コーディング技術は、既存のソフトウエアを活用するのではなく、アルゴリズムをゼロから開発する手法を取っており、そのための専門家を多く有しているとのことであった。またこのアルゴリ

図表 4　臓器分類別公的がん研究費の国際比較（交付額の多い 10 疾病）

米国	交付決定額（千円）	件数	英国	交付決定額（千円）	件数	フランス	交付決定額（千円）	件数	日本	交付決定額（千円）	件数
部位が特定できない研究	76,020,858	3,038	部位が特定できない研究	12,559,251	279	乳がん	524,268	62	部位が特定できない研究	11,889,651	727
乳がん	52,364,166	2,110	乳がん	830,447	34	部位が特定できない研究	417,375	53	肺がん	1,760,282	263
肺がん	25,311,961	967	結腸/直腸がん、大腸がん	811,029	26	白血病	301,262	36	肝臓がん	1,410,322	190
前立腺がん	24,691,923	961	白血病	691,825	23	結腸/直腸がん、大腸がん	289,091	28	結腸/直腸がん、大腸がん	1,386,781	201
結腸/直腸がん、大腸がん	23,588,980	925	卵巣がん	651,918	10	非ホジキンリンパ腫	288,749	24	乳がん	1,368,047	190
白血病	21,901,333	803	前立腺がん	484,777	25	肝臓がん	260,280	24	白血病	1,081,190	196
脳腫瘍	17,136,779	638	子宮内膜がん	313,655	2	肺がん	158,548	25	胃がん	966,010	157
非ホジキンリンパ腫	11,004,907	439	食道がん	295,037	7	悪性黒色腫	137,551	12	膵臓がん	944,329	148
悪性黒色腫	9,554,491	382	悪性黒色腫	232,876	4	脳腫瘍	136,486	24	頭頚部がん	653,759	89
卵巣がん	9,412,831	334	子宮頸がん	231,511	10	軟部組織肉腫	108,636	12	卵巣がん	611,889	85

ズムは目的に応じてカスタマイズが可能であり、様々な目的で利用可能で、実際に多くの企業や団体での利用実績があるとのことであった。なお、具体的な顧客名は開示されなかったものの、米国 NIH や WHO など国際機関、また中国の地方政府などが含まれていると思われ、さらにわが国の一般企業とも何らかの契約を締結した経験を有しているとの話であった。

現在、UberResearch 社では ICRP との契約に向けて、がん研究費分析のための CSO 及び臓器コードの自動コーディングのアルゴリズムをすでに開発・カスタマイズ済みであり、本研究で自動コーディングを依頼する場合は、このプログラムを走らせることで容易に可能であるとの説明であった。

UberResearch 社が自動コーディングなどデータ処理を受託する場合は、依頼元の事業規模や自動コーディングなどのカスタマイズの内容により、その費用は大きく異なるが、カスタマイズが少ないあるいはほとんどない場合は、ごく割安な費用で実施可能であり、CSO 自動コーディングについては費用面からはかなり割安であるとの説明を受けた。

自動コーディングにかかる作業時間は、アルゴリズム開発及び運用の専門家の作業時間次第ではあるが、3,000～4,000 件のがん研究の抄録を用いた自動コーディングであれば、機械翻訳による英訳などの準備期間も含めて 10 日前後で実施可能との話であった。

図表５ サンプルデータを用いた CSO コーディングの自動と人力の比較

CSO分類	適合	非適合	適合割合
厚労三次がん	118	44	72.8%
がん研究開発費	63	45	58.3%
文科科研費	37	35	51.4%
合計	218	124	63.7%

図表６ サンプルデータを用いた臓器コーディングの自動と人力の比較

臓器分類	適合	非適合	適合割合
厚労三次がん	116	46	71.6%
がん研究開発費	87	21	80.6%
文科科研費	57	21	73.1%
合計	260	88	74.7%

