TEXT MINING APPLICATIONS IN LEUKEMIA

A Dissertation
Presented to
The Academic Faculty

by

Armon Varmeziar

In Partial Fulfillment
of the Requirements for the Degree
Master of Science in Biomedical Engineering in the
Wallace H. Coulter Department of Biomedical Engineering

Georgia Institute of Technology & Emory University
May 2022

TEXT MINING APPLICATIONS IN LEUKEMIA

Approved by:

Dr. Cassie S. Mitchell, Advisor
Department of Biomedical Engineering
Georgia Institute of Technology & Emory University

Dr. Wilbur Lam
Department of Biomedical Engineering
Georgia Institute of Technology & Emory University

Dr. Vamsi Kota
Department of Hematology & Oncology
Medical College of Georgia & Augusta University

Date Approved: April 28th, 2022

Dedicated to my wonderful family

For their unconditional love and support

ACKNOWLEDGEMENTS

First and foremost, I would like to express my sincere gratitude and appreciation to my advisor, Dr. Cassie Mitchell, for the constant guidance, suggestions, and encouragement she provided me throughout my research and my graduate studies at Georgia Tech. She has inspired me to always push through the face of adversity and reminded to "Never, never, never give up" no matter how seemingly arduous a challenge may be. I will never forget how welcoming she was upon first meeting with her, and I will forever be grateful for her patience and understanding.

I would also like to thank Dr. Hanjoong Jo and Dr. Eva Dyer from the Department of Biomedical Engineering as well as Dr. Polo Chau and Dr. Mahdi Roozbahani from the College of Computing for their wonderful academic instruction in the classroom.

Thank you to Stephen Allegri and Sarah Bi for your friendship, help, and support throughout the courses that we took together. Thank you for all the encouragement in both the good times and the challenging times in our master's degree journey together. I am so fortunate to have met you, and I cannot wait to begin our professional careers together.

Finally, I want to give a special thanks to my family. Without their love and support none of this could have been possible: to my amazing mother and father, Mojgan and Behrooz, who inspired me to become an engineer like them since I was a young kid; to my talented sister, Sofia, for constantly raising the bar I worked so hard to set; and to our pup, Louie, for the endless happiness and joy that he selflessly shares with all of us.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS AND ABBREVIATIONS

AAPP	Amino Acid Peptide or Protein
AE	Adverse Event
BACS	Biologically Active Substance
CML	Chronic Myeloid Leukemia
DSYN	Diseases or Syndromes
GNGM	Gene or Genome
PHSU	Pharmacologic Substance
PTH	Parathyroid Hormone
TOPP	Therapeutic or Preventive Procedure
TKI	Tyrosine Kinase Inhibitor
TFR	Treatment Free Remission

SUMMARY

Text mining was utilized to improve personalized tyrosine kinase inhibitor (TKI) therapy selection for chronic myeloid leukemia (CML) patients and identify long-term adverse events for which regular surveillance is recommended. TKI therapy remains the gold standard of current treatment and has greatly improved patient survival and outcome. However, the adverse events and side effects associated with TKIs negatively impact patient quality of life and therapy compliance. Because many patients must take TKIs indefinitely, understanding the long-term adverse event profile is critical for optimal clinical management. Given many TKI are relatively new, the long-term adverse event profiles are not yet fully understood. Text mining methods such as bag-of-words, k-means clustering, cross-domain literature-based discovery, and literature meta-analysis were conducted to predict and quantify the most relevant adverse events. Three tiers of predicted adverse events were quantified through statistical analysis and machine learning. Tier 1 adverse events (top 1% of predictions) constituted hematological, glucose-related, iron-related, cardiovascular, and thyroid-related adverse events from TKI therapy. Tier 2 adverse events (top 5% of predictions) constituted kidney and inflammation related adverse events. Tier 3 adverse events (top 10% of predictions) constituted gastrointestinal neuromuscular, or other adverse events like secondary malignancies. Based on study results, tier 1 conditions were recommended for regular surveillance; tier 2 for periodic surveillance with periodicity based on patient history; tier 3 for surveillance as a function of symptoms present. These insights provide a more proactive approach for optimal CML management and personalized TKI selection.

CHAPTER 1. BACKGROUND

1.1 Text Mining & Personalized Medicine

Personalized medicine is the form of medicine that uses specific information about one's unique characteristics to make improved treatment decisions [1]. There is a plethora of literature to help distinguish important details of various diseases towards personalized medicine. For example, a simple PubMed query for common diseases like "diabetes mellitus" or "cardiovascular disease" returns hundreds of thousands of journal articles. Although there are plenty of articles to help improve one's understanding of various concepts, this abundance may cause difficulty in making important connections within the literature, which are relevant to optimal personalized patient management. Due to its inherently specialized and narrow focus, biomedical research unintentionally obstructs the establishment of relationships between the findings of different biomedical sub-fields [2]. In addition to literature, there are other data sources that provide valuable insight for improving outcomes in personalized medicine. For instance, patient medical records, standardized clinical data, administrative data (from hospitals, insurance companies, and pharmacies), and self-reported comments from individual patients all have the potential to be useful inputs. With such an enormous amount of available information, staying up to date with all new discoveries and theories in biomedical research has never been so demanding. Within the treasure trove of biomedical literature is a wealth of important information that has yet to be encapsulated into actionable insight [3].

Text mining is the discovery and extraction of interesting, non-trivial knowledge from free or unstructured text [4]. Similar to extracting insight from structured sources like databases via data mining, text mining deals with the processing of inherently unstructured and vague textual data. The applications of text mining have recently garnered interest in the biomedical domain to help alleviate the challenges associated with the profuse amount of available textual information. The overarching goal of biomedical text mining is to allow researchers to identify needed information more efficiently, uncover relationships obscured by the sheer volume of available information, and shift the burden of information overload from the human researcher to the computer. Text mining provides valuable and actionable insight for biomedical researcher and clinicians by applying algorithmic, statistical and data management methods to the vast amount of published literature and free text fields of biomedical databases [2].

1.2 Examples of Text Mining Methodology

The general framework of text mining consists of two components: text refining and knowledge distillation [5]. Text refining deals with the transformation of the unstructured text into a more organized and suitable intermediate form. Knowledge distillation then processes this intermediate form to deduce patterns or knowledge. There are three main text mining approaches that are prevalent in the biomedical domain: co-occurrence-based approaches, rule/knowledge-based approaches, and statistical/machine-learning-based approaches [6].

Co-occurrence-based approaches evaluate the frequency of words that occur in a unit of text, such as a document or an abstract. The central idea of this approach is that if there are multiple concepts that frequently appear in a unit of text, this suggests there is some relationship between these co-occurring concepts. Although more error-prone than knowledge-based and machine-learning-based approaches, co-occurrence-based approaches are often used as a baseline due to their simplicity and low computational expense [7]. One of the most popular co-occurrence-based approaches is the bag-of-words model. This model creates a vector representation of the document where each element denotes the normalized number of occurrences of a basis term in the document [8]

Knowledge-based approaches make use of a domain's pre-existing knowledge. There is a wide variety of knowledge-based approaches that range from abstract to concrete. On the one hand, linguistic and semantic analyses of the text can be used to substantiate a wide range of hypothetical relationships between specific concepts. On the other hand, there can be a hard-coded set of rules -- for instance, (gene) *plays a role in* (disease), or (treatment) *is associated with* (disease) -- to investigate concepts of domain researcher interest [6]. A convenient way to visualize such sets of rules is via a knowledge graph. In a knowledge graph, the nodes represent biomedical concepts (e.g., particular gene or disease) and the edges represent the predication between the nodes (e.g., treats, inhibits).

Machine learning-based systems construct classifiers that can be applied to various parts of textual data, from labeling parts of speech to classifying the findings of a document. These systems are used in conjunction with knowledge-based approaches to quantitatively shape the researcher's assertions. A particular instance of the combination of these two approaches is link prediction. Link prediction refers to the discovery of associations between nodes that are not directly connected in a network, such as a knowledge graph [9]. It does so by estimating the probability that an edge exists between two nodes based on the observation of existing edges. A common example of link prediction in practice is when an online social network such as Facebook or Twitter suggests friends based on one's profile.

1.3 Leukemia & Clinical Context

Mayo Clinic defines leukemia as the cancer of the body's blood-forming tissues, including the bone marrow and the lymphatic system. There are four main types of leukemia: acute myeloid leukemia, acute lymphocytic leukemia, chronic myeloid leukemia, chronic lymphocytic leukemia. Each type can be distinguished by the type of white blood cell affected (lymphoid vs. myeloid) as well as the characteristics of the disease (acute vs. chronic) [10]. In acute leukemias, the cancer develops from immature cells known as precursor cells or stem cells. In chronic leukemias however, the leukemia cells come from mature, abnormal cells. Myeloid leukemias, also known as myelogenous leukemias, develops from myeloid cells, and lymphocytic leukemias develop from cells called lymphoblasts or lymphocytes in the blood marrow. Although all four types are classified as leukemia, each type is unique is its etiology and treatment.

In this work, it is hypothesized that text mining can identify and predict long-term adverse event profiles for specific leukemia treatments. For example, in chronic myeloid leukemia (CML), there are relatively few clinical studies available that have examined long-term adverse events with tyrosine kinase inhibitor (TKI) therapy. However, text mining cross-domain literature relationships can identify adverse events that should be monitored or considered when matching a specific TKI option to a patient’s medical history and lifestyle. Thus, in this case study, text mining provides actionable insight for improving long-term surveillance of CML patients on TKI therapy.

