

How Have Cancer Clinical Trial Eligibility Criteria Evolved Over Time?

Anil Yaman, MS, Shreya Chakrabarti, PhD, Anando Sen, PhD and Chunhua Weng, PhD

Department of Biomedical Informatics, Columbia University, New York, NY USA

Abstract

Knowledge reuse of cancer trial designs may benefit from a temporal understanding of the evolution of the target populations of cancer studies over time. Therefore, we conducted a retrospective analysis of the trends of cancer trial eligibility criteria between 1999 and 2014. The yearly distributions of eligibility concepts for chemicals and drugs, procedures, observations, and medical conditions extracted from free-text eligibility criteria of 32,000 clinical trials for 89 cancer types were analyzed. We identified the concepts that trend upwards or downwards in all or selected cancer types, and the concepts that show anomalous trends for some cancers. Later, concept trends were studied in a disease-specific manner and illustrated for breast cancer. Criteria trends observed in this study are also validated and interpreted using evidence from the existing medical literature. This study contributes a method for concept trend analysis and original knowledge of the trends in cancer clinical trial eligibility criteria.

Introduction

Randomized controlled trials are fundamental to the advancement of medicine, enabling the most rigorous testing of interventions for the prevention, diagnosis and treatment of diseases and conditions. The goal of any well-designed clinical trial is to answer a clinical question as decisively and inexpensively as possible, while ensuring that a large enough sample size of patients is included in the trial. Thus, an important step in the design of clinical trial experiments is the definition of target population, or the determination of eligibility criteria, to ensure an optimal sample size of enrollment in a trial with measurable outcomes^{1,2}. Various information theoretic approaches have been investigated in order to address this design question in clinical trials^{1,3}.

Knowledge reuse and management for clinical trial eligibility criteria has been an important goal of the medical informatics research community since the 1980s⁴. Most of these efforts have been focused on improving the formalism or computability of eligibility criteria, with relatively few studies focusing on their content management, largely due to the challenges in parsing free-text eligibility criteria. One of the notable efforts for the latter is the standardization of core eligibility criteria led by Niland⁵. With the advancement of medical language processing in the medical informatics community in the past two decades, we are presented with new opportunities for efficient and scalable knowledge acquisition from free-text clinical trial eligibility criteria^{6,7}. In this paper, we aim to gain a temporal understanding of frequently used patient characteristics in the eligibility criteria of cancer clinical trials.

Trend analysis has been routinely used in clinical research to gain temporal knowledge. Huang et al. studied the surge in vitamin D deficiency patients between 2007 and 2010⁸. The surge was attributed to vitamin D deficiency screening for preventive care. Carle compared various ways of fitting multi-level models on survey data for children with special healthcare needs⁹. Bieler et al. assessed the health risk for a group of white adults who were unable to afford prescription medication¹⁰. Rosenberg confirmed the trends of lower mortality rates for mothers during childbirth (along with trends for other conditions)¹¹. These studies have generally addressed one condition or one group at a time, which may limit the generalizability of their findings. Therefore, we aim to examine multiple conditions altogether in our study.

We have previously conducted the trend and network analysis of common eligibility features in cancer trials by (a) identifying frequently used eligibility features for clinical trials of a particular cancer, or, across different cancer types and (b) by visualizing features that were adopted or abandoned over time¹². Extending that study, this study contributes a more fine-grained quantitative method to answer the following questions: (1) Can we identify concepts which are consistently increasing or decreasing in incidence across the clinical trials for all cancer types? (2) Do certain concepts show anomalous trends in use for only a particular type of cancer trial, while showing opposing trends for all other cancer types? Can this be explained in terms of available medical literature? (3) For a specific cancer type (e.g. breast cancer), can we group concepts as having similar or complementary trends? Furthermore, this study investigates 89 different cancer types in parallel to inform effective design of a large section of cancer trials.

The paper is organized as follows: The *Methods* section describes the organization and pre-processing of the dataset along with the tools used for the trend analysis. In the *Results* section, we provide numerical evidence to answer the

questions mentioned above. Possible explanations of these answers are discussed in the *Discussion* section, followed by concluding remarks.

Methods

Dataset

As of May 2015, the free-text eligibility criteria of 32,000 distinct clinical trials for 89 cancer types, each having at least one trial (see list in **Error! Reference source not found.**) between the years 1999 and 2014, were downloaded from ClinicalTrials.gov¹³. The clinical trials were grouped by these 89 cancer types based on the values in their condition fields in ClinicalTrials.gov and stamped by their start dates. The number of clinical trials belonging to a particular cancer type ranged from 5,370 (breast cancer) to 3 (AIDS-related cancer).

Preprocessing and Concept Recognition

For each trial, the eligibility criteria text was first pre-processed and split into sentences. All the sentences were further parsed using a previously reported dictionary matching algorithm^{14,15}, which utilizes the Unified Medical Language System (UMLS)¹⁶ to identify clinical concepts in an unsupervised fashion. The UMLS ontology was designed by experts in order to cross-walk major biomedical terminologies or ontologies and provides comprehensive medical knowledge. It consists of concepts from various sources and includes definitions as well as the relations and semantic mappings of the concepts. Each concept in the UMLS is identified by its Concept Unique Identifier (CUI)¹⁶ which includes a number of terms (also called atoms) that are synonymous. Moreover, the concepts are assigned to one or more semantic types that are identified by their Type Unique Identifiers (TUI)¹⁶. Example eligibility criteria and their parsing output is available at <http://is.gd/EliXR2>. For example, in clinical trial NCT00035308, for criterion “*malignant disease or immunodeficiency syndrome*,” we extracted two UMLS concepts, “malignant disease”(CUI=C0442867) and “immunodeficiency syndrome” (CUI= C002105). The dictionary matching process involves generating all possible n -grams (i.e. a sequence of n consecutive words) of up to 5 words starting from the beginning of the sentences. The generated n -grams were matched against the UMLS atoms, and the longest matching n -gram was accepted and mapped to its CUI. After this processing step, the algorithm retained 39,900 distinct concepts. For each concept, we identified the number of uses in each cancer type and each year between 1999 and 2014.