2) サンプルデータの自動コーディング結果の検討

本研究班で抽出し、分析を実施中の厚労科研費及び文科科研費の一部をサンプルデータとして抽出し、自動コーディングを試行した。

試行に用いたデータは 342 件あり、内訳としては厚労科研費の三次がん関連が 162 件、がん研究開発費が 108 件、文科科研費が 72 件であった。なお、抽出した文科科研費は研究者によって英文抄録が作成されていた研究を抽出したため、そのまま自動コーディングを実施した。抽出した厚労科研費は日本語抄録のみであったので、google 翻訳を用いて機械的に日本語から英語に翻訳し、その結果を自動コーディングした。なお、試行に用いた研究は全て人力でのコーディングを実施しており、自動コーディングと人力コーディングの結果を比較することで、自動コーディングの精度を検証した。

自動コーディング結果と人力コーディングの比較を図表５、図表６に示す。CSO の自動コーディングは全体で約 64%が人力コーディングと合致しており、臓器コーディングの約 75%と比較して低い傾向が示唆された。CSO 自動コーディングは、特に文科科研費で低い傾向が見られた。この理由としては、研究者本人による英文抄録が研究内容を的確に要約していなかったものがあることが考えられる。実際に日本語抄録と英語抄録を見比べると、その内容に大幅な差が見られた研究もあり、自動コーディングの際に研究者本人による英語抄録の活用は慎重にすべきと思われる。

また、google 翻訳を用いた自動コーディングの精度が比較的良い結果となったことも特筆すべきであろう。これは、自動コーディングが抄録に記載されたキーワードの組み合わせにより判断していると思われることから、google 翻訳ががん医療の専門用語を正しく翻訳している可能性が高いと思われる。なお、google 翻訳による日本語から英語への翻訳は、文章としては成立していないものも多くみられ、翻訳ソフトとしてはまだ発展途上であることがうかがえた。しかしながら、本研究における自動コーディングに用いる程度の精度は充分に得られたと考えられ、日本語抄録から google 翻訳と UberResearch 社の自動コーディングにより、簡便に CSO コードの付加が可能となることが示唆された。

臓器コーディングについては、人力では抄録、タイトル、キーワードに記載されているがんの部位を全て記載しており、同様の方法を自動コーディングで行うことを想定していた。そのため、今回の試行の精度はほぼ 100%に近いことを予想していたが、実際には 70%台であった。この点について UberResearch 社に確認したところ、臓器コーディングについても CSO コーディングと同様に、抄録中のキーワードの組み合わせからアルゴリズムを用いて判断させているとのことであった。例えば、「タバコに関する研究」とだけ記載されている場合は、人力コーディングでは「部位が特定できない研究」として判断されるが、自動コーディングでは「肺がん」などと判断されるとのことで、この判断が本研究に適しているかどうか、さらなる検討が必要と考えられた。

3) わが国への自動コーディングシステムの導入可能性に関する討議

UberResearch 社の自動コーディングのわが国への導入は、公的がん研究費データベースを存続させるためには最低限の作業である CSO 及び臓器コードの付加の作業量を大幅に軽減することができ、そのメリットが大きいと考えられる。ただし、現時点では自動コーディングで付加された CSO 及び臓器コードをそのまま用いるのではなく、あくまで double-blind のコード付加作業の一つとして位置付け、専門家による人力によるコード付加との整合性をとる方法が適切であろう。このように自動コーディングを活用することで、CSO 及び臓器コードの付加作業にかかる時間が半分以下となり、今後の公的がん研究費データベースの持続的な運用に大いに寄与すると考えられる。

UberResearch 社の自動コーディングの利用にあたり、日本語抄録の問題が想定されたが、google 翻訳を用いた本研究の試行により、google 翻訳などの機械翻訳を活用することでかなりの精度で自動コーディングの実施が可能であることが示唆された。今後は google 翻訳の効率的な利用などで翻訳の精度を向上させる手法について検討が必要と思われる。