CHAPTER 2. INTRODUCTION

Preface: *The work presented in Chapters 2-6 comprise a journal article in preparation that is based on the presented work. I, Armon Varmeziar, will be co-first author of the corresponding journal article. The remaining full author list has yet to be finalized at the time of this writing. Nonetheless, I want to acknowledge here project team members that directly assisted me, including Nidhi Mehra (the other co-first author), Rachel Fisher, Olivia Kronick, Xinyu Chen, and Dr. Cassie S. Mitchell.*

Chronic myeloid leukemia (CML) is an uncommon type of cancer in the bone marrow that targets early myeloid cells. It results in an increase of impartially mature cells that offsets normal myeloid cells. Comprising approximately 15\% of all leukemias, CML is cytogenetically characterized by the presence of the Philadelphia (Ph+) chromosome. Ph+ results from the fusion of the Abelson (ABL) tyrosine kinase gene of chromosome 9 and the breakpoint cluster (BCR) gene at chromosome 22. Chronic myeloid leukemia has an estimated 9,110 new cases and 1,220 new deaths in 2021 [11]. While CML is most frequently diagnosed among adults aged 65-74, more cases of pediatric, adolescent, or young adult CML has increased in recent years [12].

The advent of BCR-ABL-targeting tyrosine kinase inhibitors (TKI) as a therapeutic for CML has substantially improved patient outcomes [13, 14]. Currently, five TKIs are widely approved for CML treatment: imatinib, dasatinib, nilotinib, bosutinib, and ponotanib. Nowadays, patients with CML boast a life expectancy that is similar to that of the general population. As a result of this improved life expectancy, the goals of CML treatment are beginning to shift from short-term disease management towards long-term expectations and quality of life [15]. Namely, treatments aim to achieve deep molecular response, where blood levels of BCR ABL transcript measured via PCR are less than or equal to 0.01% on the International Scale (IS), which is equivalent to a 4-log reduction in BCR ABL transcript with ≤ 1:10,000 of measured transcripts still possessing the mutation [16]. Obtaining and maintaining a deep molecular response with ongoing TKI therapy makes CML more analogous to treating a chronic condition.

More recently, the possibility of treatment-free remission (TFR) has been a goal. TFR is where patients with a deep molecular response cease to take TKI therapy if they meet specific criteria. In the United States, the most successful guidelines for TFR patient selection include those CML patients who meet all the following criteria: patients who did not have a high Sokal score at initial diagnosis, never progressed past chronic phase, did not possess TKI resistant mutations, have been on TKI therapy for 8 years, and have had a deep molecular response for more than 2 years at the time of TFR trial initiation [17]. Several trials have commenced to assess relapse rates in patients who undergo TFR, including but not limited to the French STIM trial [18], the TWISTER study in Australia [19], the EURO-SKI study [20], the ENESTfreedom study [21], ENESTop study [22], and the DASFREE study [23].

Unfortunately, not all patients many patients who attempt TFR relapse, with the majority relapsing within 6 months; based on the many TFR studies somewhere between 40-60% of patients ultimately relapse and a required to resume TKI therapy [24]. Moreover, TKIs are not considered safe to use during pregnancy; thus, CML female patients of child-bearing age must make the difficult decision to stop therapy and risk relapse before the birth of their child [25]. Thus, based on current TKI therapies and corresponding clinical evidence, the majority of CML patients at this time require TKI therapy indefinitely. Obtaining and maintaining a deep molecular response with ongoing TKI therapy makes CML more analogous to treating a chronic condition.

While TKIs have been very effective, these powerful drugs are certainly not benign. Due to the ubiquitous nature of the drugs' targets in the body, there is a litany of adverse events (AEs) or side events associated with TKI usage [26]. Studies have shown cardiovascular, musculoskeletal, and gastrointestinal disorders associated with TKI therapy [27] [23] [28]. With such chronic utilization, these undesired effects can negatively impact patient quality of life of a patient. Poor quality of life can result in decreased patient TKI compliance, which may lead to drug resistance [29] [30] [31]. In summary, there are a number of patients, which must remain on TKI therapy for a long period of time, if not indefinitely. Thus, understanding the long-term adverse events or side effects is critical for patient management, including patient monitoring, selection of the optimal TKI given the patient's medical history, and for improving future, more targeted TKI therapies [28].

Since TKI therapy is a relatively new treatment, there are limited studies of its long-term effects. CML is also a relatively uncommon leukemia, resulting in limited clinical cohort studies on the less common AEs associated with TKI therapy. The applications of text mining and machine learning pose an innovative opportunity to circumvent the challenges associated with the lack of large-scale studies across underexplored physiology. SemNet, for instance, is a literature-based discovery tool that uses text mining to extract relationships from millions of indexed abstracts in the PubMed database [32] [33]. SemNet can extract relationships to reveal cross-domain connections in the scientific literature.

The central objective of this study was to utilize innovative text mining techniques to identify cross-domain literature relationships of physiology impacted by tyrosine kinase. The text mining techniques are performed to identify and rank adverse events relevant to TKI therapy for CML. A follow-up meta-analysis is conducted on a subset of predicted hematological AEs to quantitatively assess text mining predictions as a form of validation. An outline of this study's research process is shown in Figure 1. The benefits gained from uncovering connections are threefold: 1) knowing what long-term AEs are most probable assists in patient follow-up care by suggesting prophylactic testing to identify and treat AEs as early as possible; 2) further understanding and improving the design of TKIs so that their AEs can be attenuated or even eliminated; 3) improving personalized medicine for optimal TKI selection by considering a patient's medical history/antecedent disease, genetic predispositions, and lifestyle, as well as considering drug-specific TKI AEs, when selecting a TKI therapy.

CHAPTER 3. METHODOLOGY

The overall goal of this work was to use innovative text mining to identify and rank potential adverse events (AEs) or side effects due to long-term TKI therapy in CML patients. Methods consisted of three major four analyses: 1) Preliminary cluster analysis to connect adverse events to specific TKI drug types based on text mining of CML-specific literature; 2) Literature based discovery simulations with SemNet to examine highly ranked cross-domain relationships connected to tyrosine kinase physiology using a knowledge graph consisting of over 30+ million PubMed articles; 3) Quantitative meta-analysis with real CML TKI clinical data to validate a subset of text mined predictions for which there was sufficient existing quantitative clinical data.

3.1 Cluster Analysis of TKI-Specific Adverse Events

Peer-reviewed CML journal articles in PubMed were extracted to assess the statistical associations between highly occurring adverse events and specific TKI drugs (imatinib, nilotinib, dasatinib, bosutinib, ponotinib) based on the presence of specific terms in a document (a specific journal article) and the overall corpus (all included CML articles). After conducting a PubMed search using the query ‘CML AND side effects OR adverse events’, 2575 abstracts were retrieved. The abstracts were parsed, common stop words (a, an, the, but, etc.) and punctuation removed, tokenized, lemmatized, and cleaned in Python using the Natural Language Toolkit library. A bag-of-words technique [34] was used to vectorize abstracts and transform them into a term frequency-inverse document frequency (TF-IDF) matrix. TF-IDF calculates how relevant a word is in a corpus. The meaning of TF-IDF increases proportionally to the number of times in the text a word appears but is compensated by the word frequency in the corpus. The TF-IDF matrix for thi study had dimensions of (2575, 1621). Thus, 1621 terms occurring in between 1% and 85% of the abstracts were included for analysis. Note that only abstracts were processed for inclusion, rather than the entire journal article, in order to reduce bias from repetitive or less relevant text in sections of an article not specific to that study's results.

K-means clustering [35] was used as to showcase the distinct adverse event and side effect profiles associated with specific TKIs used to treat CML. The TF-IDF matrix was used as the input for k-means clustering, which was implemented using Scikit-learn in Python. The data was reduced to 50 components using TruncatedSVD [36]. Finally, t-Distributed Stochastic Neighbor Embedding (t-SNE) was used to further reduce the dimensionality for 2D visualization [36].

After examining the top 200 terms with the highest term frequency in-verse document frequency (TF-IDF) scores in each cluster, clusters 4, 7, 8, 10, and 16 were identified to be relevant to specific TKIs and their associated undesired effects. A list of n terms related to AEs and toxicity was identified from visually inspecting the aforementioned terms in each of the five clusters. The highest frequency of occurrence for each term in the clusters was noted to identify specific adverse event profiles associated with each of the TKIs. For each of these clusters, fold differences were calculated for each term to compare its frequency within the cluster to its frequency within the corpus of abstracts. Pairwise Chi-Square tests were used to determine whether each of the five clusters identified were significantly different from each other. A family-wise alpha of 0.05 was utilized, and a Bonferroni correction for multiple comparisons was used to lower the threshold for statistical significance to prevent a Type I error.

3.2 SemNet Simulation of Cross-Domain Conditions linked to Tyrosine Kinase

The TKI-specific clusters from the bag-of-word analysis of peer-reviewed CML journal article abstracts are helpful for assessing known drug-specific TKI AEs. However, to assess potential long-term AEs that may not have been yet measured in mentioned in CML literature, a different approach is needed. One way to assess distant interconnected relationships across domains is to utilize a knowledge graph. A knowledge graph uses extracted text relationships to create a flowchart of connections. A scientist or clinician can typically deduce connections that are a few connections away from the target or disease of interest if those connections fall within their domain specialty. However, more distant cross-domain relationships or relationships simply further from the target are much more difficult to deduce. Machine learning algorithms can be used to rank relationships within a large knowledge graph to one or more targets of interest. Here, we utilized the SemNet software for this purpose [32]. This study used the original SemNet software given work was completed before the release of SemNet 2 [33]. Two types of knowledge graph text mining were performed: the standard unsupervised rank aggregation algorithm of SemNet to assess published relationships in the knowledge graph [32] and a link prediction module to assess predicted unpublished relationships [37]. Relationships predicted as important suggest possible conditions that should be monitored as part of long-term AEs with TKI therapy. Figure 1 illustrates the knowledge graph pipeline utilized in this work.