The number of newly opened cancer trials over years showed an increasing trend as demonstrated in Figure 1. Since the concepts are more likely to appear in the recent years, we normalized the yearly frequencies of the concepts by the total number of trials conducted each year and the number of disease pairs. We further utilized the semantic mapping in the UMLS to classify the concepts from the semantic groups *Chemicals and Drugs*, *Procedures*, *Observation*, and *Disorders and Conditions*.

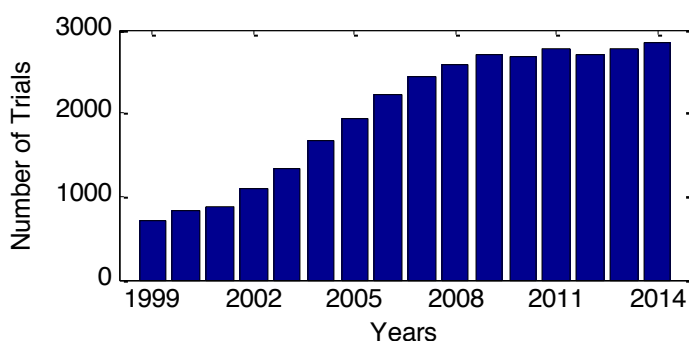


Figure 1: The distribution of the number of newly opened cancer trials every year between the years 1999 and 2014.

Assessing the Trend of the Concepts

We are interested in the temporal dynamics of the presence of the concepts in clinical trial eligibility criteria. Parametric and non-parametric statistical tests such as the Mann-Kendall test and Linear Regression are routinely used to identify the existence of a trend within time dependent data, and thus assess the positive or negative change over time. The Mann-Kendall test is a nonparametric statistical test for assessing monotonic trends in the data¹⁷⁻¹⁹. Using the standard design of hypothesis tests, we assume H_0 to be the null hypothesis that no monotonic trend is present and H_A to be the alternative hypothesis. The null hypothesis is rejected at a significance level α if the corresponding p -value is less than α . Linear regression²⁰, on the other hand, is a parametric method where the significance of the linear

relationship between an independent variable X and a dependent variable Y is tested. The relationship is described by a linear regression line, shown in equation (1). Significance testing is performed to test for the slope β_1 of the regression line to be non-zero, which would indicate a significantly increasing or decreasing trend in the use of a concept over time.

$$Y = \beta_0 + \beta_1 X \quad (1)$$

Unlike the Mann-Kendall test, linear regression can provide the slope (rate of change) information of the trend. Therefore, we considered ‘year’ as an independent variable (X) and the corresponding normalized frequencies of the concepts as the dependent variable (Y); and assessed the slope using linear regression. We tested the null hypothesis $H_0: \beta_1 = 0$ using the F-statistic at a significance level of $\alpha = 0.05$. The H_0 null hypothesis is rejected if the p -value is smaller than 0.05, and H_A which is $\beta_1 \neq 0$ is accepted²⁰. The sign of the slope is interpreted as the direction of the trend (positive: up; negative: down); and the magnitude of its value is interpreted as its strength. For the cases in which the significance test failed to reject the null hypothesis, it is assumed that there is no change in trends.

Diseases Using Concepts that have Similar Trends

We generated a feature vector for each cancer type which encodes the direction of the trend of the concepts. Each disease feature vector is an n -dimensional array where n is the total number of concepts. The position of each concept within the feature vector is marked by -1, +1 and 0 if the slope of the concept is negative, positive, or equal to 0 respectively. The similarity between two feature vectors is measured using the Jaccard distance measure²¹ which can take a value between 0 and 1 signifying the strength of the similarity. In our case, this is calculated by dividing the number of similarly trending concepts between two disease vectors by the number of concepts appearing to be trending in either of the vectors, and subtracting the result from 1. The lower the Jaccard distance measure is between two diseases, the more is the number of concepts with similar trends shared by them.

Results

Table 1. The top 5 concepts from each semantic group ranked by the percentage (and the number) of cancer types for which they exhibit upward or downward trend. For example, HIV vaccine exhibits an upward trend in 65 or 73% of the 89 cancer types, while creatinine exhibits a downward trend in 68 or 76% of the 89 cancer types.

Chemicals & Drugs			
Upward Trending Concepts	Percentage	Downward Trending Concepts	Percentage
HIV Vaccine	73% (N=65)	Creatinine	76% (N=68)
Cytochrome P450 3A4	49% (N=44)	Mitomycins	61% (N=54)
Crotoxin	46% (N=41)	Nitrosoureas Antineoplastic Alkylating Agents	45% (N=40)
Blood Group Antigen C	45% (N=40)	Hormone Receptor	39% (N=35)
Cocaine	45% (N=40)	Antiemetics	38% (N=34)
Procedure			
Interventional Procedure	51% (N=45)	Hormone Therapy	92% (N=82)
Cognitive Therapy	46% (N=41)	Biological Response Modifier Therapy	89% (N=79)
Contraception Barrier	43% (N=38)	Combined Modality Therapy	84% (N=75)
Nursing Therapy	40% (N=36)	Pharmacotherapeutic	73% (N=65)
Hysterectomy	40% (N=36)	Chemotherapy Regimen	72% (N=64)
Observation			
History of Present Illness	65% (N=65)	Performance Status	82% (N=73)
HIV Seropositivity	64% (N=64)	Pregnancy Test Negative	81% (N=72)
Pregnancy Test	46% (N=34)	Platelet Count	79% (N=70)
Kidney Function Test	35% (N=32)	Granulocyte Count	64% (N=57)
Electrocardiography	35% (N=32)	White blood Count	50% (N=45)
Disorders & Conditions			
Pregnancy	76% (N=68)	Urologic Diseases	91% (N=81)
HIV Infections	73% (N=65)	Not Pregnant	89% (N=79)
Adverse Event	66% (N=59)	HIV Negative	64% (N=57)
Breastfeeding	63% (N=56)	Skin Carcinoma	21% (N=18)
Heart Problem	56% (N=50)	Primary Malignant Neoplasm	21% (N=18)

After the trend assessment process, we identified 6,287 distinct concepts that have significant slopes for at least one of the cancer types. Out of all the concepts, we identified 2,937 concepts which constituted four semantic groups of interest. Out of these, 736 concepts belonged to the semantic group *Chemicals and Drugs*, 1328 to the *Disorders and Conditions*, 449 to the semantic group *Observation*, and 424 of them to *Procedures*. The remaining concepts were not included in this trend analysis.