UberResearch 社の自動コーディングにかかる費用については、年度ごとに抽出したがん研究の件数が 4,000 件前後と想定すると、年間 50〜100 万円程度で受託可能との回答を口頭で得た。この金額を含め、公的がん研究費データベースの運用にかかるコストについて、今後検討する必要があろう。

(3)わが国における公的がん研究費データベースの活用方法の検討

公的がん研究費データベースの持続的な運営と活用について、昨年度研究で厚生労働省と文部科学省との協議を実施した。

厚生労働省とは、2015 年 2 月 27 日にがん対策・健康増進課との協議を行い、公的がん研究費データベースに収載する研究項目の追加についてアドバイスをいただいたほか、希少がんに関する情報収集と分析のリクエストがあった。

文部科学省とは、2015 年 3 月 12 日に研究振興局との協議を行い、アウトカム情報

など今後の研究費配分の政策立案に有用な情報について協議を行った。また、両省から国立研究開発法人日本医療研究開発機構（AMED）との協議を勧められた。

本年度研究では、昨年度研究の成果を踏まえ、2015年9月8日にAMEDの戦略推進部がん研究課との協議を行った。協議内容は、本研究班の紹介と2015年度ICRP総会の報告、さらに公的がん研究費データベースの活用を含めた今後の協力体制についてであった。本研究班で構築した公的がん研究費データベースは、AMEDにとって重要な情報であり、今後アウトカム情報の活用の検討などを含めて、公的がん研究費データベースの持続的な運用について、引き続き協議をすることで同意した。

また、UberResearch社代表のMr. Higgsとの面談の際に、NIHのOffice of Portfolio AnalysisのDirectorであるDr. George Santangeloと面談する機会をいただき、本研究班で実施したわが国のがん研究費の分析や、本研究で実施した国際比較の概要などを説明した。NIHとしても研究費の適正な配分に向けた各種のデータ整備や分析に取り組んでおり、今後国際的な協力体制も強化したいとのことで、今後本研究班のみならず、AMEDなどわが国の医学研究にかかる機関との協力体制を構築したいとのことであった。また、本研究班で実施している研究費の科学的な分析にはNIHとしても今後とも取り組みたいとのことで、分析手法の検討などについても協力したいとの申し入れがあった。

D. 考察

本研究は、公的がん研究費データベースの活用と持続性について、同データベースを活用した国際比較分析の試行や、データベースの効率的な運用と将来のあり方について分析を実施した。

本研究により構築した公的がん研究費データベースにより、また諸外国で活用が進んでいるCSO分類を用いたことで、諸外国のがん研究費との比較が可能になることが明らかになった。また、その成果を用いてわが国ならではのがん研究費の配分のあり方を検討し、その結果を踏まえた政策立案が可能になると思われる。

公的がん研究費データベースを継続的に活用するためには、データ整備やデータベース構築をできる限り省力化し、効率の良い運用が求められる。そのためには、現在最も人手と時間がかかっている各研究へのCSO及び臓器コードの付加作業の軽減は必須である。その対策の一つとして、UberResearch社の提供する自動コーディングの活用は極めて有効と思われる。また、公的がん研究費データベースを、研究班の成果物のみならず、わが国のがん政策立案に幅広く活用するためには、持続的な公的がん研究費データベースの運用体制の確保が必須と考えられる。本年度はAMEDとの協議を通じ、公的がん研究費データベースの持続的な運用について検討することができた。来年度も引き続き公的がん研究費データベースの活用について、検討を重ねたいと考えている。

E. 結論

本研究により、公的がん研究費データベースの活用について、データベース運用体制やデータ整備、また分析を通じた活用の検討を実施した。今後ともICRPやAMED、NIHなどとの連携を密にすることで、より効果的なCSO分類の活用とがん研究費の配分が可能になると考えられる。

