**U.S. Patent and Trademark Office**

**OFFICE OF CHIEF ECONOMIST**

**Data Science Big Data**

USPTO Cancer Moonshot Patent Data

Jesse Frumkin, *Patent Examiner*

Amanda F. Myers, *Economist*

August 2016

For more information about the USPTO’s Office of Chief Economist, visit www.uspto.gov/economics.



Table of Contents

[I. Introduction 3](#_Toc459023875)

[A. Background 3](#_Toc459023876)

[B. Objective 3](#_Toc459023877)

[II. Method 4](#_Toc459023878)

[A. Classification-based queries 4](#_Toc459023879)

[1. Official Patent Classification 4](#_Toc459023880)

[2. Derwent World Patents Index 6](#_Toc459023881)

[a. Derwent Chemical Patents Index and Electronic/Electrical Patents Index 6](#_Toc459023882)

[b. Derwent Chemical Index Classification 7](#_Toc459023883)

[c. Japanese F-Terms and F-Index 7](#_Toc459023884)

[B. Keyword-based queries 8](#_Toc459023885)

[1. Define search terms 8](#_Toc459023886)

[2. Query US document title/abstract 9](#_Toc459023887)

[3. Query Derwent title and abstract 9](#_Toc459023888)

[III. Search Method Assessment 9](#_Toc459023889)

[A. Science and Technology Information Center 9](#_Toc459023890)

[B. Chemical Abstract Plus query 10](#_Toc459023891)

[C. Preliminary Visualizations 10](#_Toc459023892)

[IV. Results 10](#_Toc459023893)

[V. Patent Data 13](#_Toc459023894)

[A. CPC-based technology categories 13](#_Toc459023895)

[B. DNA, RNA, or Protein Sequence Indicator 13](#_Toc459023896)

[VI. External Data 13](#_Toc459023897)

[A. NIH RePORTER 14](#_Toc459023898)

[B. FDA Orange Book 14](#_Toc459023899)

[VII. Conclusion 15](#_Toc459023900)

[VIII. Appendix I: Data Set Dictionary 16](#_Toc459023901)

[IX. Appendix II: Cancer Classification Codes 18](#_Toc459023902)

[X. Appendix III: Cancer EPI & CPI Codes 20](#_Toc459023903)

[XI. Appendix IV: Cancer Japanese F-Terms and F-Index Codes 21](#_Toc459023904)

[XII. Appendix V: Cancer Keyword Stems 23](#_Toc459023905)

Introduction

## Background

In his 2016 State of the Union Address, President Obama called on Vice President Biden to lead a new national effort to eliminate cancer. The resulting National Cancer Moonshot works to accelerate research efforts by enhancing data access and facilitating collaborations among researchers, doctors, philanthropists, patients, and innovators.[[1]](#footnote-1) The Cancer Moonshot Task Force, comprised of individuals from several Federal agencies including the Department of Commerce and the United States Patent and Trademark Office (USPTO), has since announced various initiatives to drive investment, improve policy, and build private sector partnerships. The Task Force is focused on catalyzing scientific breakthroughs, unleashing the power of data, accelerating access to new therapies, strengthening prevention and diagnosis, and enhancing patient access and care.[[2]](#footnote-2)

One USPTO initiative, the “Horizon Scanning Tool”, focuses on leveraging patent data to reveal new insights regarding Federal funding, research and development (R&D), and commercialization of cancer-related innovations. This tool houses patent data sets that are intended to empower users to identify promising R&D on the horizon in diagnostics, therapeutics, data analytics, and model biological systems. Users are able to build rich visualizations of patent data, often an early indicator of successful R&D, and combine them with other related indicators.

## Objective

The purpose of this report is to document the methods used to generate the USPTO Cancer Moonshot Patent Data (hereafter “data set”) for the Horizon Scanning Tool. We generate the data set using USPTO examiner tools to execute a series of queries designed to identify cancer-specific patents and patent applications. We apply several query approaches to ensure coverage of the various fields and subject matter that cancer-related innovations encompass. These include drugs, diagnostics, cell lines, mouse models, radiation-based devices, surgical devices, image analytics, data analytics, and genomic-based inventions. In designing our search method, we emphasize precision over recall. That is, we aim to minimize false positives at the potential cost of excluding some documents with potential applicability to cancer. We do this to ensure that the data set contains the most relevant cancer inventions.

The final data set consist of 269,353 patent documents (published patent applications and granted patents) spanning the 1976 to 2016 period. For this final set, we compile key patent data fields, including patent number, title, classification, and grant date. We also construct a set of fields based on the patent classification to indicate whether the invention falls within certain high-level technology categories. We then combine the data set with two external data sources – the National Institutes of Health (NIH) Research Portfolio Online Reporting Tools (RePORTER) and the U.S. Food and Drug Administration (FDA) *Approved Drug Products with Therapeutic Equivalence Evaluations,* commonly known as the Orange Book, – to facilitate users seeking to link data on cancer-related patents to data on upstream funding and R&D as well as downstream commercialization efforts. The data set is available for download at: [insert link]. Appendix I contains a data dictionary with names and definitions for all fields in the data set.

The remainder of this report is structured as follows. Section II details the query methods, highlighting the merits and limitations of each. Section III describes the internal validation steps. Section VI summarizes the search results. Section V details the data set fields constructed to facilitate use of the tool. Section VI discusses the data extracted from NIH RePORTER and FDA Orange Book, as described above. Section VII concludes.

Method

## Classification-based queries

The first set of queries relies on different patent technology classification schemes. Generally, classifications are based on the technological and functional features of inventions and encompass the entire spectrum of subject matter that can be claimed. Within any patent classification scheme, there is no single class that encompasses all cancer-related innovations. However, for some types of invention, there are cancer-specific subclasses, subgroups, or subcategories that we exploit to identify cancer-related patent documents.