Increasing or Decreasing Trends

We found that 37.32% of all concepts from four semantic groups (N=1096) exhibit a downward trend, and 79.61% of them (N=2338) exhibit an upward trend in at least one cancer type. The concepts were ranked based on the percentage of the cancer types for which they exhibit upward or downward trends. The list of concepts shown in Table 1 was generated by taking the top 5 concepts from each group based on the percentage of the cancer types that they exhibit increasing or decreasing trends for. A detailed discussion of these trends is presented in the next section. Some of the concepts appeared to have a significant increasing or decreasing trend in only one type of cancer type. We identified them as disease specific concepts, and these constituted 32.55% of all concepts (N=956).

Figure 2 shows an example of a concept demonstrating a sharply increasing trend, namely, “Pregnancy” (a) and a concept showing a strong decreasing trend, namely, “Hormone Therapy” (b) in the 15 types of cancers that had the highest number of trials. Each cell in the figures shows the normalized frequency of a concept for a given year (x-axis) and a cancer type (y-axis). The cells are color coded based on the magnitude of the frequency values as shown on the vertical bar. The concept “Pregnancy” shows a sharply increasing trend in almost all cancer types with the exception of prostate cancer, which is a purely male-based cancer type. It can also be seen that the incidence of this concept starts spiking up around the years 2003-2004, after which it increases sharply in the most common types of cancer. On the right (b), the concept “Hormone Therapy” shows a strong downward trend up to the year 2004, after which it is relatively rare in incidence among these types of cancer.

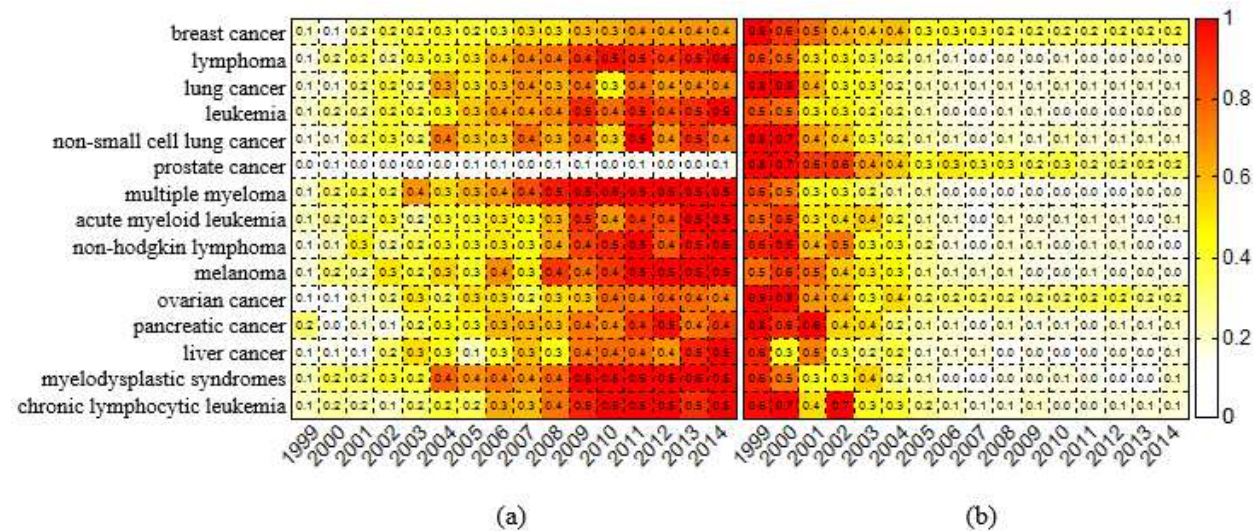


Figure 2. The distribution of yearly frequencies of the concepts “Pregnancy” (a) and “Hormone Therapy” (b) over the topmost 15 types of cancers, ranked based on the number of trials they have.

Anomalous Trends

An interesting aspect of our study was to see how some of the concepts generally exhibit a trend for most cancers but the opposing trend for a particular type of cancer. “SGOT - Glutamate oxaloacetate transaminase” is one such concept which decreases for 31% (N=28) of cancers and increases only for hairy cell leukemia. Hairy cell leukemia is a very rare form of cancer and there are relatively few trials addressing it. Hence, this increase may be attributed to chance. “Corticosteroids” decreases in use for 13% (N=12) of the cancers and increases only in multiple myeloma trials. The decrease is due to the fact that this drug weakens the immune system and puts patients at risk for infections. However, in the case of multiple myeloma, corticosteroids form an important part of the treatment which may explain its elevated use²². Similarly, “Adrenal Cortex Hormones” decreases for 11% (N=10) of the cancers and increases only for multiple

myeloma. Some studies have shown that multiple myeloma can be treated by massive amounts of adrenal cortex hormone²³ while others have warned about its side effects²⁴.