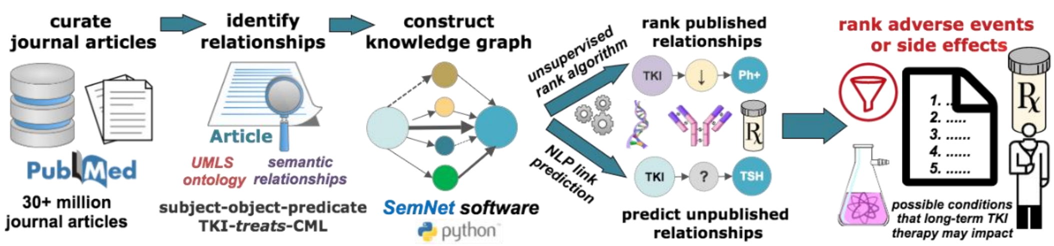

Figure 1. Overview of cross-domain text mining relationships extracted from peer-reviewed articles in PubMed and connected into a knowledge graph. Subject-object-predicate triples are extracted from the text of journal article, which relate two or more nodes or "concepts" of interest. The knowledge graph ontology is specified by the Unified Medical Language System (UMLS). Example node types include pharmacological substance, gene or genome, neoplastic process, organic chemical, cell component and amino acid/ protein. Predication types include "interacts", "associated", "part of", "treats", "inhibits", "stimulates", "causes", and "is a". The open source Python-based biomedical literature based discovery software, SemNet [32], was used to identify rank relationships between possible adverse events or side effects due to tyrosine kinase inhibition. A standard SemNet simulation was used to rank published relationships connecting undesired conditions to tyrosine kinase. A natural language processing (NLP) link prediction algorithm [37] was used in combination with SemNet to identify relationships that are not explicitly in the knowledge graph. Rather, link prediction uses existing published relationships in the knowledge graph to predict novel relationships, which are either unpublished or, minimally, do not exist in the graph. Results are then synthesized and ranked to suggest adverse events or side effects that should be carefully and prophylactically monitored as part of a long-term TKI therapy

3.2.1 *Initial examination of highest rank nodes to CML and tyrosine kinase*

SemNet is a semantic inference network and biomedical concept graph that uses unsupervised ranking algorithms to reveal hidden connections from PubMed's nearly 30 million indexed abstracts [32][33]. A user can run a simulation for a particular node of interest, which is specified as the "target node", for which SemNet will search the knowledge graph and return a list of "source nodes" that are connected to the target node(s). Each connection has both a predication type that represents the association between the two nodes, as well as a magnitude that represents the prevalence of the connection within the PubMed literature. Each node has a particular node type to help the user methodically filter for desired source nodes. For instance, if a user wanted to examine all the published genetic factors involved in chronic myeloid leukemia, they could run a simulation with a target node of "chronic myeloid leukemia" and filter the resulting source nodes for the "gene or genome (GNGM)" node type.

SemNet Simulations were utilized to examine diseases and therapeutic substances that might be associated with TKI therapy for CML. The simulation parameters were set to filter for two particular node types in the knowledge graph: Diseases or Syndromes (DSYN) and Therapeutic or Preventive Procedure (TOPP). These parameters were specified to focus on any immediate diseases or therapeutic substances within one connection of the specified target node: CML. After the simulation was ran, a comprehensive list of 308 treatments and diseases related to CML were retrieved. The list was subsequently analyzed to look for existing trends or novel relationships derived from the current literature. The results of this simulation identified conditions of interest, including node types from Amino Acid Peptide or Protein (AAPP), Biologically Active Substance (BACS), Disease or Syndrome (DSYN), Gene or Genome (GNGM), and Pharmacologic Substance (PHSU). HeteSim scores were utilized to rank conditions that were of most importance to CML and tyrosine kinase.

Largely due to the inclusion of the CML node as a target, predominantly hematological conditions were returned as the highest ranked related nodes. Those especially featured included: aplastic anemia, anemia, neutropenia, pancytopenia, thrombocytopenia and the more general node, myelosuppression. These nodes were utilized as inputs into more specific link prediction cross-domain simulations to identify potentially underrepresented but very relevant conditions for which long-term TKI therapy may cause or exacerbate.

3.2.2 Link prediction of distant or less published connections to tyrosine kinase

While SemNet is able to summarize existing connections in the knowledge graph, it cannot predict a connection that does not already exist in the literature. Thus, another text mining method called link prediction was used in conjunction with SemNet to predict potentially novel relationships of interest between possible AEs and long-term CML TKI therapy. An end-to-end link prediction module originally used to identify repurposed drugs for COVID-19 [37] was adapted to predict potential AEs with long-term CML TKI therapy. The link prediction used the following format, head node-relationship-tail node. For example, tyrosine kinase inhibitor-causes-myelosuppression.

A node neighborhood analysis was performed to gain a better understanding predicted relationships and conditions of interest. A node neighborhood is defined as all nodes connected to a specific node of interest. The objectives of this analysis include: 1) the comparison of node neighborhoods between conditions of interest and; 2) to identify missing nodes and edges from these neighborhoods. The node neighborhood analysis used the 6 nodes from the initial SemNet analysis as hub nodes for making the neighborhoods: aplastic anemia, anemia, neutropenia, pancytopenia, thrombocytopenia and the more general node, myelosuppression. Entity prediction was used to generate a list of nodes with a specified relationship type to the six node hubs. The included relationships were CAUSES, PRECEDES, AUGMENTS, DISRUPTS, STIMULATES, and AFFECTS, and included the node types were AAPP, BACS, DSYN, GNGM, PHSU). The unsupervised graph algorithm, TransE [37] generated a list of nodes predicted to have a relationship with each hub node. A Neo4j query (head node-relationship-tail node) was used to construct subgraphs around each hub node. Each output of the query was accompanied by a degree normalized weight value, which represented the strength of the relationship between the head and tail node. The weight values were used to prune the number of nodes included in the neighborhood for standardization across the generated hub node neighborhoods.

The node neighborhood results were used to set the targets for additional SemNet simulations for DSYN conditions associated with tyrosine kinase. These simulations generated pairwise comparisons between mutually highly ranked nodes related to tyrosine kinase and the hub node(s). A pictorial diagram to illustrate this concept is shown in Figure 2. To ensure that the resultant nodes were maximally informative, the baseline nodes linked to "hematopoiesis" were removed as well as overly generic nodes, like "protein". To enable comparison and aggregation across hub node neighborhood simulations, the HeteSim scores were recorded and normalized to fall between 0 and 1. Approximately the top 10% of aggregated hub node neighborhood simulations were isolated for human-in-the-loop validation. The top 10% were chosen on the basis of the most novelty (using the SemNet novelty score) and relevance (using the SemNet HeteSim score). Finally, three tiers were used to classify foci of conditions that are predicted to be relevant to long-term TKI therapy with CML. Tier 1 included the top 1% of predicted aggregates, tier 2 the top 5% of predicted aggregates, and tier 3 the top 10% of predicted aggregates.

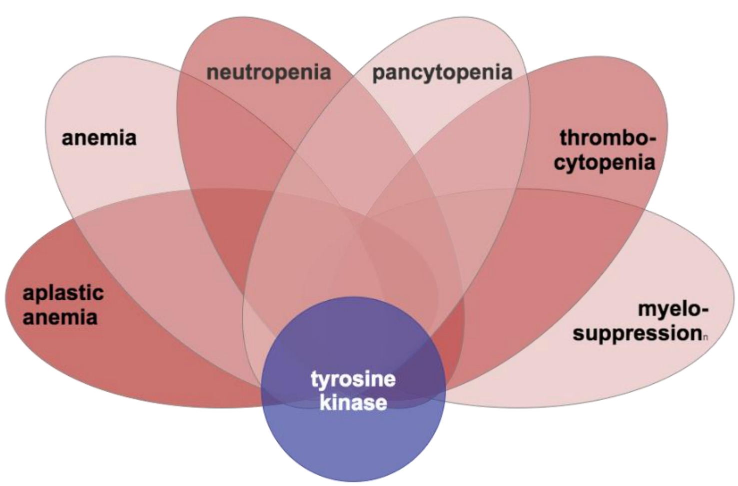

Figure 2. Conceptual overview of hub node neighborhood analysis to find more distant or less represented connections to tyrosine kinase in the knowledge graph. The 6 leaves shown represent the hub nodes that were selected using the initial SemNet simulations (aplastic anemia, anemia, neutropenia, pancytopenia, thrombocytopenia, and myelosuppression) to identify high ranking conditions tied to the CML target node. Here link prediction and SemNet simulation using the hub nodes is used to find nodes related to tyrosine kinase that could be relevant to CML patients on long-term TKI therapy.

3.2.3 Human-in-the-loop validation of potential relevance of predicted conditions

The human-in-the-loop validation protocol published in [37] was utilized to determine whether there was sufficient tangential literature evidence to support or refute a predicted relationship. Recall link prediction looks for links based on suggested similar patterns but not directly published explicit relationships. Thus, tangential evidence could include relationships from connecting domains. For example, a predicted relationship such as tyrosine kinase-affects-glucose may be found to be present in the diabetes literature but not explicitly in the CML literature. Each predicted relationship was classified by a human curator as “agrees”, “disagrees”, or “missing”. Agrees was defined as a relationship for which sufficient literature evidence was available to make the predicted link likely. Disagrees was defined as relationship for which sufficient literature evidence was available to make the predicted link unlikely. Missing was defined as relationship for which there was insufficient information to make a determination. Each relationship was reviewed by an independent quality control team and any differences in classification label resolved [37].