### Official Patent Classification

For most recent patent documents, USPTO assigns three official patent classifications: i) the Cooperative Patent Classification (CPC)[[3]](#footnote-3); ii) U.S. Patent Classification (USPC)[[4]](#footnote-4); and iii) International Patent Classification (IPC)[[5]](#footnote-5). Each scheme includes subgroup definitions or subclass titles that explicitly reference cancer-specific innovations. For example, the description for CPC subgroup C12Q 1/6886 reads “Cancer (DNA/RNA hybridization probes)”. However, cancer-specific codes are too restrictive to capture all documents with cancer relevant subject matter. Thus, we apply the following iterative approach to expand the set of cancer-specific codes:

*Step 1*: Specify a list of codes with definition statements or titles that expressly reference a cancer-specific invention.

*Step 2*: Search the Examiner's Automated Search Tool (EAST) for patent documents with codes identified in Step 1.

*Step 3*: Based on the search results from Step 2, calculate document frequencies for all codes to identify codes not initially identified in Step 1, but that are specific to an aspect of cancer.

*Step 4*: For frequently occurring codes not initially identified, review a small set of documents and the CPC definitions to verify that codes capture innovations relevant to cancer with high precision.

*Step 5*: Add codes verified as cancer-related in Step 4 to the list of cancer codes.

We perform multiple iterations of the above process for each of the three official classification schemes. This results in a final list of 30 USPC, 69 CPC, and 7 IPC cancer-related codes. Note that this list includes varying levels of granularity (e.g. group and subgroup). Searching the EAST system on the final list of cancer classification codes yields 81,555 U.S. documents. Table 1 presents the most frequently occurring cancer classification codes for each scheme in our results. See Appendix II for the complete lists.

Table 1: U.S. Patent Document Frequency by Cancer CPC, USPC, and IPC Codes

|  |  |  |
| --- | --- | --- |
| IPC Classification | Description | Document Frequency |
| A61P 35/00 | Antineoplastic agents (medicinal) | 15,784 |
| G01N 33/574 | Cancer (immunoassays) | 2,774 |
| A61P 35/02 | Leukemia (antineoplastic medicines) | 1,916 |
| A61P 35/04 | Metastasis (antineoplastic medicines) | 1,491 |
| C07K 16/30 | Tumor cell (monoclonal antibodies) | 939 |

|  |  |  |
| --- | --- | --- |
| CPC Classification | Description | Document Frequency |
| C12Q 1/6886 | Cancer (DNA/RNA hybridization probes) | 6,108 |
| A61K 39/0011 | Cancer antigens (medicinal) | 2,095 |
| C07K 16/30 | Tumor cell (monoclonal antibodies) | 1,975 |
| G01N 33/574 | Cancer (immunoassays) | 1,625 |
| G01N 33/5011 | Testing antineoplastic activity | 1,501 |

|  |  |  |
| --- | --- | --- |
| USPC Classification (Retired) | Description | Document Frequency |
| 435/6.14 | Detecting cancer (chemically) | 6,673 |
| 435/7.23 | Tumor or cancer cell (testing) | 4,914 |
| 514/19.3 | Cancer (drugs) | 4,456 |
| 424/155.1 | Cancer (monoclonal antibodies) | 1,265 |
| 530/388.8 | Cancer-cell binding (antibody) | 807 |

### Derwent World Patents Index

Users of the EAST system can access searchable indexes supported by the Thomson Reuters® Derwent Worlds Patent Index (DWPI). The DWPI searchable indexes are generally derived from Derwent enhanced patent documents, which contain annotated titles and abstracts written by Thomson Reuters analysts that summarize each unique invention of a multi-jurisdiction Derwent patent family.[[6]](#footnote-6) The abstract annotates the invention's advantages, uses, figure captions, and potential novelty. We specify a set of queries utilizing three of the DWPI searchable indexes. Query results consist of Derwent patent families, which we limit to only those containing U.S. documents.

### a. Chemical Patents Index and Electronic/Electrical Patents Index

DWPI contains proprietary patent classification codes, called Thomson Reuters Manual Codes, in three broad technology areas, which are divided into 21 sections and further subdivided into classes. We build queries based on two of the broad categories of Manual Codes – the Chemical Patents Index (CPI)[[7]](#footnote-7), which encompasses the chemical and pharmaceutical sections, and the Electronic/Electrical Patents Index (EPI)[[8]](#footnote-8), which covers the electronic and electrical sections.[[9]](#footnote-9) We include the latter to ensure coverage of patented devices, often used in cancer diagnostics and treatment. As with the official patent classification schemes, a subset of CPI and EPI classes expressly reference cancer-specific inventions. For example, CPI code B14-H01 covers “Anticancer (pharmaceuticals)”. We keyword search the DWPI within EAST to identify the set of 166 cancer-specific CPI and EPI codes.

Searching DWPI via EAST, we identify 105,565 Derwent patent families with a cancer-specific CPI or EPI code. Table 2 presents the most frequently occurring cancer CPI and EPI codes in our results. See Appendix III for the complete list of identified cancer-related CPI and EPI Codes.

Table 2: Derwent Family Frequency by Cancer CPI and EPI Codes

|  |  |  |
| --- | --- | --- |
| CPI Classification | Description | Document Families |
| B14-H01 | Anticancer (pharmaceutical) | 72,377 |
| B12-K04A1 | Diagnosis of tumors, cancer | 16,749 |
| B14-H01B | Antiproliferative, inhibitor of cell division | 13,909 |
| C14-H01 | Leukaemia treatment (agricultural) | 5,091 |
| B04-F02A | Cancer cells/carcinoma | 4,289 |

|  |  |  |
| --- | --- | --- |
| EPI Classification | Description | Document Families |
| S05-G02G4 | Treatment planning systems (radiation) | 739 |

### b. Chemical Index Classification

Chemical Index Classification in DWPI is designed to designate if certain chemical structures are present among the compounds claimed or disclosed in a chemical patent. We keyword search DWPI documentation to identify the four classification codes capturing cancer-related chemical structures (see Table 3). We then search DWPI against these cancer Chemical codes, resulting in 104,172 Derwent patent families.