The concept “*X-Ray Computed Tomography*” shows a decreasing trend only for two cancer types, i.e., brain cancer and astrocytomas, but increases in use for 21.3% of cancers (N=19). This may be attributed to the use of more sophisticated techniques for brain cancer and astrocytoma detection so that “*X-Ray Computed Tomography*” gets retired. Also, “*Platelet Count measurement*” is found to show an increasing trend only for multiple myeloma, whereas it decreases for 63% (N=56) of cancer types. Multiple myeloma is a relatively rare type of cancer, and platelet count has been found to be one of the most important prognostic factors for it²⁵. In the *Disorders and Conditions* semantic group, two concepts “*carcinoma in situ*” and “*squamous cell carcinoma of skin*” showed anomalous trends. “*Carcinoma in situ*” is a very general term and in fact several concepts relating to carcinoma in situ for a particular organ were a part of our study. This concept trends upwards for 8% (N=7) and downwards for 37% (N=33) of cancer types. There were no anomalous trends detected for any of the organ specific carcinomas in situ and hence this disparity can be a result of the concept being too generic. The concept “*squamous cell carcinoma*” showed an increase only for multiple myeloma while it was decreasing for 31.5% (N=28) other cancer types. This can be attributed to this concept being organ-specific to the skin.

Diseases with Similarly Trending Concepts

Figure 3 and Figure 4 show the groups of diseases with similar trending concepts for the semantic groups *Chemicals and Drugs*, and *Procedures* respectively. The graphs are generated by measuring the distance between the feature vectors of the cancer types (as described in Methods section) using the Jaccard distance measure. Only the connections between the diseases with a Jaccard distance smaller than 0.65 and 0.55 respectively are shown. The thresholds for the Jaccard distances were chosen independently in each figure for visual clarity.

We hypothesized that some cancers would be clustered together for sharing many concepts with similar trends. This is well confirmed in Figure 3, in which we showed four clusters divided by the red lines: one for the various types of leukemia, one for two types of lymphoma, one for sarcoma, and the last for other cancers. It is also interesting to note here that the fourth cluster shows the two types of lung cancer grouped together with breast cancer and two types of ovarian cancer. We investigated this cluster in detail and noticed that certain concepts such as “*creatinine*” and “*filgrastim*” decrease in their use significantly for breast cancer and the two types of lung cancer, and certain others such as “*inhibitors*”, “*baseline dental cement*”, and “*HIV vaccine*” increase for these cancer types as well. These concepts are more generic concepts, and not very disease-specific, and thus, we consider the fourth cluster as being associated with a more generic concept trend, shared by the three common cancer types, namely, ovarian, breast and lung cancer.

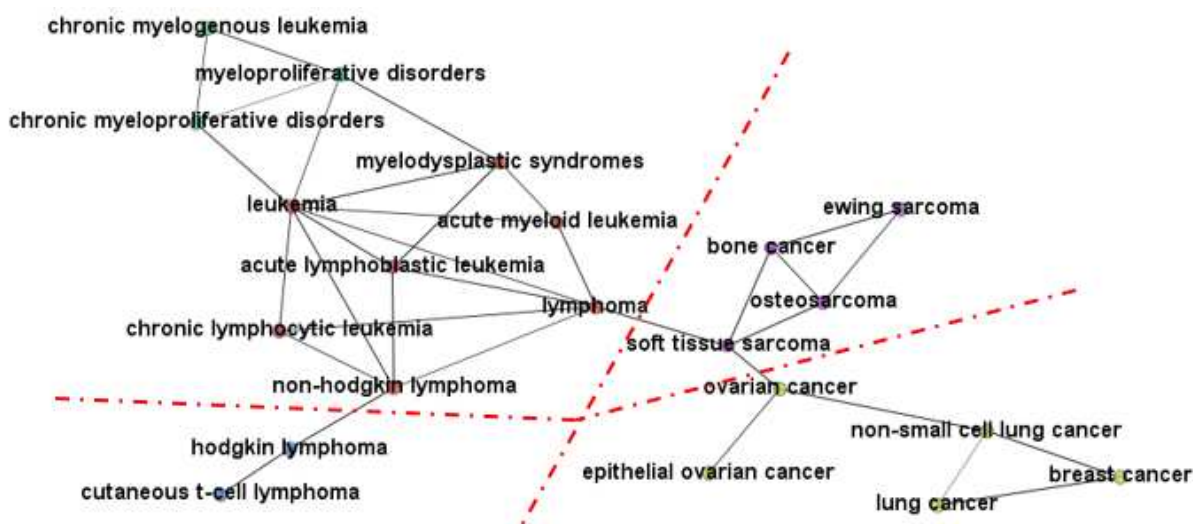


Figure 3. A network that shows shared concepts belonging to the *Chemicals and Drugs* semantic group with similar trend in behavior between cancer diseases. Each node represents a cancer type; and each edge represents the value of Jaccard distance measure between the two diseases it connects. Only the edges where the Jaccard measure is less than 0.65 are shown.

It can also be seen that the diseases were similarly clustered for the *Chemicals and Drugs* and the *Procedures* semantic groups. An interesting observation is that lymphoma is the most connected cancer, showing similarity in concept trends with most other cancers, as shown in Figure 4.