3.3 Statistical Meta-Analysis

The text mining analyses all concurred that hematological conditions were the most common adverse event or side effect for TKI therapy. In order to assess the prevalence and correlation of Tyrosine Kinase Inhibitor (TKI) treatment with hematological conditions, searches on Pub-Med.gov and ClinicalTrials.gov were performed in 2021 to discover relevant peer-reviewed data sources for performing a cross-cohort meta-analysis. The prevalence of each hematological condition was examined compared across specific TKI treatments and different CML disease stages. Ultimately, sufficient data was available to examine and compare the prevalence of anemia, neutropenia, or thrombocytopenia while being treated with imatinib, dasatinib, nilotinib, or busutinib for chronic phase, accelerated phase, or blast phase CML.

3.3.1 Inclusion Criteria

The key words used to search are as follows:

a. ("chronic myeloid leukemia" OR "Chronic myelogenous leukemia" OR "CML) AND ("imatinib" OR "dasatinib" OR "nilotinib" OR "bosutinib" OR "ponatinib" OR "asciminib" OR "radotinib" OR "ruxolitinib" OR "tipifarnib")

b. ("chronic myeloid leukemia" OR "Chronic myelogenous leukemia" OR "CML) AND ("Aplastic anemia" OR "anemia" OR "neutropenia" OR "thrombocytopenia" OR "myelosuppression" OR "pancytopenia").

Before curation, the number of studies found were: 6 for aplastic anemia, 16 for anemia, 27 for neutropenia, 24 for thrombocytopenia, 56 for myelosuppression, and 15 for pancytopenia.

3.3.2 Exclusion Criteria

After gathering all found literature through open access or Georgia Tech Library, studies without one or more of the following attributes were excluded: quantitative incidence data available, studies published in non-English language, studies not comprising diagnosed CML patients, CML treatments that did not solely comprise the use of an included TKI. In the case of duplicate patient populations identified in multiple peer-reviewed academic articles (such as data published with duplicated NCT numbers), the clinical data for the cohort was only recorded once to prevent analytical bias.

3.3.3 Data Curation and Quality Control

Data initially recorded from each study included the following: PMID (paper dentification), number of patients under treatment, number of patients under treatment with condition, fraction of patients with hematological adverse event, TKI type [imatinib, dasatinib, nilotinib, bosutinib, ponatinib, asciminib, radotinib, ruxolitinib, tipifarnib], type of hematological condition [aplastic anemia, anemia, neutropenia, thrombocytopenia, myelosuppression, pancytopenia], grade of condition [all grades, grade 3-4], phase of CML [chronic, accelerated, blast], and data source. Data was collected and aggregated for "all grades" of the hematological condition. The exact source of each study was traced, and data published with duplicated NCT numbers were excluded. Lastly, all data curated was independently validated by a quality control team [28] [38].

During curation, studies without relevant data relating to CML patients on TKI treatment showing hematological conditions, along with studies that were not TKI treatment only, were further excluded. At the end of the curation process, the number of unique data points collected were as follows: 3 for aplastic anemia, 50 for anemia, 55 for neutropenia, 61 for thrombocytopenia, 13 for myelosuppression, and 16 for pancytopenia. Many studies found included data for multiple hematological conditions at the same time, making the number of unique studies different from the number of unique data points. Ultimately, the number of available studies and/or the aggregate sample size of patients for myelosuppression, aplastic anemia, and pancytopenia was too small to be included in the pairwise statistical analysis. Similarly, the number of available studies and/or the aggregate sample size of patients taking ponatinib, asciminib, radotinib, ruxolitinib, tipifarnib was too small to be included in the pairwise statistical analysis.

3.4 Statistics

Two statistical techniques were used in this meta-analysis: pairwise statistical analysis and odds ratio analysis. Pairwise statistical analysis was performed to examine hematological adverse event incidence among CML patients taking TKIs. Pairwise comparisons were performed to examine potential significant differences in hematological AEs as a function of specific TKI therapy. Pairwise comparisons were also performed to examine potential significant differences in hematological AEs as a function of CML phase (chronic, accelerated, and blast phase). The pairwise test performed determined if the proportion of patients with the hematological AE was significantly different than the proportion of patients who did not have the hematological AE. The family-wise alpha was set at 0.05. However, the final p-value threshold for significance lowered using a

Bonferroni correction to account for multiple comparisons and prevent a Type I error. Additionally, to compare hematological conditions as a function of specific TKI therapy, an odds ratio (OR) with 95% confidence interval was calculated using imatinib as the control population. The objective of performing the OR test was to assess how dasatinib, nilotinib, or bosutinib impacted the odds of getting a specific hematological AE compared to imatinib. Imatinib was chosen as the control cohort because it is a first generation TKI and is still viewed as the most popular first-line standard of care worldwide.

CHAPTER 4. RESULTS

The goal of this study was to explore the unintentional adverse events or side effects with TKI therapy and, especially, long-term TKI therapy. A secondary goal was to provide an evidence-based framework to assist clinicians in the optimal selection of a specific TKI based on patient medical history and to help monitor, diagnose, and treat potential AEs. Three types of analysis are performed. First, a text mined analysis of CML TKI literature was performed to quantify know adverse events or side effects specific to a TKI therapy. Second, a cross-domain text mining analysis using a knowledge graph of nearly 30+ million PubMed articles was performed to assess which less represented or even unpublished relationships are predicted to be most relevant to long-term TKI therapy. Third, a quantitative meta-analysis was performed to quantify the prevalence of the most common hematological conditions identified in both text mining analyses as a function hematological AE type, TKI type, and CML disease stage.

4.1 K-means Clustering of Adverse Events in the CML and TKI literature

K-means clustering was performed to identify potential connections between adverse events or side effects and specific TKI drugs used for CML treatment. Figure 3 depicts the results of the k-means clustering algorithm, which was applied to 2575 CML abstracts retrieved from a Pub-Med search query for ‘CML AND side effects OR adverse events’. Of the thirty clusters generated, five were identified to be pertinent to tyrosine kinase inhibitors based on the terms with high frequency within each cluster, as discussed in the Methods section and shown in Figure 4. Cluster 4 (blue) pertains to bosutinib, cluster 7 (gold) pertains to imatinib, cluster 8 (green) pertains to ponatinib, cluster 10 (red) pertains to dasatinib, and cluster 16 (pink) pertains to nilotinib. In each of these clusters, named specific TKI occurs at a drastically higher frequency than the other highly occurring terms. To statistically analyze if each cluster was significantly different, pairwise Chi-squared tests were conducted. The family-wise alpha was set to 0.05, and a Bonferroni correction was applied to appropriately adjust the p-value threshold for significance for multiple comparisons. All five clusters were significantly different from each other, as depicted in Figure 4.

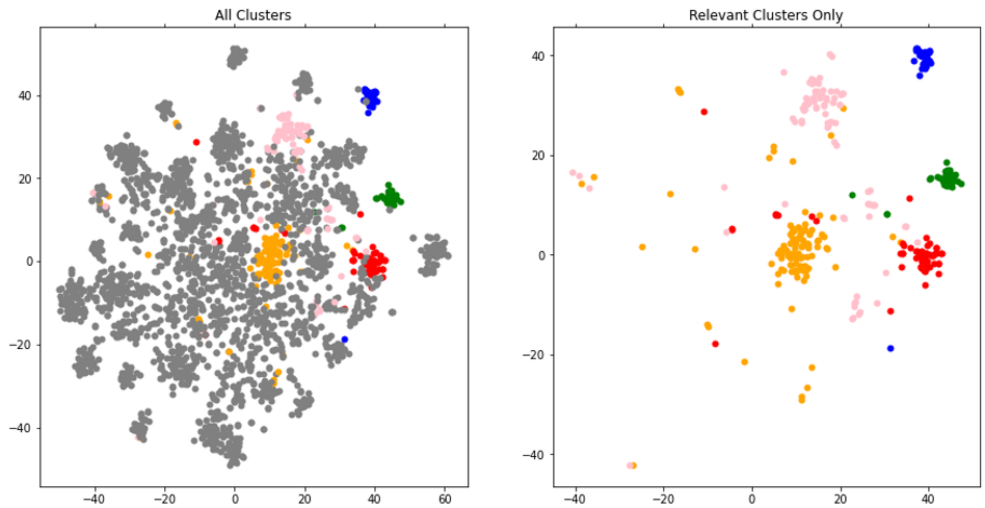

Figure 3. Left: Visual representation of the k-means clustering output when applied to PubMed abstracts retrieved for CML and side effects. Each point on the plot represents a distinct abstract. Gray dots are the abstracts unrelated to TKIs. The colored dots represent the abstracts in the clusters pertaining to TKIs. Right: Output of the k-means clustering algorithm without the unrelated abstracts.