Table 3: Derwent Family Frequency by Cancer Chemical Index Classification

|  |  |  |
| --- | --- | --- |
| Chemical Index Classification | Description | Document Families |
| P631 | Inhibitor of cell division (in cancer) | 27,259 |
| P632 | Leukaemia treatment | 17,200 |
| P633 | Other antitumour agent | 97,264 |
| P634 | Tumour inducing agent | 2,306 |

### c. Japanese F-Terms and F-Index

Japanese F-Terms and F-Index Codes are extensions of IPC classification used for indexing Japanese patent documents. Since many U.S. patent documents are members of international patent families, we query DWPI against Japanese F-Terms and F-Index terms in order to locate U.S. patent family members associated with the Japanese documents. We keyword search English translations of the definitions of Japanese F-Terms and F-Index codes to identify a set of 20 cancer-related F-Terms and F-Index codes. We then query DWPI for these codes, resulting in 21,247 Derwent patent families. Table 4 presents the most frequently occurring cancer F-Terms and F-Index Codes in our results. See Appendix IV for the complete list of F-Terms and F-Index Codes identified as cancer-related.

Table 4: Derwent Family Frequency by Cancer-related Japanese F-Terms and F-Index Codes

|  |  |  |
| --- | --- | --- |
| F-Term | Description | Document Families |
| 4C086 ZB26 | Medicine for tumours | 17,437 |
| 4C085 LL18 | Cancer Diagnostics | 1,108 |
| 2G045 AA26 | Carcinostatic substances | 1,089 |
| 4C085 AA21 | Chemotherapy | 648 |
| 4C088 ZB26 | Medicines to combat tumours (plant) | 482 |

|  |  |  |
| --- | --- | --- |
| F-Index | Description | Document Families |
| A61K39/395 T | Cancer (medicinal antibodies) | 1,504 |
| G01N33/574 | Immunoassays (for cancer) | 1,449 |
| G01N33/574 A | Cancer-related proteins (immunoassays) | 1,046 |
| G01N33/574 D | Cancer-related saccharides (immunoassays) | 338 |
| C07K16/30 | Tumour cell (monoclonal antibodies) | 363 |

## Keyword-based queries

For our second set of queries, we specify a set of key terms associated with cancer for searching.

### Define search terms

We define our keyword set based on two external sources: the U.S. National Library of Medicine’s Medical Subject Headings (MeSH®) and DWPI Title Terms.

#### MeSH terms

MeSH is an annually updated, controlled vocabulary (thesaurus) that provides uniformity and consistency to the indexing and cataloguing of biomedical literature.[[10]](#footnote-10) MeSH headings are organized in a hierarchical tree structure consisting of 16 main branches, which are divided into multiple levels of sub-branches. The tree structure is designed to allow for a search of a broader term (e.g., neoplasms) to automatically expand to include narrower terms (e.g., melanoma) in all branches. We identified cancer-specific MeSH keywords by browsing the MeSH hierarchy for terms narrower than "neoplasms", excluding terms known to represent benign and precancerous conditions.[[11]](#footnote-11) We then test the identified keywords against DWPI to determine whether the terms would retrieve Derwent patent documents about unrelated (non-cancer) conditions or benign conditions. We replace keywords having multiple meanings with more precise phrases to improve the query specificity. We do not employ an iterative approach to selecting MeSH keywords because we specifically validate these terms (see Section III.B).

#### DWPI Title Terms

Derwent enhanced patent documents include an annotated title. These titles contain preferred forms of words and are generally more standardized than the titles listed on actual patent documents which can contain numerous terms to indicate the same technology or subject matter. To identify cancer relevant terms in the Derwent titles, we perform preliminary searches of DWPI for cancer-specific documents. We then extract titles from the resulting Derwent documents and parse each word to construct a set of individual terms. We eliminate duplicates and stop terms and then manually select individual terms that are relevant to cancer.

We combine cancer-specific terms from MeSH and Derwent titles to create our final set of 127 cancer keyword stems. We design our query to capture various expansions of the keyword stems. For example the query $cancer$ (where "$" represents left and right truncation of the base term) expands to "cancer", "cancerous", "anticancer", "anti-cancer", etc. Table 5 presents the most frequently occurring keywords in our results. See Appendix V for the complete list of cancer keyword stems.

Table 5: U.S. Patent Document & Derwent Family Frequency by Cancer Keywords

|  |  |  |
| --- | --- | --- |
| Keyword (QUERY) | Patent Documents | Derwent Patent Families |
| cancer | 56,982 | 94,021 |
| tumor | 23,781 | 47,430 |
| tumour | 1,614 | 22,604 |
| leukemia | 3,206 | 19,018 |
| carcinoma | 2,981 | 16,631 |

### Query US document title/abstract

We query the final set of keyword stems against U.S. patent document titles and abstracts in EAST. We initially searched against the full text of U.S. patent documents but found that queries resulted in a large number of false positives. This is because claims can include clauses describing a broad range of possible applications of the invention. By limiting our search to titles and abstracts, we ensure our results include cancer-specific innovations and excludes patent documents that only reference a cancer-relevant term in the detailed description of the invention. Our cancer keyword query of U.S. patent document titles and abstract results in 100,635 patent documents.

### Query Derwent title and abstract

We also query the final set of keyword stems against the title and abstract of Derwent patent documents in DWPI via EAST. Again, our DWPI search results consist of Derwent patent families, which we limit to those containing U.S. documents. Our cancer keyword query of Derwent titles and abstracts results in a total of 157,206 patent families.

Search Method Assessment

We rely on internal resources available to patent examiners to assess our search strategy.