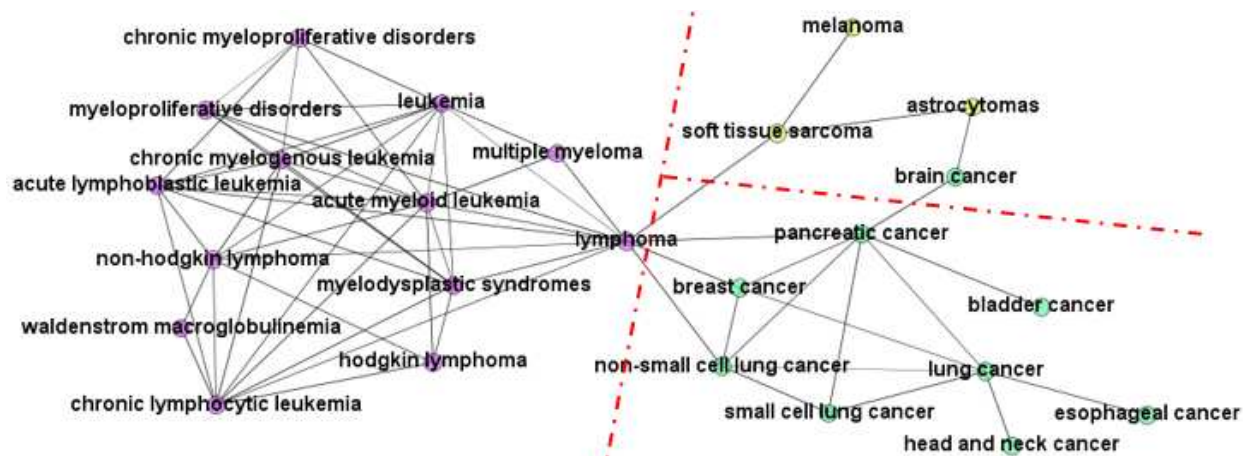


Figure 4. A network that shows shared concepts belonging to the *Procedures* semantic group with similar trend between cancer diseases. Each node represents a cancer type; and each edge represents the value of Jaccard distance measure between the two diseases it connects. Only the edges where the Jaccard measure is less than 0.55 are shown.

Also, the threshold used for the Jaccard similarity is much lower for the *Procedures* semantic group than it is for the *Chemicals and Drugs* semantic group. In spite of this, the number of cancer types which share similar trends in *Procedures* is much higher than that for the latter semantic group, as can be seen on comparison of Figure 3 and Figure 4. This can indicate that concepts from the semantic group *Procedures* are more shared among cancer types, i.e. they represent more generic concepts. Although we cannot illustrate this with a figure due to the lack of space, the concepts from the semantic group *Chemicals and Drugs* were more disease-specific, and hence less shared among cancer types.

Concepts with Complementary Trends within a Specific Disease

The next part of the analysis was to assess the trends in the four most common cancer types: breast cancer, prostate cancer, lung cancer and skin cancer. The concepts which were increasing or decreasing for prostate, lung and skin cancer were found to show the same trend for most other cancer types.

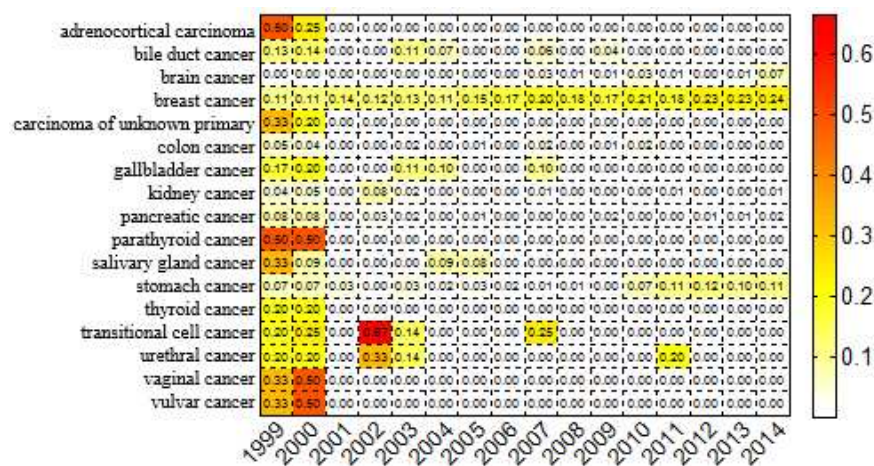


Figure 5. The yearly distribution of the frequencies of the concept “*erbB2 Receptor*” on 17 cancer types between the years from 1999 to 2014.

The concept “*hormone receptor*” showed the strongest decline in breast cancer. However, the concepts “*estrogen receptor*” and “*progesteron receptor*” showed substantial increase. This phenomenon can be attributed to the fact that *hormone* has been replaced by the more specific hormones “*estrogen*” and “*progesteron*”²⁶ in practice. The concept “*erbB-2 receptor*” demonstrated the sharpest increase for breast cancer. This is shown in **Figure 5** where the frequency

change of the concept “*erbB-2 receptor*” is shown along with 17 other types of cancer which are the only ones that the concept exhibits significant trends for. Amplification of this protein converts cultured cells into a cancerous phenotype, which is known to occur in 15-20% of breast cancers²⁷. The increase of “*Erbium*” in breast cancer (along with epithelial ovarian cancer and uterine sarcoma) may be due to the development of the vaginal erbium laser, which is used for thermotherapy in breast cancer survivors²⁸. It was also interesting to note that the concepts “*Alanine Transaminase*” and “*Calcium*” had rising trends for about half of the cancers.

Discussion

Interpretation of the Trends

We tried to identify related medical literature to help interpret the trends shown in Table 1. The upward trending concepts can be potentially attributed to rising awareness of the related patient characteristics or screening methods. Several of the concepts trending upward are about cancer prevention agents and a few others are related to general health. The development of a Human Immunodeficiency Virus (HIV) vaccine has received tremendous funding²⁹. Clinical trials for HIV vaccines are often used to screen cancer patients according to the National Institutes of Health website³⁰. Due to the increased awareness about drug abuse (resulting in lower use as reported by Substance Abuse and Mental Health Service Administration³¹), the use of drugs such as cocaine is becoming an important patient screening criterion in any clinical study.

The recent discovery of the structure of the enzyme CYP3A4 has led to active research on it³². It was found that “*Cytochrome P450 3A4*” (in combination with the bioreductive drug, AQ4N) enhances the anti-tumor effects of radiation and chemotherapy drugs³³. “*Crotoxin*” is the venom of a South American rattle snake which has been found to have anti-tumor actions³⁴. This has led to its increased use in several cancer trials, such as the one discussed in Cura et al³⁵. “*Catechin*”, another concept which shows an increasing trend, is a constituent in green tea and has proved to be a cancer prevention agent³⁶. There have been several clinical trials to validate this property³⁷.