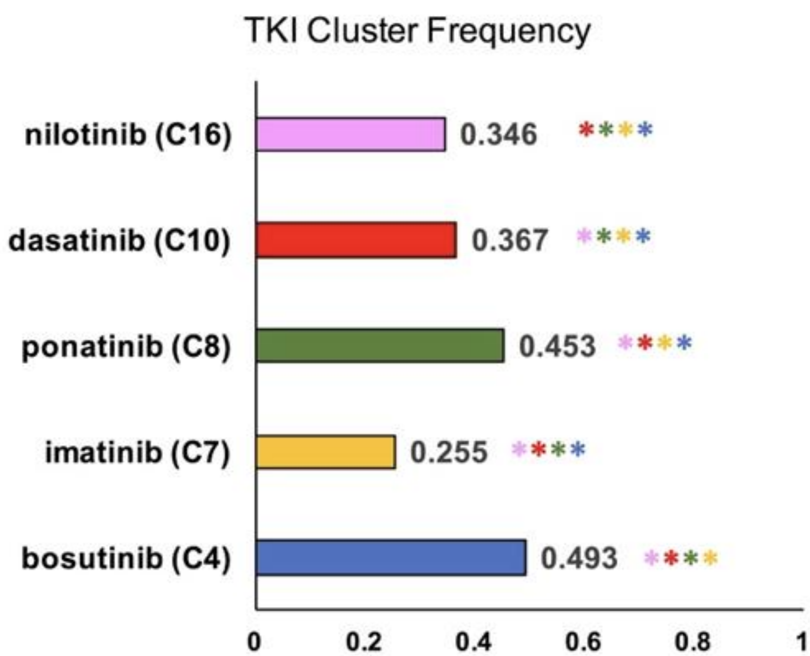

Figure 4. Bar chart mapping each TKI to its corresponding cluster. The color of each bar corresponds to the color of the cluster in the scatter plot (Figure 3). Frequency of occurrence of each TKI in its related cluster is printed above each bar. A chi-squared test was performed on clusters 4, 7, 8,

Table 1 illustrates highly occurring terms within each of the clusters identified using k-mean clustering. The frequency of TKIs and side effects or adverse events within each of the resulting clusters are shown in Table 1 Panel B and Panel C, respectively. Terms pertaining to gastrointestinal adverse events, such as “diarrhea”, “nausea”, and “vomiting” occurred most frequently in cluster 4 (bosutinib). Terms pertaining to hematological and dermatological adverse events, such as “edema”, “rash”, “myelosuppression”, “hematological”, and “lesion” occurred most frequently in cluster 7 (imatinib). Additionally, the terms “hepatitis” and “liver” also occurred most frequently in cluster 7 (imatinib). Terms such as “cardiovascular” and “vascular” occurred most frequently in cluster 8 (ponatinib); “platelet” and “arterial” also occurred at high frequencies in cluster 8. Terms such as “pulmonary”, “pleural”, “platelet”, “hypertension”, and “arterial” occurred at the highest frequency in cluster 10 (dasatinib). The only term to occur at the highest frequency in cluster 16 (nilotinib) was “diabetes”; however, terms including “cardiovascular”, “vascular”, and “arterial” also occurred at relatively higher frequencies. Finally, terms pertaining to hematological conditions appeared frequently within all five clusters, and thus were associated with all TKIs. “Myelosuppression” appeared in all clusters except cluster 16 (cluster most associated with nilotinib). The term “hematological” and “platelet” appeared within all clusters except cluster 4 (bosutinib). “Thrombocytopenia” appeared in all clusters.

Table 1. Color coding corresponds to colors of clusters in the cluster plot (Figure 3). Using abstracts retrieved from a PubMed search involving CML and side effects or adverse effects, k-means clustering was performed and clustering frequency computed. Based on the cluster frequency, the top 200 terms by term frequency - inverse document frequency (TF-IDF) scores were examined. Clusters 4, 7, 8, 10, 16 were identified to each pertain to a different tyrosine kinase inhibitor (TKI) therapy and its associated adverse events or side effects. This table shows the generated cluster frequency (refers to the frequency at which a term occurs within a given cluster) data for relevant terms in these clusters pertaining to specific TKIs.

Cluster	Term	Cluster Frequency
4	bosutinib	0.493
	imatinib	0.053
	diarrhea	0.045
	nilotinib	0.027
	gastrointestinal	0.017
	nausea	0.016
	vomiting	0.014
7	imatinib	0.255
	gastrointestinal	0.034
	stromal	0.032
	liver	0.022
	edema	0.020
	lesion	0.020
	platelet	0.019
	hematological	0.011
	cardiac	0.011
8	ponatinib	0.453
	cardiovascular	0.051
	nilotinib	0.037
	vascular	0.035
	bosutinib	0.034
	cardiac	0.025
	imatinib	0.024
	dasatinib	0.023
	platelet	0.021
	hypertension	0.019
10	dasatinib	0.367
	pulmonary	0.066
	hypertension	0.037
	imatinib	0.033
	effusion	0.032
	arterial	0.029
	pleural	0.029
	platelet	0.028
	heart	0.024
	retention	0.023

	vascular	0.022
	gastrointestinal	0.020
	pneumonia	0.012
	inflammation	0.012
	nilotinib	0.011
	rash	0.011
16	nilotinib	0.346
	imatinib	0.078
	dasatinib	0.062
	cardiovascular	0.031
	vascular	0.025
	arterial	0.024
	diabetes	0.013
	kidney	0.013
	liver	0.013
	lesion	0.013

Table 2. Color coding corresponds to colors of clusters in the cluster plot (Figure 3). The terms and clustering frequency from Table 1 are reorganized to show frequency of appearance of each TKI type across each cluster examined. Maximum frequency across all clusters is indicated with boldfaced text and dark border.

TKI/Cluster	4	7	8	10	16
bosutinib	**0.493**	0.003	0.034	0.007	0.004
imatinib	0.053	**0.255**	0.024	0.033	0.078
ponatinib	0.003	0.002	**0.453**	0.000	0.004
dasatinib	0.026	0.009	0.023	**0.367**	0.062
nilotinib	0.027	0.008	0.037	0.011	**0.346**

Table 3. Color coding corresponds to colors of clusters in the cluster plot (Figure 3). The terms and clustering frequency from Table 1 are reorganized to show frequency of appearance of different adverse events across each cluster examined. Maximum frequency across all clusters is indicated with boldfaced text and dark border.

Event	4	7	8	10	16
diarrhea	**0.045**	0.005	0.000	0.010	0.001
gastrointestinal	0.017	**0.034**	0.000	0.020	0.006
nausea	**0.016**	0.007	0.000	0.002	0.001
vomiting	**0.014**	0.002	0.000	0.000	0.001
cardiovascular	0.014	0.000	**0.051**	0.004	0.031
pleural	0.011	0.003	0.000	**0.029**	0.002
effusion	0.011	0.005	0.000	**0.032**	0.002
pulmonary	0.009	0.001	0.000	**0.066**	0.002

edema	0.004	**0.020**	0.000	0.001	0.006
rash	0.003	**0.028**	0.002	0.011	0.009
liver	0.008	**0.022**	0.006	0.000	0.013
platelet	0.000	0.019	0.021	**0.028**	0.005
myelosuppression	0.003	**0.012**	0.002	0.007	0.000
hepatitis	0.000	**0.012**	0.000	0.003	0.000
hematological	0.000	**0.011**	0.002	0.004	0.006
hypertension	0.000	0.000	0.019	**0.037**	0.003
thrombocytopenia	0.005	0.007	0.004	0.004	0.001
pneumonia	0.000	0.000	0.000	0.012	0.005
inflammation	0.000	0.007	0.000	**0.012**	0.002
vascular	0.004	0.002	**0.035**	0.022	0.025
arterial	0.000	0.000	0.021	**0.029**	0.024
diabetes	0.000	0.000	0.000	0.000	**0.013**
lesion	0.000	**0.020**	0.002	0.000	0.013

4.2 Initial SemNet Analysis of Relationships to CML

Initial SemNet simulations were conducted to assess which conditions were most associated with CML and TKI therapy in a knowledge graph comprised of nearly 30+ million PubMed abstracts. With CML as the target node, and DSYN and TOPP as the source node types, a list of 308 treatments and diseases related to CML were generated. The two SemNet features that were examined were HeteSim score and novelty score. HeteSim score represents the strength of the relationship to the target. The novelty score indicates how potentially novel a connection may be. The resultant TOPP (therapeutic or preventive procedure) source nodes yielded a list of broad procedures without specific processes or drug names, and thus were not of interest. The resultant DSYN (disease or syndrome) nodes were sorted by their HeteSim scores and novelty scores and compared. The top 20 nodes by HeteSim and the top 10 nodes by novelty are shown in Figure 5. Seven out of the top 10 resulting nodes were hematological conditions when ranked by HeteSim.

These nodes were aplastic anemia, pancytopenia, thrombocythemia, thrombocytosis, hematological disease, neutropenia, and myelosuppression. Similarly, 6 out of the top 10 resulting nodes were hematological conditions when ranked by novelty. These nodes were thrombocythemia, myelosuppression, thrombocytosis, aplastic anemia, and pancytopenia. The hematological conditions that were found to overlap in both HeteSim-ranked results and novelty-ranked results were aplastic anemia, pancytopenia, thrombocytopenia, neutropenia, and myelosuppression.

The k-means clustering examined only abstracts that specifically included CML and a TKI drug, whereas SemNet examined the entire knowledge graph of cross-domain relationships from all of PubMed. Yet, hematological conditions were frequent in both analyses. Some of these conditions shown in Figure 5 are related to the presentation of CML, itself (shown in purple bar), some are specific to CML patients taking TKIs (aqua), and for some conditions, it is not clear whether the relationship can be explicitly attributed (light blue). The grey bars in Figure 5 represent artifacts in SemNet due to the frequency of highly occurring general nodes in the UMLS, such as “complication, infection”. These general nodes do not add value to the analysis and thus were disregarded.