## Science and Technology Information Center

The Science and Technology Information Center (STIC) provides materials, resources, and services to assist patent examiners find prior art, obtain foreign patent information, and identify journal articles and books to support patentability decisions.[[12]](#footnote-12) We request two STIC searchers that specialize in Technology Center 1600 (Biotechnology and Organic Chemistry) review the set of cancer-related keywords and classification codes as well as the search strategies. We also consult with select Supervisory Patent Examiners in Technology Center 1600 on the search strategy.[[13]](#footnote-13)

## Chemical Abstract Plus query

Chemical Abstract Plus (CAplus) is a commercial database of chemical journal and patent publications available to examiners.[[14]](#footnote-14) To test the coverage of our keyword set, we query our list of cancer-specific MeSH terms against the CAplus indexing terms. About 93% of the patent documents retrieved from CAplus are in our final results set. We spot-check the remaining 5,021 documents identified in the CAplus query but not in our results set and find them to be less directly related to cancer. We omit the additional documents in the CAPlus query from the final result set despite greater sensitivity in the CAPlus query due to lower specificity in the non-overlapping set.

## Preliminary Visualizations

The USPTO has a license to use InnovationQ from IP.com, which facilitates visualizations based on classifications, assignees, inventors, text, dates, countries of origin, and enforceability. We create preliminary visualizations (not reported) to ensure that the assignees, inventors, text, and date distributions were reasonable. In addition, we construct visualizations (not reported) using the PatSeq tool from Lens.org to ensure that the distributions of non-patent literature references and genomic species-of-origin in our results were reasonable.

Results

We combine results from all query methods to obtain our final set of cancer relevant U.S. patent documents. However, because classification and keyword searches against DWPI yield Derwent patent families, we need to retrieve all U.S. documents numbers that fall within those identified families. Derwent patent family results include a reference to the earliest known patent document within the family. When that references is a U.S. document, we merge the Derwent patent family results into the EAST database by batch entering the U.S. document numbers into EAST's bulk Image Search & Retrieval form. This allows us to retrieve the CPC family ID, which USPTO uses to indicate related applications within USPTO databases. For some Derwent families, the earliest document in DWPI contains the WIPO identifier. We query these identifiers against the Foreign Patent Retrieval Service in EAST to retrieve the CPC family IDs. For the remaining Derwent families, we could only obtain foreign document identifiers. For these, we slowly retrieve the earliest U.S. document and CPC family ID using EAST's "Image Document" feature. Lastly, we use an internal USPTO table to link CPC family IDs to all U.S. document numbers to ensure we retrieve all U.S. documents associated with the Derwent patent families.

Our final combined set of search results contains 269,353 cancer-related U.S. patent documents spanning the 1976 to 2016[[15]](#footnote-15) time period.[[16]](#footnote-16) Figure 1 depicts the results by query method and data source. The results from each query method are not mutually exclusive, however each approach identifies some cancer-related patent documents not identified by the other approaches.

Figure 1: Query Results by Method and Source

**EAST**

**DWPI**

Patent Data

For our final data set, we extract key patent data fields from the EAST system, including patent title, patent CPC family ID, patent or publication ID, application ID number and filing date, publication or grant date, and patent classification codes (CPC, USPC, and IPC). Note that, for older documents, patent data coverage may be limited to ID numbers, patent grant date, patent tile and classification. We also construct some fields to assist data users.

## CPC-based technology categories

We generate a set of taxonomy fields based on the CPC classification to indicate the technological and functional features of cancer-related innovations. We group CPC classes and subclasses into seven high-level technology categories: Drugs & Chemistry (including small molecule and large protein-based therapeutics)[[17]](#footnote-17), Diagnostic & Surgical Devices (including in vitro diagnostics and medical devices)[[18]](#footnote-18), Radiation Measurement (including a subset of radiation therapy inventions)[[19]](#footnote-19), Data Science (including computer-related and imaging technologies)[[20]](#footnote-20), Food & Nutrition (including cancer prevention foods)[[21]](#footnote-21), Model Systems & Animals (including genetically modified cell lines and animals for testing of therapeutics and other experimentation)[[22]](#footnote-22), Cells & Enzymes (emphasizing cellular and molecular biology)[[23]](#footnote-23), and Other & Pre-classification (interdisciplinary inventions and documents which have not yet been classified)[[24]](#footnote-24). Since a patented innovation can span multiple technology fields, we create a set of seven indicator variables that designate whether the invention falls within a particular category.

## DNA, RNA, or Protein Sequence Indicator

Applicants that disclose nucleic acid or amino acid sequences in their patent documents are generally required to follow particular sequence rules.[[25]](#footnote-25) We generate a field to indicate whether the patent document includes such sequence data to assist users that want to focus on personalized medicine and genomics.

External Data

Lastly, we extract data from two external sources to assist users desiring to link the data set to upstream funding and R&D activity and downstream commercialization efforts.

## NIH RePORTER

NIH RePORTER provides detailed information on NIH-funded research that has resulted in patented inventions.[[26]](#footnote-26) We join our results to the RePORTER Patents Information[[27]](#footnote-27) on patent numbers. We then retrieve the NIH Federal Grant Number and NIH Grant Recipient Organization, which may differ from the Assignee or Applicant listed on the face of the patent document, to add to our final data set.

## FDA Orange Book

The Orange Book identifies drug products approved on the basis of safety and effectiveness by the FDA under the Federal Food, Drug, and Cosmetic Act.[[28]](#footnote-28) Each edition of the Orange Book contains a list of unexpired patents covering approved drug products.

We merge the data set to panel data constructed from the current edition of the Orange Book data[[29]](#footnote-29) as well as data from prior editions. We utilize the prior editions to ensure full coverage of patents that have previously expired. We join on patent number and extract the following fields from the Orange Book for each match:

1. FDA\_Application\_Number – FDA drug application numbers
2. FDA\_Drug\_Trade\_Name – Official trade name of the drug to which the application pertains
3. FDA\_Approval\_Date – Approval date for each FDA drug application
4. FDA\_Applicant – Name of the applicant for each FDA drug application
5. FDA\_Ingredient – Active ingredient for a specific indication approved by the FDA. Multiple ingredients are separated by a comma.