Further, pain alleviation is an important aspect of cancer treatment and therapy, and many clinical techniques exist for this. Recent years has seen a surge in the use of interventional³⁸ and cognitive therapies³⁹ for pain management and relief for cancer patients, which could suggest the increase in the incidence of these two types of therapies in cancer trial eligibility criteria text, as seen in Table 1. The use of barrier contraceptive methods as a form of birth control, as compared to oral contraceptive methods, has increased over the last few years⁴⁰, which may confirm the trend observed in the increasing incidence of the concept “*Contraception Barrier*”.

The concepts that demonstrated the largest decreasing trends across most cancer types can be associated with certain research results for these chemicals. We give some possible explanations for these trends observed in Table 1. Although “*creatinine*” has been found to be strongly correlated with cancer detection, a recent study has shown that serum creatinine is being replaced by biomarkers such as serum cystatin C for the early detection of renal impairment in cancer patients⁴¹. This may explain the decrease in the use of “*creatinine*” as a concept for cancer trial eligibility criteria. Another possible reason is that “*creatinine clearance*” is considered superior to “*creatinine*” for measuring kidney function therefore the latter has been gradually replaced by superior alternatives such as “*creatinine clearance*”.

The radiosensitizer “*mitomycin*” has also been shown to be rarely used for certain types of cancer because of certain adverse effects that were observed⁴², which may explain its decreasing incidence in cancer trial eligibility criteria. Some other concepts which showed the largest downward trends across all cancer types were “*bilirubin*”, “*anti-emetics*” and “*hormone receptor*”. The chemical “*hormone receptor*” shows a decreasing trend, while the chemical “*inhibitor*” shows an increasing incidence of use, in an almost equal percentage of cancer disorders. One interesting observation is that many of the concepts that show decreasing trends are related to chemotherapy-induced effects. For example, “*creatinine*” is used for renal failure detection caused by chemotherapy, “*mitomycins*” are popular as chemotherapeutic agents, “*bilirubin*” is often studied as an indicator of liver failure as a result of chemotherapy, and “*anti-emetic*” drugs are prescribed to prevent emesis caused by chemotherapy.

All procedure-based concepts which showed the strongest declines in their use relate to cancer therapies with adverse side effects or risks. “*Chemotherapy regimens*” have long been used as a cancer therapy but recent trends show increase in the number of intermediate risk regimens and usage in early stage cancer⁴³. Combining with radiotherapy (“*combined modality therapy*”) amplifies the side effects but has not shown additional benefit⁴⁴. Several sources such as⁴⁵ and⁴⁶ have warned about the side effects of “*hormone therapy*” and “*biological response modifier therapy*” which might require hospital stays. “*Pharmacotherapeutics*” are associated with weakening of the immune system⁴⁷.

Limitations

Although this study represents a vital step in eligibility criteria trend analysis, there is still a large scope for extending this work in future studies. Firstly, clinical concepts were identified in this study using the UMLS ontology through a dictionary matching approach for named entity recognition (NER). Although this approach is straightforward in its implementation, it suffers from various limitations such as false positives caused by the ambiguity of names and false negatives created by differences in spelling and synonymous concepts⁴⁸. Another disadvantage of the dictionary-based approach is that it is based on the assumption of completeness of the existing UMLS dictionary, since this approach can only recognize concepts listed in UMLS. However, dictionaries are often not complete or comprehensive as they may not contain uncatalogued and emerging concepts⁴⁹. A future work would involve the use of machine learning approaches to create more comprehensive and intelligent algorithms for NER and using these methods for concept recognition, and in turn, trend analysis of more accurately extracted concepts.

Secondly, concept recognition in this study was based purely on the appearances of the concepts in the clinical texts associated with cancer trials. Hence, there was no distinction between concepts appearing in the inclusion and exclusion criteria for clinical trials. Consideration of this distinction and application of sophisticated negation detection techniques to review concepts that are present only in either the inclusion criteria or the exclusion criteria might represent a more thorough approach to study the trend analysis of the concepts in clinical trials.

Furthermore, this study focused only on the concepts belonging to the *Chemicals and Drugs*, *Procedures*, *Observations*, and *Disorders and Conditions* semantic groups. However, clinical trial texts may include eligibility criteria from other semantic groups. Including concepts from these semantic groups will lead to a more holistic study of trend analysis, with direct applications to all domains of cancer trial design. Finally, this study can be extended to address trials from many domains of medical research, other than just cancer trials that this paper focused on.

Conclusions

This study demonstrates some interesting trends in the use of certain eligibility criteria for screening patients into cancer clinical trials. Possible medical evidences related to the observed increases, decreases and anomalies in the use of cancer trial concepts were discussed. In the future, validation of these findings by clinicians and further interpretation would be a beneficial addendum. The results from this study may translate well to future clinical studies so that clinicians and medical researchers can utilize this prior knowledge of concept trends to come up with better designs for clinical trials. This study can also be extended in the future to include concepts from other semantic groups for cancer trials, to study inclusion and exclusion criteria separately, and also expand to fields other than cancer disorders. This could enable a more holistic trend analysis of concepts, and ultimately, to better aid clinical research design. It would be of interest to see if the time of incidence of discovery of the various medical prognoses and interventional measures mentioned in the previous section correlates with start time of the rising trend. This would show us how long it takes for discoveries to be translated into clinical research practice and their wide adoption.