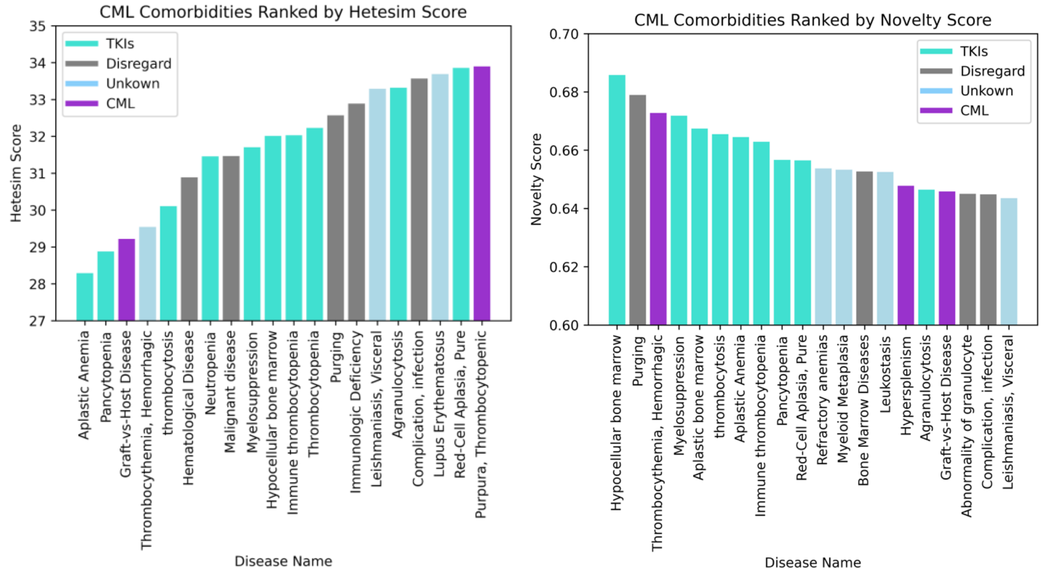

Figure 5. *Left*: The top 20 highly scoring HeteSim nodes representing comorbidities or hematological adverse conditions are shown, as identified from the SemNet simulations. A lower HeteSim score correlates to a stronger relationship to the target, so from left to right the correlation decreases. A majority of these highest ranking nodes are related to hematological deficiency. The color code was marked after literature search: green, conditions observed in patients on TKIs; orange, conditions observed in CML patients on other treatments; blue, conditions where no established relationship to CML was found; gray: conditions too broad to be informative. *Right*: The top 10 highly scoring novelty nodes representing DSYN are shown, as identified from the SemNet simulations. A higher novelty score correlates to a potentially stronger novel connection, so from left to right the novelty decreases. A majority of these highest-ranking nodes are related to hematological deficiency.

4.3 Prediction of Distant or Less Published Connections to Tyrosine Kinase

Initial SemNet simulations looked at relationships tied to the target node, CML. However, the inclusion of CML as a target constrained the identification of conditions related to tyrosine kinase pathophysiology that may not have yet been published in the CML literature. In order to examine deeper relationships of potential importance, link prediction [37] and node neighborhoods were utilized. Node neighborhoods were created using six hub nodes identified from the initial CML simulation results: aplastic anemia, anemia, neutropenia, pancytopenia, thrombocytopenia, and myelosuppression. Analysis of mutually highly ranked nodes across the hub node neighborhoods predicted relevant relationships to tyrosine kinase that are not necessarily directly connected to CML in the knowledge graph.

Cross-domain SemNet simulation results and relationship predictions were pooled for each hub node neighborhood, and the HeteSim scores were recorded and normalized to fall between 0 and 1. The mean occurrence rate and normalized HeteSim scores were aggregated to determine which conditions were most likely to be relevant to long-term TKI therapy. Foci were determined based on three tiers of aggregate rankings: tier 1 approximated the top 1%, tier 2 the approximated the top 5% and tier 3 approximated the top 10%. Foci names were not based on UMLS but rather functional classifications that best represented the nodes comprised. Tier 1 included: hematology condition (1), glucose regulation (2), iron homeostasis (3), cardiovascular conditions (4), and thyroid conditions (5). Tier 2 included inflammation (6) and kidney disorders (7). Tier 3 included gastrointestinal disorders (8) neurological disorders (9), and other (10). “Other” denoted conditions that ranked within the top 10% of aggregated and normalized nodes but could not be reasonably described by a single cohesive category. A visual representation of aggregated cross-domain text mined predictions of conditions potentially relevant to TKI therapy are shown in Figure 6. For each foci, the three highest ranked conditions are listed.

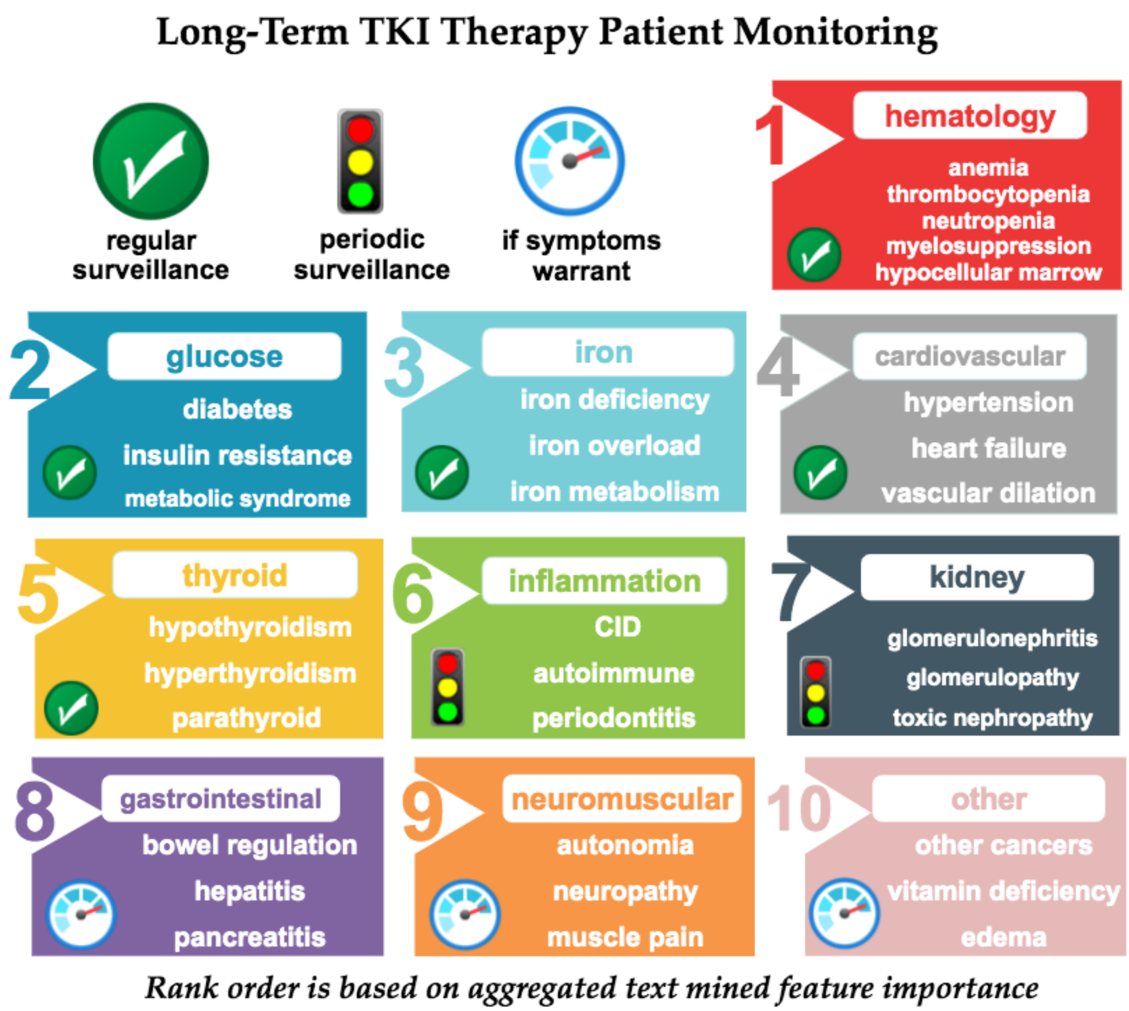

Figure 6. Conditions predicted to be most relevant to long-term TKI therapy based on aggregated text mined feature importance. Predictions are based on cross-domain literature-based discovery across all of PubMed using SemNet. A residual analysis was performed to examine the higher ranked outputted and/or intersecting nodes from SemNet simulations examining related conditions of interest with tyrosine kinase. Rankings are based on normalized and aggregated HeteSim feature importance scores. Tier 1 comprised approximately the top 1% of aggregated nodes: hematological conditions (1), glucose (2), iron homeostasis (3), cardiovascular disorders (4), and thyroid disorders (5). The second tier comprised approximately the top 5% of aggregated nodes: inflammation (6) and kidney disorders (7). The third tier comprised approximately the top 10% of aggregated nodes: neurological disorders (9) and other (10). "Other" was used to denote conditions that ranked within the top 10% of aggregated and normalized related nodes but did could not be described by a single cohesive functional category. The surveillance categories were based on the predicted importance by tier, with tier 1 warranting regularly scheduled surveillance, tier 2 warranting periodic surveillance based on additional factors (risk, age, symptoms, etc.), and tier 3 warranting monitoring predominantly based on the presence of overt symptoms and/or monitoring similar to a non-TKI treated patient in an age and gender-matched population.

Hematologic, cardiovascular, and gastrointestinal conditions were expected to be present based on the results of k-means clustering of frequent terms in the CML specific literature. However, some of the other conditions identified as relevant to tyrosine kinase in the cross-domain literature analysis were less obvious. For example, diabetes and kidney disorders only appeared in a single k-means CML literature cluster (cluster 16, nilotinib). Thyroid disorders, iron homeostasis and inflammation were not among the most frequent terms in the k-means clustering of CML specific literature.