There can be multiple FDA applications for the drug covered by a single patent as well as multiple applicants, trade numbers, approval dates, and/or ingredients per FDA application number. For example, there are two FDA applications, identified by application numbers 22277 and 21029, for U.S. patent 5,260,291. Both applications were filed for the drug Temodar (active ingredient Temozolomide) by Merck Sharp Dohme. The first FDA application (22277) was approved on February 27, 2009, but the second FDA application (21029) has two FDA approval dates: August 11, 1999 and October 19, 2006.

To condense the FDA Orange Book data to the unique patent level, we grouped the data in an embedded list within each FDA variable. Multiple observations at the patent level are separated by a pipe (|). Thus, in our prior example, the FDA application number field in the data set contains “22277|21029” (see Figure 2). Multiple observations at the application level are separated by a semi-colon (;). Again, in our prior example, the FDA approval data field in the data set contains “Feb 27, 2009|Aug 11, 1999; Oct 19, 2006” (see Figure 2). Please note that order does matter.

Figure 2: Example of Embedded List in FDA Fields

|  |  |  |
| --- | --- | --- |
| Patent\_or\_Publication\_ID | FDA\_Application\_Number | FDA\_Drug\_Trade\_Name |
| US 5260291 A | 22277|21029 | TEMODAR|TEMODAR;TEMODAR |

|  |  |
| --- | --- |
| FDA\_Approval\_Date | FDA\_Applicant |
| Feb 27, 2009|Aug 11, 1999;Oct 19, 2006 | MERCK SHARP DOHME|MERCK SHARP DOHME;MERCK SHARP DOHME |

|  |
| --- |
| FDA\_Ingredient |
| TEMOZOLOMIDE|TEMOZOLOMIDE;TEMOZOLOMIDE |

Conclusion

Sharing knowledge and improving access to data is a key element of the National Cancer Moonshot. USPTO’s Horizon Scanning Tool contributes to that objective by removing barriers to accessing patent data relevant to cancer. In this report, we document the methods used to generate the USPTO Cancer Moonshot Patent Data for the Horizon Scanning Tool. This release is an initial step in leveraging patent data to reveal new insights into the prevention, diagnosis, and treatment of cancer. Moreover, we encourage external users to merge, extend, and refine the data set as well as provide feedback on its usefulness. We also plan to periodically update the data to include the most recent documents and emerging innovations that service the national effort to eliminate cancer.

# Appendix I: Data Set Dictionary

|  |  |
| --- | --- |
| Field Name | Definition |
| Family\_ID | CPC Family ID, identifies all related U.S. patent documents |
| Patent\_or\_Publication\_ID | U.S. patent number (8-character) for utility patent grants; U.S. publication number (11-digit) for published patent applications; U.S. patent number (7-character alphanumeric code starting with “D”) for design patents; Reissue patent number (7-character alphanumeric code starting with “RE”) for reissued patents |
| Application\_Number | U.S. patent application number |
| Filing\_Date | U.S. patent application filing date |
| Grant\_or\_Publication\_Date | Grant issue date for patent grants or publications date for published applications |
| CPC\_Inventive | CPC Inventive classification to capture the inventive subject matter |
| CPC\_Additional | CPC Additional classification to capture additional information for future searching |
| IPC\_Primary | IPC Primary classification (as of publication or grant date), single primary classification of the inventive subject matter |
| IPC\_Secondary | IPC Secondary classification (as of publication or grant date), non-primary classification covering inventive subject matter and additional information to facilitate future searching  |
| USPC \_Original | Original USPC classification (retired as of 1/1/2013), single principal classification  |
| USPC \_Cross\_Reference | Cross Reference USPC classification(retired as of 1/1/2013), additional classifications |
| Patent\_Title | Title of the patent or patent application |
| Drugs\_and\_Chemistry | CPC-based technology category, 1 indicates inventive subject matter pertaining to drugs and chemistry (CPC Inventive classification A61K or C07) |
| Diagnostic\_and\_Surgical\_Devices | CPC-based technology category, 1 indicates inventive subject matter pertaining to diagnostic and surgical devices (CPC Inventive classification C12Q, G01N, A61B, A61L, or B01) |
| Radiation\_Measurement | CPC-based technology category, 1 indicates inventive subject matter pertaining to radiation measurement (CPC Inventive classification G01T) |
| Data\_Science | CPC-based technology category, 1 indicates inventive subject matter pertaining to data science (CPC Inventive classification G06Q, G06F, or G06T) |
| Food\_and\_Nutrition | CPC-based technology category, 1 indicates inventive subject matter pertaining to food and nutrition (CPC Inventive classification A23, A21, or A22) |
| Model\_Systems\_and\_Animals | CPC-based technology category, 1 indicates inventive subject matter pertaining to model systems and animals (CPC Inventive classification A01K) |
| Cells\_and\_Enzymes | CPC-based technology category, 1 indicates inventive subject matter pertaining to cells and enzymes (CPC Inventive classification C12N) |
| Other\_and\_Preclassification | CPC-based technology category, 1 indicates inventive subject matter not fitting into the specified categories or not yet classified |
| DNA\_RNA\_or\_Protein\_Sequence | DNA, RNA, or Protein Sequence, 1 indicates presents of DNA, RNA, or protein sequence disclosure in the patent document |
| NIH\_Federal\_Grant\_Number | NIH Federal Grant Number (source: NIH RePORTER) |
| NIH\_Grant\_Recipient\_Organization | NIH Federal Recipient Organization (source: NIH RePORTER) |
| FDA\_Application\_Number | FDA Application Number (Multiple applications numbers are separated by a pipe; within application multiples are separated by a semicolon) (source: FDA Orange Book) |
| FDA\_Drug\_Trade\_Name | Official trade name of the drug to which the FDA application pertains (source: FDA Orange Book) |
| FDA\_Approval\_Date | Approval date for FDA drug application (source: FDA Orange Book) |
| FDA\_Applicant | Name of the applicant that filed the FDA drug application (source: FDA Orange Book) |
| FDA\_Ingredient | Active ingredient for a specific indication approved by the FDA (Multiple ingredients are separated by a comma) (source: FDA Orange Book) |