Acknowledgments

We would like to thank Tian Kang for useful discussion of the methodology of this study. This study is sponsored by National Library of Medicine Grant **R01LM009886** (PI: Weng) and National Center for Advancing Translational Sciences **UL1TR000040** (PI: Ginsberg). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

1. Willan AR, Pinto EM. The value of information and optimal clinical trial design. *Stat Med*. 2005;24(12):1791-1806. doi:10.1002/sim.2069.
2. Collins JF, Williford WO, Weiss DG, Bingham SF, Klett CJ. Planning patient recruitment: fantasy and reality. *Stat Med*. 1984;3(4):435-443. doi:10.1002/sim.4780030425.
3. Weng C, Embi PJ. Informatics Approaches to Participant Recruitment. *Heal Informatics*. 2012:81-93. doi:10.1007/978-1-84882-448-5.
4. Weng C, Tu SW, Sim I, Richesson R. Formal representation of eligibility criteria: A literature review. *J Biomed Inform*. 2010;43(3):451-467. doi:10.1016/j.jbi.2009.12.004.
5. Niland J, Eoyang G, Grama L, et al. ASPIRE: Agreement on Standardized Protocol Inclusion Requirements for Eligibility. 2008. <http://hssp-cohort.wikispaces.com/file/view/ASPIRE+CDISC+Intrachange+July+10+2007+Final.ppt> accessed February 2009.

6. Weng C, Wu X, Luo Z, Boland MR, Theodoratos D, Johnson SB. EliXR: an approach to eligibility criteria extraction and representation. *J Am Med Informatics Assoc.* 2011;18(Suppl 1):i116-i124. doi:10.1136/amiajnl-2011-000321.
7. Luo Z, Johnson SB, Weng C. Semi-Automatically Inducing Semantic Classes of Clinical Research Eligibility Criteria Using UMLS and Hierarchical Clustering. In: *Proc of AMIA 2010 Fall Symp.*; 2010:487-491.
8. Huang KE, Milliron B-J, Davis S A., Feldman SR. Surge in US Outpatient Vitamin D Deficiency Diagnoses: National Ambulatory Medical Care Survey Analysis. *South Med J.* 2014;107(4):214-217. doi:10.1097/SMJ.0000000000000085.
9. Carle AC. Fitting multilevel models in complex survey data with design weights: Recommendations. *BMC Med Res Methodol.* 2009;9:49. doi:10.1186/1471-2288-9-49.
10. Bieler GS, Brown GG, Williams RL, Brogan DJ. Estimating model-adjusted risks, risk differences, and risk ratios from complex survey data. *Am J Epidemiol.* 2010;171(5):618-623. doi:10.1093/aje/kwp440.
11. Rosenberg D. Trend analysis and interpretation. Key concepts and methods for maternal and child health professionals. 1997:39.
12. Weng C, Yaman A, Lin K, He Z. Trend and Network Analysis of Common Eligibility Features for Cancer Trials in ClinicalTrials.gov. In: *Smart Health Lecture Notes in Computer Science.* Vol 8549. Springer; 2014:130-141.
13. ClinicalTrials.gov. <https://www.clinicaltrials.gov/>. Accessed May 10, 2015.
14. Levy-Fix G, Yaman A, Weng C. Structuring Clinical Trial Eligibility Criteria with the Common Data Model. In: *Proceedings of 2015 AMIA Joint Summits for Translational Science.* San Francisco; 2015:194-198.
15. Miotto R, Weng C. Unsupervised mining of frequent tags for clinical eligibility text indexing. *J Biomed Inform.* 2013;46(6):1145-1151. doi:10.1016/j.jbi.2013.08.012.
16. UMLS Terminology Services. <https://uts.nlm.nih.gov/home.html>. Accessed June 19, 2015.
17. Mann H. Nonparametric tests against trend. *Econometrica.* 1945;12:245-249. doi:10.2307/1907187.
18. Kendall MG, Gobbons JD. *Rank Correlation Methods.*; 1990.
19. Gilbert RO. *Statistical Methods for Environmental Pollution Monitoring.*; 1987.
20. Draper NR, Smith H. *Applied Regression Analysis.*; 1998. doi:10.2307/2987167.
21. Jaccard P. The distribution of the flora in the alpine zone. *New Phytol.* 1912;11(2):37-50. doi:10.1111/j.1469-8137.1912.tb05611.x.
22. Treatment of multiple myeloma with massive doses of adrenal cortex hormones. <http://www.cancer.org/cancer/multiplemyeloma/detailedguide/multiple-myeloma-treating-chemotherapy>.
23. Yamasaki K, Ito S, Nomura S, Uetani T, Horikoshi N. Treatment of multiple myeloma with massive doses of adrenal cortex hormones. *Saishin Igaku.* 1967;22(7):1573-1585.
24. Nohara N, Tsukinoki K, Toda M. the Problem of Side-Effects in Adrenal Cortex Hormone Treatment of Skin Diseases. *Nisshin Igaku Jpn J Med Prog.* 1963;50:225-243.
25. Kyle R a, Gertz M a, Witzig TE, et al. Review of 1027 patients with newly diagnosed multiple myeloma. *Mayo Clin Proc.* 2003;78(1):21-33. doi:10.4065/78.1.21.
26. Breastcancer.org - Hormone Status. http://www.breastcancer.org/symptoms/diagnosis/hormone_status.
27. Burstein HJ. The distinctive nature of HER2-positive breast cancers. *N Engl J Med.* 2005;353(16):1652-1654. doi:10.1056/NEJMp058197.
28. Gambacciani M, Levancini M. *Vaginal Erbium Laser: The Second Generation Thermotherapy for the Genitourinary Syndrome of Menopause (GSM) in Breast Cancer Survivors. A Preliminary Report of a Pilot Study.*; 2015. doi:DOI: 10.1016/j.maturitas.2015.02.105.
29. *HIV Prevention Research & Development Funding Trends. 2000–2014: Investment Priorities To Fund Innovation In An Evolving Global Health and Development Landscape;* 2015.
30. Cancer and HIV. <https://aidsinfo.nih.gov/clinical-trials/search/q/1/category/4/cancer/6/cancer-and-hiv---all-trials>.
31. SAMHSA. <http://blog.samhsa.gov/2012/03/29/u-s-sees-downward-trend-in-cocaine-use/#.VftdFixVhBc>.
32. Scott EE, Halpert JR. Structures of cytochrome P450 3A4. *Trends Biochem Sci.* 2005;30(1):5-7. doi:10.1016/j.tibs.2004.11.004.
33. McCarthy HO, Yakkundi A, McErlane V, et al. Bioreductive GDEPT using cytochrome P450 3A4 in combination with AQ4N. *Cancer Gene Ther.* 2003;10(1):40-48. doi:10.1038/sj.cgt.7700522.
34. Sampaio SC, Hyslop S, Fontes MRM, et al. Crotoxin: Novel activities for a classic β -neurotoxin. *Toxicol.* 2010;55(6):1045-1060. doi:10.1016/j.toxicol.2010.01.011.
35. Cura JE, Blanzaco DP, Brisson C, et al. Phase I and pharmacokinetics study of crotoxin (cytotoxic PLA2, NSC-624244) in patients with advanced cancer. *Clin Cancer Res.* 2002;8(4):1033-1041.