F. 健康危険情報

なし

G. 研究発表

H. 知的財産権の出願・登録状況

なし

1．論文発表

2．学会発表

III. 研究成果の刊行に関する一覧表

雑誌

発表者氏名	論文タイトル名	発表誌名	巻号	ページ	出版年

IV. 研究成果の刊行物・別刷

資　　料

CSO コード一覧

Biology

1.1 Normal Functioning
1.2 Cancer Initiation: Alterations in Chromosomes
1.3 Cancer Initiation: Oncogenes and Tumor Suppressor Genes
1.4 Cancer Progression and Metastasis
1.5 Resources and Infrastructure related to biology

Etiology

2.1 Exogenous Factors in the Origin and Cause of Cancer
2.2 Endogenous Factors in the Origin and Cause of Cancer
2.3 Interactions of Genes and/or Genetic Polymorphisms with Exogenous and/or Endogenous Factors
2.4 Resources and Infrastructure Related to Etiology

Prevention

3.1 Interventions to Prevent Cancer: Personal Behaviors that Affect Cancer Risk
3.2 Nutritional Science in Cancer Prevention
3.3 Chemoprevention
3.4 Vaccines
3.5 Complementary and Alternative Prevention Approaches
3.6 Resources and Infrastructure Related to Prevention

Early Detection, Diagnosis and Prognosis

4.1 Technology Development and/or Marker Discovery
4.2 Technology and/or Marker Evaluation with Respect to Fundamental Parameters of Method
4.3 Technology and/or Marker Testing in a Clinical Setting
4.4 Resources and Infrastructure Related to Early Detection, Diagnosis or Prognosis

Treatment

5.1 Localized Therapies - Discovery and Development
5.2 Localized Therapies - Clinical Applications
5.3 Systemic Therapies - Discovery and Development
5.4 Systemic Therapies - Clinical Applications
5.5 Combinations of Localized and Systemic Therapies
5.6 Complementary and Alternative Treatment Approaches
5.7 Resources and Infrastructure Related to Treatment

Cancer Control, Survivorship and Outcomes Research

6.1 Patient Care and Survivorship Issues
6.2 Surveillance
6.3 Behavior
6.4 Cost Analyses and Health Care Delivery
6.5 Education and Communication
6.6 End-of-Life Care
6.7 Ethics and Confidentiality in Cancer Research
6.8 Complementary and Alternative Approaches for Supportive Care of Patients and Survivors
6.9 Resources and Infrastructure Related to Cancer Control, Survivorship, and Outcomes Research

臓器コード

ICRP Site	ICRP Code	ICRP Description	*Equivalent ICD-10 Code (for information only DO NOT USE ON ICRP TEMPLATE)*
Adrenocortical Cancer	0		*C74.0*
Anal Cancer	103		*C21*
Bladder Cancer	3		*C67*
Bone Cancer	4	Includes Osteosarcoma, Malignant Fibrous Histiocytoma, Ewing's sarcoma and all other bone/cartilaginous tumors.	*C40, C41*
Brain Tumor	6	Includes Chordoma	*C71*
Breast Cancer	7		*C50*
Cardiotoxicity / Heart Cancer	8		*C38.0*
Cervical Cancer	9		*C53*
Colon and Rectal Cancer	64		*C18, C19, C20*
Ear Cancer	10		*C30.1*
Endometrial Cancer	11		*C54*
Esophageal / Oesophageal Ca	12		*C15*
Eye Cancer	13	Not including Retinoblastoma (45)	*C69 (excluding C69.2)*
Gallbladder Cancer	14		*C23*
Hodgkin's Disease	24		*C81*
Kaposi's Sarcoma	46		*C46*
Kidney Cancer	25	Includes Kidney cancer and Wilms' tumor (60)	*C64*
Laryngeal Cancer	26		*C32*
Leukemia / Leukaemia	27	Including ALL, AML, CLL, CML & Hairy Cell Leukaemia, Myelodysplastic Syndrome and Myeloproliferative disorders	*C91, C92, C93, C94, C95*
Liver Cancer	23	Including Bile Duct	*C22*
Lung Cancer	28	Including Mesothelioma	*C34, C45*
Melanoma	29		*C43*
Myeloma	30	Including Multiple Myeloma	*C90*
Nasal Cavity and Paranasal Si	31		*C30.0, C31*
Neuroblastoma	32		*C74.9*
Non-Hodgkin's Lymphoma	35		*C82, C83, C84, C85, C96.3*
Oral Cavity and Lip Cancer	36		*C00, C01, C02, C03, C04, C05, C06, C09*
Ovarian Cancer	66		*C56*
Pancreatic Cancer	37		*C25*
Parathyroid Cancer	38		*C75.0*
Penile Cancer	39		*C60*
Pharyngeal Cancer	61		*C14.0*
Pituitary Tumor	40		*C75.1*
Primary CNS Lymphoma	104		--
Primary of Unknown Origin	102		--
Prostate Cancer	42		*C61*
Retinoblastoma	45		*C69.2*
Salivary Gland Cancer	63		*C07, C08*
Sarcoma (soft tissue)	105	Includes Fibrosarcoma, Rhabdomyosarcoma, leiomyosarcoma, liposarcoma, muscle and other Soft Tissue Sarcoma (but not Ewing's Sarcoma or other bone/cartilaginous tumors (4), or Kaposi's Sarcoma (46))	*C49*
Skin Cancer (non-melanoma)	49		*C44*
Small Intestine Cancer	50		*C17*
Stomach Cancer	51		*C16*
Testicular Cancer	52		*C62*
Thymoma, Malignant	53		*C37*
Thyroid Cancer	54		*C73*
Vaginal Cancer	57		*C52*
Vulva	101		*C51*