# Appendix II: Cancer Classification Codes

|  |  |  |
| --- | --- | --- |
| CPC | USPC | IPC |
| C12Q1/6886 | 514/19.3 | A61P35/00 |
| G01N33/574 | 435/6.14 | G01N33/574 |
| A61K39/0011 | 435/7.23 | A61P35/02 |
| G01N33/5011 | 424/155.1 | A61P35/04 |
| G01N33/57484 | 424/277.1 | C07K16/30 |
| C07K16/30 | 424/174.1 | C12N5/09 |
| G01N33/57492 | 514/19.4 | A61K35/13 |
| A61L2300/416 | 424/138.1 |  |
| G01N33/57434 | 530/388.8 |  |
| G01N33/53 | 514/19.5 |  |
| G01N33/57423 | 530/387.7 |  |
| G01N33/57419 | 514/19.2 |  |
| G01N33/57415 | 514/19.6 |  |
| G01N33/57407 | 424/181.1 |  |
| G01N33/57488 | 530/389.7 |  |
| C12N15/1135 | 514/19.8 |  |
| A61K47/48569 | 436/64 |  |
| G01N33/57438 | 530/388.85 |  |
| G06T2207/30096 | 424/156.1 |  |
| G01N33/57496 | 514/19.9 |  |
| G01N33/57426 | 435/330 |  |
| C07K16/3053 | 800/10 |  |
| C07K16/3069 | 435/344 |  |
| G01N33/57449 | 436/813 |  |
| C12N5/0693 | 435/344.1 |  |
| A23V2200/308 | 514/19.7 |  |
| C07K16/3015 | 977/911 |  |
| A61K47/48584 | 424/573 |  |
| C07K16/3007 | 977/912 |  |
| C07K16/3046 | 607/901 |  |
| C07K16/3061 |  |  |
| G01N33/57446 |  |  |
| A61K2039/5152 |  |  |
| A61K2039/585 |  |  |
| C07K16/303 |  |  |
| A61K51/1045 |  |  |
| G01N2800/7028 |  |  |
| G01N33/5743 |  |  |
| A61K35/13 |  |  |
| A61K47/48638 |  |  |
| C07K16/3023 |  |  |
| G01N33/5017 |  |  |
| G01N33/5748 |  |  |
| A61K35/768 |  |  |
| A61K47/4863 |  |  |
| G01N2333/82 |  |  |
| G01N33/57411 |  |  |
| C07K16/3076 |  |  |
| A61K47/48615 |  |  |
| G01N33/57442 |  |  |
| Y10S436/813 |  |  |
| A61K47/48576 |  |  |
| A61K51/1072 |  |  |
| C07K16/3038 |  |  |
| A61K51/1057 |  |  |
| A61K51/1048 |  |  |
| C12N5/0093 |  |  |
| A61K47/48607 |  |  |
| Y10S977/911 |  |  |
| A61K38/1764 |  |  |
| G01N33/57476 |  |  |
| G01N33/57469 |  |  |
| A61K51/106 |  |  |
| A61K45/05 |  |  |
| G01N2033/57403 |  |  |
| G01N2033/57461 |  |  |
| G01N2033/57465 |  |  |
| G01N2033/57453 |  |  |
| G01N2033/57457 |  |  |

# Appendix III: Cancer EPI & CPI Codes

|  |
| --- |
| Derwent EPI & CPI Codes |
| B04-B04C4  | B14-H01E6  | C14-H01F6  |
| B04-B04C8  | B14-H01E7  | C14-H01G  |
| B04-F02A  | B14-H01E8  | C14-H01H  |
| B04-F02A0E  | B14-H01F  | C14-H01H1  |
| B04-G05  | B14-H01F1  | C14-H01J  |
| B04-G0500E  | B14-H01F2  | C14-H01J1  |
| B04-K01S  | B14-H01F3  | C14-H01K  |
| B04-K01S0E  | B14-H01F4  | C14-H01K  |
| B12-G07  | B14-H01F5  | C14-H01K1  |
| B12-K04A1  | B14-H01F6  | C14-H01K2  |
| B12-K04G2A  | B14-H01G  | C14-H01K3  |
| B14-H  | B14-H01G  | C14-H01L  |
| B14-H00X  | B14-H01H  | C14-H01M  |
| B14-H01  | B14-H01H1  | C14-H01N  |
| B14-H01C  | B14-H01J  | C14-H01R  |
| B14-H01D  | B14-H01J1  | C14-H01S  |
| B14-H01D1  | B14-H01K  | C14-H01T  |
| B14-H01D2  | B14-H01K1  | C14-H01U  |
| B14-H01E  | B14-H01K2  | C14-H01V  |
| B14-H01E1  | B14-H01K3  | C14-H01X  |
| B14-H01E2  | B14-H01L  | C14-H01Y  |
| B14-H01E3  | B14-H01M  | C14-H01Z  |
| B14-H01E4  | B14-H01N  | C14-H05  |
| B14-H01E5  | B14-H01R  | C14-H06  |
| B14-H01E6  | B14-H01S  | C14-S11C  |
| B14-H01E7  | B14-H01T  | C14-S11E  |
| B14-H01E8  | B14-H01U  | D05-H15A  |
| B14-H01F  | B14-H01V  | S05-G02G4  |
| B14-H01F1  | B14-H01X  |  |
| B14-H01F2  | B14-H01Y  |  |
| B14-H01F3  | B14-H01Z  |  |
| B14-H01F4  | B14-H05  |  |
| B14-H01F5  | B14-H06  |  |
| B14-H01F6  | B14-S11C  |  |
| B14-H01G  | B14-S11E  |  |
| B14-H01G  | C04-B04C4  |  |
| B14-H01H  | C04-B04C8  |  |
| B14-H01H1  | C04-E05  |  |
| B14-H01J  | C04-F02A  |  |
| B14-H01J1  | C04-F02A0E  |  |
| B14-H01K  | C04-G05  |  |
| B14-H01K1  | C04-G0500E  |  |
| B14-H01K2  | C04-K01S  |  |
| B14-H01K3  | C04-K01S0E  |  |
| B14-H01L  | C12-G07  |  |
| B04-B04C4  | C12-K04A1  |  |
| B04-B04C8  | C12-K04G2A  |  |
| B04-F02A  | C14-H  |  |
| B04-F02A0E  | C14-H00X  |  |
| B04-G05  | C14-H01  |  |
| B04-G0500E  | C14-H01C  |  |
| B04-K01S  | C14-H01D  |  |
| B04-K01S0E  | C14-H01D1  |  |
| B12-G07  | C14-H01D2  |  |
| B12-K04A1  | C14-H01E  |  |
| B12-K04G2A  | C14-H01E1  |  |
| B14-H  | C14-H01E2  |  |
| B14-H00X  | C14-H01E3  |  |
| B14-H01  | C14-H01E4  |  |
| B14-H01C  | C14-H01E5  |  |
| B14-H01D  | C14-H01E6  |  |
| B14-H01D1  | C14-H01E7  |  |
| B14-H01D2  | C14-H01E8  |  |
| B14-H01E  | C14-H01F  |  |
| B14-H01E1  | C14-H01F1  |  |
| B14-H01E2  | C14-H01F2  |  |
| B14-H01E3  | C14-H01F3  |  |
| B14-H01E4  | C14-H01F4  |  |
| B14-H01E5  | C14-H01F5  |  |