36. Suganuma M, Saha A, Fujiki H. New cancer treatment strategy using combination of green tea catechins and anticancer drugs. *Cancer Sci.* 2011;102(2):317-323. doi:10.1111/j.1349-7006.2010.01805.x.
37. Kumar NB, Pow-Sang J, Egan KM, et al. Randomized, Placebo-Controlled Trial of Green Tea Catechins for Prostate Cancer Prevention. *Cancer Prev Res.* 2015.
38. Sloan P a. The evolving role of interventional pain management in oncology. *J Support Oncol.* 2(6):491-500, 503.
39. McCracken L, Turk D. Behavioral and Cognitive–Behavioral Treatment for Chronic Pain: Outcome, Predictors of Outcome, and Treatment Process. *Spine (Phila Pa 1976).* 2002;27(22):2564-2573 ST - Behavioral and Cognitive - Behaviora.
40. Finer LB, Jerman J, Kavanaugh ML. Changes in use of long-acting contraceptive methods in the United States, 2007-2009. *Fertil Steril.* 2012;98(4):893-897. doi:10.1016/j.fertnstert.2012.06.027.
41. Štabuc B, Vrhovec L, Štabuc-Šilih M, Cizej TE. Improved prediction of decreased creatinine clearance by serum cystatin C: use in cancer patients before and during chemotherapy. *Clin Chem.* 2000;46(2):193-197.
42. Seiwert TY, Salama JK, Vokes EE. The chemoradiation paradigm in head and neck cancer. *Nat Clin Pract Oncol.* 2007;4(3):156-171. doi:10.1038/nponc0750.
43. Neutropenic Risk. <http://www.neutropeniarisk.com/neutropenic-cascade/chemotherapy-trends/>.
44. Murphy SB, Hustu HO. A randomized trial of combined modality therapy of childhood non-Hodgkin's lymphoma. *Cancer.* 1980;45(4):630-637.
45. Cancer Research UK. <http://www.cancerresearchuk.org/about-cancer/cancers-in-general/treatment/hormone/general-side-effects-of-hormone-therapy>.
46. MedicineNet. <http://www.medicinenet.com/script/main/art.asp?articlekey=2464>.
47. Dembic Z. Pharmaco-therapeutic challenges in cancer biology with focus on the immune-system related risk factors. *Curr Pharm Des.* 2014;20(42):6652-6659.
48. Zhong H, Hu X. Disease Named Entity Recognition by Machine Learning Using Semantic Type of Metathesaurus. *Int J Mach Learn Comput.* 2013;3(6):494-498. doi:10.7763/IJMLC.2013.V3.367.
49. Shatkay H, Craven M. Mining the Biomedical Literature. 2012.

Appendix -1

The cancer types included in this study are:

acute lymphoblastic leukemia, acute myeloid leukemia, adrenocortical carcinoma, aids-related cancer, anal cancer, appendix cancer, astrocytomas, atypical teratoid rhabdoid tumor, bile duct cancer, bladder cancer, bone cancer, brain cancer, brain stem glioma, breast cancer, carcinoma of unknown primary, cervical cancer, childhood cancer, chronic lymphocytic leukemia, chronic myelogenous leukemia, chronic myeloproliferative disorders, colon cancer, craniopharyngioma, cutaneous t-cell lymphoma, endometrial cancer, ependymoma, epithelial ovarian cancer, esophageal cancer, ewing sarcoma, gallbladder cancer, gastrointestinal cancer, gastrointestinal stromal tumors, gestational trophoblastic disease, hairy cell leukemia, head and neck cancer, hodgkin lymphoma, hypopharyngeal cancer, intraocular melanoma, kaposi sarcoma, kidney cancer, laryngeal cancer, leukemia, lip and oral cavity cancer, liver cancer, lung cancer, lymphoma, malignant fibrous histiocytoma, melanoma, mesothelioma malignant, metastatic squamous neck cancer, mouth cancer, multiple myeloma, myelodysplastic syndromes, myeloproliferative disorders, nasal cavity and paranasal sinus cancer, nasopharyngeal cancer, neuroblastoma, non-hodgkin lymphoma, nonmelanoma skin cancer, non-small cell lung cancer, oral cancer, oropharyngeal cancer, osteosarcoma, ovarian cancer, pancreatic cancer, parathyroid cancer, penis cancer, pharyngeal cancer, pheochromocytoma, prostate cancer, rectal cancer, retinoblastoma, rhabdomyosarcoma, salivary gland cancer, skin cancer, small cell lung cancer, small intestine cancer, soft tissue sarcoma, stomach cancer, testicular cancer, throat cancer, thyroid cancer, transitional cell cancer, urethral cancer, uterine sarcoma, vaginal cancer, vulvar cancer, waldenstrom macroglobulinemia, wilms tumor, women cancers.