The three tiers reflect both the prevalence of conditions in the text mined foci, as well as their predicted relevance to tyrosine kinase. As such, the assigned tier was used to suggest the level of clinical surveillance for CML patients undergoing long-term TKI therapy. Surveillance entails testing for the presence and degree of clinical or subclinical conditions that are associated with or worsened by TKI therapy. Tier 1 foci are predicted in the present analysis to prevalent and relevant enough that regularly scheduled surveillance is likely warranted regardless of patient history. Tier 2 foci are predicted to need periodic surveillance where the periodicity is expected to be less frequent than tier 1, unless a patient has relevant antecedent disease or co-morbidities that increase susceptibility to tier 2 conditions. Based on text mined predictions for prevalence and relevance to TKI therapy, tier 3 foci likely only need tested when active and/or persistent symptoms are present, unless a patient has relevant antecedent disease, co-morbidities, or other factors that increase susceptibility to tier 3 conditions.

4.4 Meta-Analysis to Quantify Anemia, Neutropenia, and Thrombocytopenia with TKI usage

Both the k-means text mining analysis of CML abstracts and the larger cross-domain text mining performed with SemNet showed several hematological conditions were associated to TKI treatments for CML. TKIs are meant to reduce cancerous white blood cells. Ironically, while myelosuppression is a known possibility for a TKI, there have not been that many cross-cohort analyses to better quantify the prevalence and degree to which this happens. Therefore, a statistical meta-analysis was conducted to validate the text mining predictions and to quantitatively analyze the prevalence of hematological conditions with TKI therapy for CML. Imatinib, dasatinib, nilotinib, and bosutinib were the TKI therapies that had sufficient data to be included for pairwise statistical analysis. Anemia, thrombocytopenia, and neutropenia were the hematological conditions that had sufficient data to be included for pairwise statistical comparison.

Figure 7 illustrates the descriptive statistics of hematological adverse event incidence as a function of hematological AE type, patient CML phase, and patient TKI therapy. More details on the patient sample sizes and citations for the extracted data are found in Table X. Figure 7A illustrates the percentage of patients with AE grouped by thrombocytopenia (green hues), neutropenia (brown hues), and anemia (blue hues). The darkest pigment intensities represent blast phase, medium pigment intensities represents accelerated phase, and the lightest pigment intensities represent chronic phase. The mean AE rates across all stages is: thrombocytopenia, 46.4%; neutropenia, 42.4%; anemia, 50.7%. Figure 7B illustrates the percentage of patients with AE grouped by CML stage, with colors and pigments matching those of panel A. The mean AE rates for each AE grouped by CML stage is: thrombocytopenia (chronic = 30.8%, accelerated = 65.4%, blast = 43.3%); neutropenia (chronic 27.0%, accelerated = 63.8%, blast = 36.4%); anemia (chronic = 29.5%, accelerated = 66.9%, blast = 55.8%). Thus, accelerated phase has the highest incidence of each hematological AE, and chronic phase has the lowest incidence of each hematological AE. Figure 7C illustrates hematological AE incidence (all severities) as a function of TKI therapy. The mean AE rates (includes all severities of AE and all phases of CML) are: bosutinib, 44%; nilotinib, 47.6%, dasatinib, 57%. imatinib, 36.8%). Figure 7D illustrates severe AE (defined as stage 3 or stage 4) as a function each TKI therapy. Insufficient data was available for bosutinib on severe AE. The mean severe AE incidence was 26.2% for nilotinib, 22.2% for dasatinib, and 14.9% for imatinib.

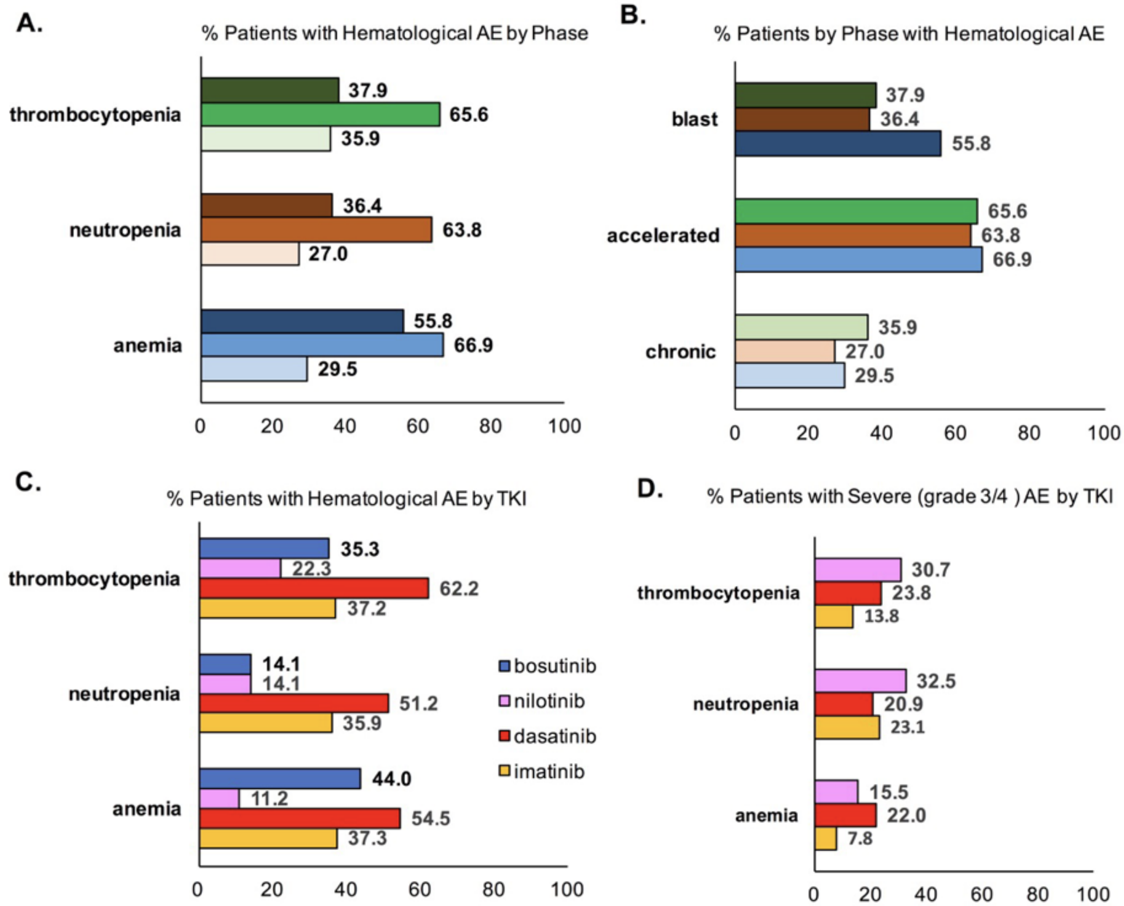

Figure 7. Descriptive statistics for aggregate incidence of hematological adverse events in CML patients treated with TKI therapies. A. Percent of patients (i.e. aggregate incidence) of each type of hematological adverse event, including all severities: thrombocytopenia (monochromatic green hues), neutropenia (monochromatic brown hues), and anemia (monochromatic blue hues). The pigment intensity or "degree of darkness" represents patient phase with chronic phase having the lightest pigments, accelerated phase the medium pigments, and blast phase the darkest pigments. B. Percent of patients (i.e. aggregate incidence) of hematological adverse events of any severity as function of patient CML phase (blast, accelerated, or chronic phase). Bar colors and pigment intensity match those described in panel A. C. Percent of patients (i.e. aggregate incidence) with hematological AE based on type of TKI therapy. D. Percent of patients (i.e. aggregate incidence) with severe (grade 3 or 4) hematological adverse events as a function of TKI therapy. Data was unavailable for bosutinib for severe hematological AEs.

Table 4. A complete list of the data acquired during the literature-based meta-analysis, as described in the Method and Results section. N/A: not applicable, or no applicable data source found. Example: To calculate the % of patients with anemia for all grades, imatinib, (#WithCondition) / (Total#) = 1036 / (1036+1738) = 37.3%.

Hematological condition	**TKI Type**	**# of Patients with Condition** From studies of: (allGrades ǀ Grade3-4Only)	**# of Patients Without Condition** From studies of: (allGrades ǀ Grade3-4Only)	**% of Patients With Condition** From studies of: (allGrades ǀ Grade3-4Only)	**Reference Number**

Anemia	Imatinib	1036	54	1738	636	37.3	7.8	4~21
	Dasatinib	1244	121	1040	430	54.4	22.0	13, 22~30
	Nilotinib	299	71	2376	387	11.2	15.5	31~35
	Bosutinib	343	N/A	437	N/A	44.0	N/A	18, 20, 36
	Ponatinib	105	N/A	312	N/A	25.2	N/A	37
	Asciminib	17	N/A	133	N/A	11.3	N/A	38
	Radotinib	48	4	112	73	30.0	5.2	14, 39
	Ruxolitinib	24	N/A	36	N/A	40.0	N/A	40, 41
	Tipifarnib	17	N/A	9	N/A	65.4	N/A	42
Neutropenia	Imatinib	921	253	1643	841	35.9	23.1	5~19, 21, 43~46
	Dasatinib	736	235	702	890	51.2	20.9	13, 22~23, 25~30, 47~48
	Nilotinib	362	251	2206	521	14.1	32.5	31, 33~35, 49~51
	Bosutinib	110	N/A	670	N/A	14.1	N/A	18, 36, 45
	Ponatinib	126	N/A	480	N/A	20.8	N/A	37, 43, 52
	Asciminib	16	N/A	134	N/A	10.7	N/A	38
	Radotinib	61	1	99	76	38.1	1.3	14, 39
	Ruxolitinib	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	Tipifarnib	12	N/A	14	N/A	46.2	N/A	42
	Imatinib	963	173	1629	1083	37.2	13.8	5~21, 43~44, 46, 53~55