# Appendix IV: Cancer Japanese F-Terms and F-Index Codes

|  |  |
| --- | --- |
| F-Term | F-Index |
| 2G045 AA26 | G01N33/574 |
| 4B018 ME08 | G01N33/574 A |
| 4C082 MA01 | G01N33/574 B |
| 4C085 AA21 | G01N33/574 C |
| 4C085 LL18 | G01N33/574 D |
| 4C086 ZB26 | C07K16/30 |
| 4C088 ZB26 | C07K16/32 |
| 4H048 AB28 | C12N5/09 |
|  | A61B10/00 T |
|  | A61K39/395 E |
|  | A61K39/395 T |
|  | A61M1/36 555 |

# Appendix V: Cancer Keyword Stems

|  |
| --- |
| Query Patterns for Keywords ($=wildcard) |
| $ACANTHOMA | $NEUROTHEKEOMA$ |
| $ADAMANTINOMA$ | $ODONTOMA$ |
| $ADENOCARCINOM$ | $OPSOCLONUS$1MYOCLONUS$ |
| $ADENOMATOID$ | $OSTEOCHONDROMATOSIS$ |
| $ADENOMATOUS$ | $PARAGANGLIOMA$ |
| $ADENOMYOMA$ | $PERICYTOM$ |
| $ANGIOFIBROMA$ | $PEUTZ$1JEGHERS$ |
| $ANGIOKERATOMA$ | $PHEOCHROMOCYTOMA$ |
| $ANGIOLIPOMA$ | $PILOMATRIXOMA$ |
| $ANGIOMYOLIPOMA$ | $PINEALOMA$ |
| $APUDOMA | $PLASMACYTOMA$ |
| $BIRT$1HOGG$1DUBE$ | $PROLACTINOMA$ |
| $BLASTOMA$ | $PSEUDOMYXOMA$ |
| $BRACHYTHERAP$ | $RHABDOID$ |
| $BRAF-RESISTANT$ | $SARCOMA$ |
| $CANCER$ | $SOMATOSTATINOMA$ |
| $CANCEROSTA$ | $TERATOMA$ |
| $CARCINOGEN$ | $THECOMA |
| $CARCINOID$ | $THELIOMA$ |
| $CARCINOM$ | $THYMOMA$ |
| $CARCINOSARCOMA$ | $TUMOR$ |
| $CEMENTOMA$ | $TUMOUR$ |
| $CHONDROMA | $ZOLLINGER$1ELLISON$ |
| $CHORDOMA$ | ENCEPHALITIS" |
| $CONDROMA$ | ERYTHROPLASIA |
| $CRANIOPHARYNGIOMA$ | EXOSTOSES |
| $CYSTADENOMA | FIBROMA |
| $DENYS$1DRASH$ | GLIOMA |
| $ENDOTHELIOM$ | LIPOMA |
| $EPITHELIOMA$ | NEVUS |
| $ERYTHROLEUKEM$ | OSTEOMA |
| $FRAUMENI$ | POLYPS |
| $GANGLIONEUROMA$ | VIPOMA |
| $GERMINOMA$ | WAGR |
| $GLIOBLASTOM$ |  "WILMS TUMOR" |
| $HAMARTOMA | "ABERRANT CRYPT FOCI" |
| $HYPERPLASIA$ | "ACTH SYNDROME" |
| $LAMBERT$1EATON$ | "ACTINIC KERATOSIS" |
| $LEUKAEMI$ | "ADENOMATOUS POLYPOSIS" |
| $LEUKEMI$ | "ATYPICAL SQUAMOUS" |
| $LEUKOPLAKIA$ | "CERVICAL DYSPLASIA" |
| $LEUKOSIS$ | "ECTOPIC ACTH" |
| $LIPOBLASTOMA$ | "EWING SARCOMA" |
| $LUTEOMA | "EWINGS SARCOMA" |
| $LYMPHANGIOLEIOMYOMATOSIS$ | "EWING'S SARCOMA" |
| $LYMPHANGIOMA$ | "HODGKIN LYMPHOMA" |
| $LYMPHANGIOMYOMA$ | "HODGKINS LYMPHOMA" |
| $LYMPHOM$ | "HODGKIN'S LYMPHOMA" |
| $MACROGLOBULINEMIA$ | "LIMBIC |
| $MALIGNAN$ | "LYMPHOMATOID GRANULOMATOSIS" |
| $MASTOCYTOMA$ | "LYPHOMATOID GRANULOMATOSIS"  |
| $MASTOCYTOSIS$ | "PAGET DISEASE" |
| $MELANOM$ | "SMALL ADENOMA PLEOMORPHIC" |
| $MELANOTIC$ | "SQUAMOUS INTRAEPITHELIAL" |
| $MENINGIOMA$ | "TRANSVERSE MYELITIS" |
| $MESENCHYMOMA$ | "TUBEROUS SCLEROSIS" |
| $MESONEPHROMA$ | "URTICARIA PIGMENTOSA" |
| $METASTASE$ | "XERODERMA PIGMENTOSUM" |
| $METASTASI$ |  |
| $MYELOCYTIC$ |  |
| $MYELOLIPOMA$ |  |
| $MYELOMA$ |  |
| $NEOPLASIA$ |  |
| $NEOPLASM$ |  |
| $NEOPLASTIC$ |  |
| $NEPHROMA$ |  |
| $NEURILEMMOMA$ |  |
| $NEUROCYTOMA$ |  |
| $NEUROMA |  |