Cancer types, not otherwise specified **ICRP Site**	**ICRP Code**	**ICRP Description**	*Equivalent ICD-10 Code (for information only DO NOT USE ON ICRP TEMPLATE)*
Blood Cancer	67	Use this code for Blood Cancers other than: Hodgkin's Disease (24), Leukemia / Leukaemia (27), Myeloma (30), Non-Hodgkin's Lymphoma (35)	*C88, C96 (excluding C96.2, C96.3)*
Gastrointestinal Tract	15	Use this code for GI cancers other than: Colon and Rectal (64), Esophogeal /Oesophageal (12) , Gallbladder (14), Liver (23), Pancreatic (37), Small Intestine (50), Stomach (51). The computer program will automatically map these sites to GI cancers.	*C26.9*
Genital System, Female	17	Use this code for genital system, female cancers other than: Cervical (9), Endometrial (11), Ovarian (66), Vaginal (57), Vulva (101). The computer program will automatically map these sites to this category.	*C57*
Genital System, Male	19	Use this code for genital system, male cancers other than: Penile (39), Prostate (42), Testicular (52) cancers. The computer program will automatically map these cancer sites to this category.	*C63*
Head and Neck Cancer	21	Use this code for head and neck cancers other than: Laryngeal (26), Nasal Cavity and Paranasal Sinus (31), Oral Cavity and Lip (36), Parathyroid (38), Pharyngeal (61), Salivary Gland (63), and Thyroid (54) cancers. The computer program will automatically map these cancer sites to this category.	*C76.0*
Nervous System	33	Use this for nervous system cancers other than: Brain (6), Eye (16), Neuroblastoma (32), Pituitary (40), Primary CNS Lymphoma (104) and Retinoblastoma (45). The computer program will automatically map these cancers to this category.	--
Not Site-Specific Cancer	2	Includes fundamental research (fluids, secretions, milk lymph, blood components, cell lines and cell fractions, etc.) and research that applies to all types of cancer.	-
Respiratory System	43	Use this code for respiratory cancers other than: Lung (28), Nasal Cavity & Paranasal Sinus (31) cancers. The computer program will automatically map these cancers to this category.	*C39*
Urinary System	55	Use this code for urinary cancers other than: Bladder (3), Kidney or Wilms' tumor (25). The computer program will automatically map these cancer sites to this category.	*C65, C66, C68*