1. https://www.whitehouse.gov/CancerMoonshot [↑](#footnote-ref-1)
2. https://www.whitehouse.gov/the-press-office/2016/06/28/fact-sheet-cancer-moonshot-summit-vice-president-biden-announces-new [↑](#footnote-ref-2)
3. http://www.cooperativepatentclassification.org/index.html [↑](#footnote-ref-3)
4. http://www.uspto.gov/web/patents/classification/selectnumwithtitle.htm. USPC classification was officially replaced by CPC classification as of January 1, 2013. The USPTO continued to assign USPC to incoming applications after that date for routing purposes only. [↑](#footnote-ref-4)
5. http://www.wipo.int/classifications/ipc/en/ [↑](#footnote-ref-5)
6. Derwent patent families contain patent documents related to a single invention from multiple jurisdictions, which may include U.S., European, and Japanese patent applications and grants. Each patent family is grouped around a Basic patent, which is usually the first published example of the invention. All subsequent filings, termed Equivalent patents, are referred back to the Basic patent. For more on the most commonly used definitions of patent families, see https://www.oecd.org/sti/inno/44604939.pdf. [↑](#footnote-ref-6)
7. http://ip-science.thomsonreuters.com/m/pdfs/cpi\_manualcodes.pdf [↑](#footnote-ref-7)
8. http://ip-science.thomsonreuters.com/m/pdfs/epi\_manualcodes1.pdf [↑](#footnote-ref-8)
9. We do not utilize the third broad category which covers engineering patents. [↑](#footnote-ref-9)
10. https://www.nlm.nih.gov/bsd/disted/meshtutorial/introduction/02.html [↑](#footnote-ref-10)
11. www.ncbi.nlm.nih.gov/mesh/68009369 [↑](#footnote-ref-11)
12. http://www.uspto.gov/patent/initiatives/scientific-and-technical-information-center-stic-awareness-campaign [↑](#footnote-ref-12)
13. One SPE suggested that we expand the search to include international documents using multiple languages with the Cross-Lingual-Expansion feature of WIPO’s Patentscope. We decided not to expand our search to foreign-language documents because the aim of the Horizon Scanning Tool is to visualize patenting activity within the United States. We may revisit this in the future. [↑](#footnote-ref-13)
14. http://www.cas.org/File%20Library/Training/STN/DBSS/caplus.pdf [↑](#footnote-ref-14)
15. The data set contains roughly 560 patents (or 0.2 percent) granted prior to 1976. [↑](#footnote-ref-15)
16. We removed four older patents from the final data set due to irrelevance to cancer innovation: U. S. patent 35,097; U.S. patent 419,324; U.S. patent 527,040; and U.S. patent 1,924,793. [↑](#footnote-ref-16)
17. Based on CPC Inventive classification A61K or C07. [↑](#footnote-ref-17)
18. Based on CPC Inventive classification C12Q, G01N, A61B, A61L, or B01. [↑](#footnote-ref-18)
19. Based on CPC Inventive classification G01T. [↑](#footnote-ref-19)
20. Based on CPC Inventive classification G06Q, G06F, or G06T. [↑](#footnote-ref-20)
21. Based on CPC Inventive classification A23, A21, or A22. [↑](#footnote-ref-21)
22. Based on CPC Inventive classification A01K. [↑](#footnote-ref-22)
23. Based on CPC Inventive classification C12N. [↑](#footnote-ref-23)
24. Indicates inventive subject matter not fitting into the specified categories or not yet classified. [↑](#footnote-ref-24)
25. https://www.uspto.gov/web/offices/pac/mpep/s2421.html [↑](#footnote-ref-25)
26. For detailed information on NIH RePORTER, see https://report.nih.gov/faq.aspx?sid=2#36. [↑](#footnote-ref-26)
27. We utilized the “Patents Information with project references for all Fiscal years” data file updated through 8/1/2016. It is available to download from the ExPORTER site at https://exporter.nih.gov/ExPORTER\_Catalog.aspx?sid=0&index=3 [↑](#footnote-ref-27)
28. For detailed information on the FDA Orange Book, see http://www.fda.gov/Drugs/DevelopmentApprovalProcess/ucm079068.htm [↑](#footnote-ref-28)
29. Downloaded from www.accessdata.fda.gov/scripts/cder/ob/ [↑](#footnote-ref-29)