Thrombo-cytopenia	Dasatinib	706	365	429	1171	62.2	23.8	13, 22~23, 25~30, 47~48, 53, 56
	Nilotinib	601	237	2095	535	22.3	30.7	31~35, 49, 50
	Bosutinib	188	N/A	344	N/A	35.3	N/A	18, 20, 36
	Ponatinib	262	N/A	344	N/A	43.2	N/A	37, 43, 52
	Asciminib	33	N/A	117	N/A	22.0	N/A	38
	Radotinib	142	N/A	95	N/A	59.9	N/A	14, 39
	Ruxolitinib	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	Tipifarnib	10	N/A	16	N/A	38.5	N/A	42

Statistical analysis was performed to examine significant differences in AE incidence (all severities) as a function of CML phase (chronic, accelerated, blast) in Figure 8. A pairwise statistical comparison was performed at an alpha of 0.05. Bonferroni correction for multiple comparisons lowered the p-value threshold of significance to $p < 0.016$ to avoid a Type I error. Anemia (Figure 8A) has significant differences between each phase. Neutropenia (Figure 8B) has significant differences in incidence between blast and accelerated phase as well as chronic and accelerated. Thrombocytopenia (Figure 8C) has significant differences in incidence between blast phase and accelerated, as well as chronic phase and accelerated phase. When comparing differences in hematological AE in all chronic phase patients (Figure 8D), there is a significant difference between anemia and thrombocytopenia, as well between thrombocytopenia and neutropenia. There were no pairwise significant differences between hematological AE type in accelerated phase patients (Figure 8E). Likewise, there were no pairwise significant differences between hematological AE type in blast phase patients. The p-values indicate that anemia is trending towards being significantly more prevalent than neutropenia ($p = 0.03$) or thrombocytopenia ($p = 0.05$). The relationships are shown as insignificant here since they exceed the Bonferroni threshold for significance in multiple comparisons. Nonetheless, with a less conservative post-hoc correction or a larger blast phase patient sample size, anemia (55.8%) may become significantly more prevalent in blast phase patients compared to either thrombocytopenia (37.9%) or neutropenia (36.4%).

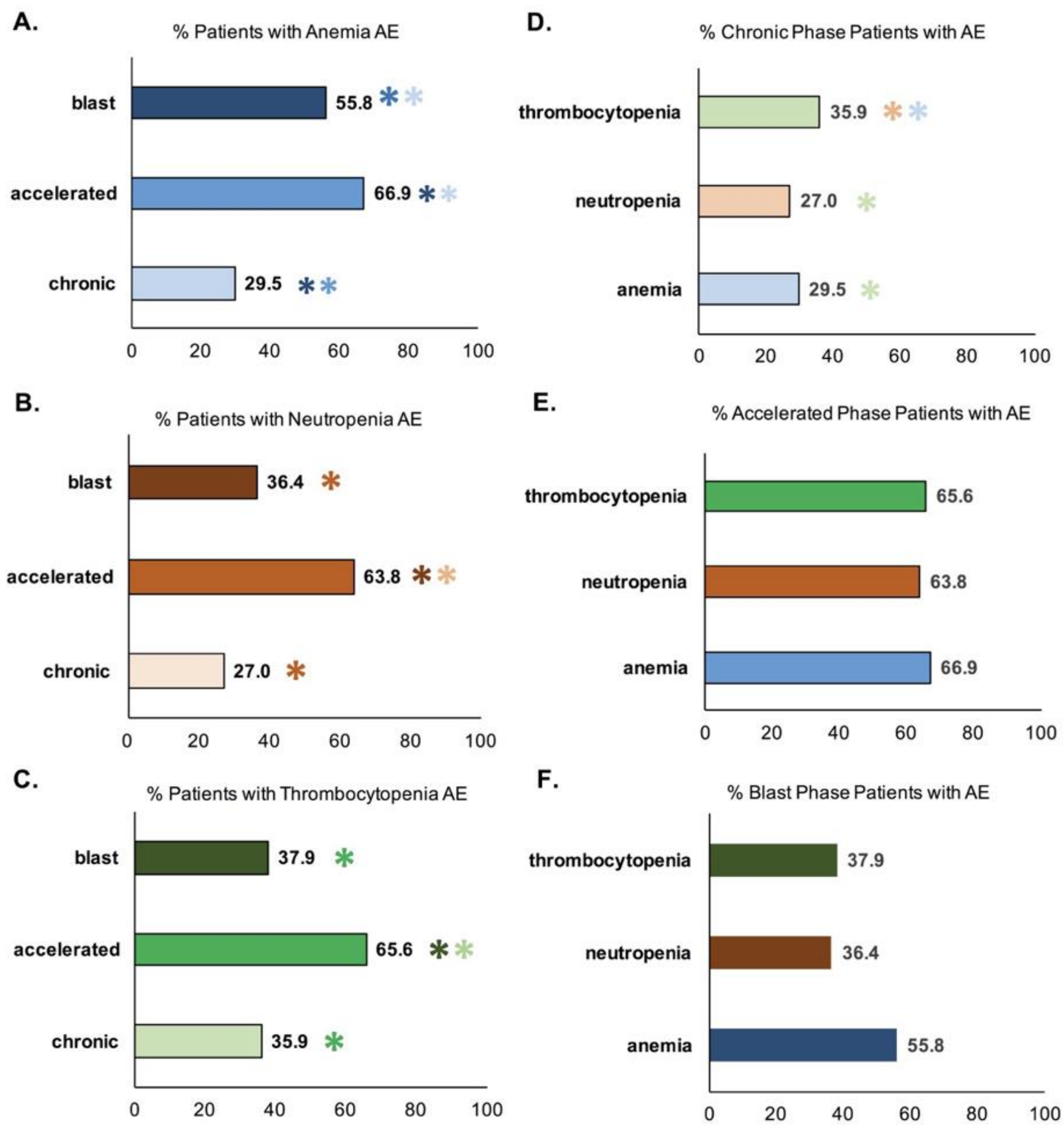

Figure 8. Pairwise statistical comparison hematological adverse events (all severities) in CML patients treated with TKIs as a function of CML stage (chronic, accelerated, or blast phase). Percentage is aggregate incidence of the sample. For pairwise analysis at an alpha of 0.05, significance less than the Bonferroni-corrected p-value of $p < 0.016$ is illustrated with an asterisk (*). The * are color-coded to illustrate with which other category there is a significant pairwise connection. Multiple asterisks (**) of different colors indicate multiple significant pairwise relationships. **A.** Percentage of patients with anemia. There is a significant difference in anemia between every phase pair. **B.** Percentage of patients (aggregate incidence) with neutropenia. Neutropenia is significantly different between blast and accelerated phase, as well as chronic and accelerated phase. **C.** Percentage of patients with thrombocytopenia. Thrombocytopenia is significantly different between blast and accelerated phase, as well as chronic and accelerated phase. **D**. Percent of chronic phase patients with hematological AE. There is a significant pairwise difference between thrombocytopenia and neutropenia, as well as thrombocytopenia and anemia. **E.** Percent of accelerated phase patients with hematological AE. There are no statistical differences between any pair of hematological AEs. **F.** Percent of blast phase patients with hematological AE. There are no statistical differences between any pair of hematological AEs. However, with a larger sample size, there could be significance between anemia and neutropenia or thrombocytopenia.

Next, statistical analysis was performed to examine significant differences in AE incidence (all severities) as a function of TKI therapy in Figure 9. A pairwise statistical comparison was performed at an alpha of 0.05. Bonferroni correction for multiple comparisons lowered the p-value threshold of significance to $p < 0.005$ to avoid a Type I error. Additionally, an odds ratio with 95% confidence interval was calculated to determine whether a TKI increase or decreased the odds of hematological AE compared to imatinib. Imatinib was used as the control in the odds ratio analysis since it is the oldest and most popular standard of care treatment. Figure 9A-C illustrates pairwise statistical significance analysis between drugs for each hematological condition. Figure 9D-F illustrates the corresponding odds ratio analysis for each hematological condition.

Incidence of anemia (all severities, all CML phases) was significantly different between every TKI therapy (Figure 9A). The descending ordering of incidence per TKI was dasatinib (54.5%), bosutinib (44.0%), imatinib (37.3%), nilotinib (11.2%). Likewise, the odds ratio analysis for anemia (Figure 9B) agreed with the pairwise analysis. The odds ratios and their 95% confidence intervals exceed 1 for bosutinib (OR =1.18) and dasatinib (OR =1.46), meaning both had significantly increased odds of anemia over imatinib. In contrast, nilotinib (OR = 0.3) significantly decreases the odds of anemia compared to imatinib. Percent of patients with neutropenia (all severities, all CML phases) was significantly higher between dasatinib and all other drugs and significantly lower between imatinib and all other drugs (Figure 9C). The odds ratio analysis agrees for neutropenia (Figure 9D) and agrees with the pairwise analysis. The OR and corresponding confidence interval was greater than 1 for dasatinib (OR = 1.42), meaning dasatinib significantly increased odds of neutropenia over imatinib. In contrast, bosutinib (OR = 0.39) and nilotinib (OR = 0.39) significantly decreased the odds of neutropenia compared to imatinib. Percent of patients with thrombocytopenia was significantly different in every pairwise comparison except bosutinib and dasatinib (Figure 9E). The OR and corresponding confidence intervals for thrombocytopenia (Figure 9F) illustrated no significant difference between bosutinib and imatinib. While bosutinib odds ratio (OR =0.95) is less than 1, its confidence interval crosses 1, meaning there was no significant difference in odds of getting thrombocytopenia with bosutinib compared to imatinib. Nilotinib (OR = 0.60) decreased odds of thrombocytopenia compared to imatinib, and dasatinib (OR = 1.67) increased odds of thrombocytopenia compared to imatinib.

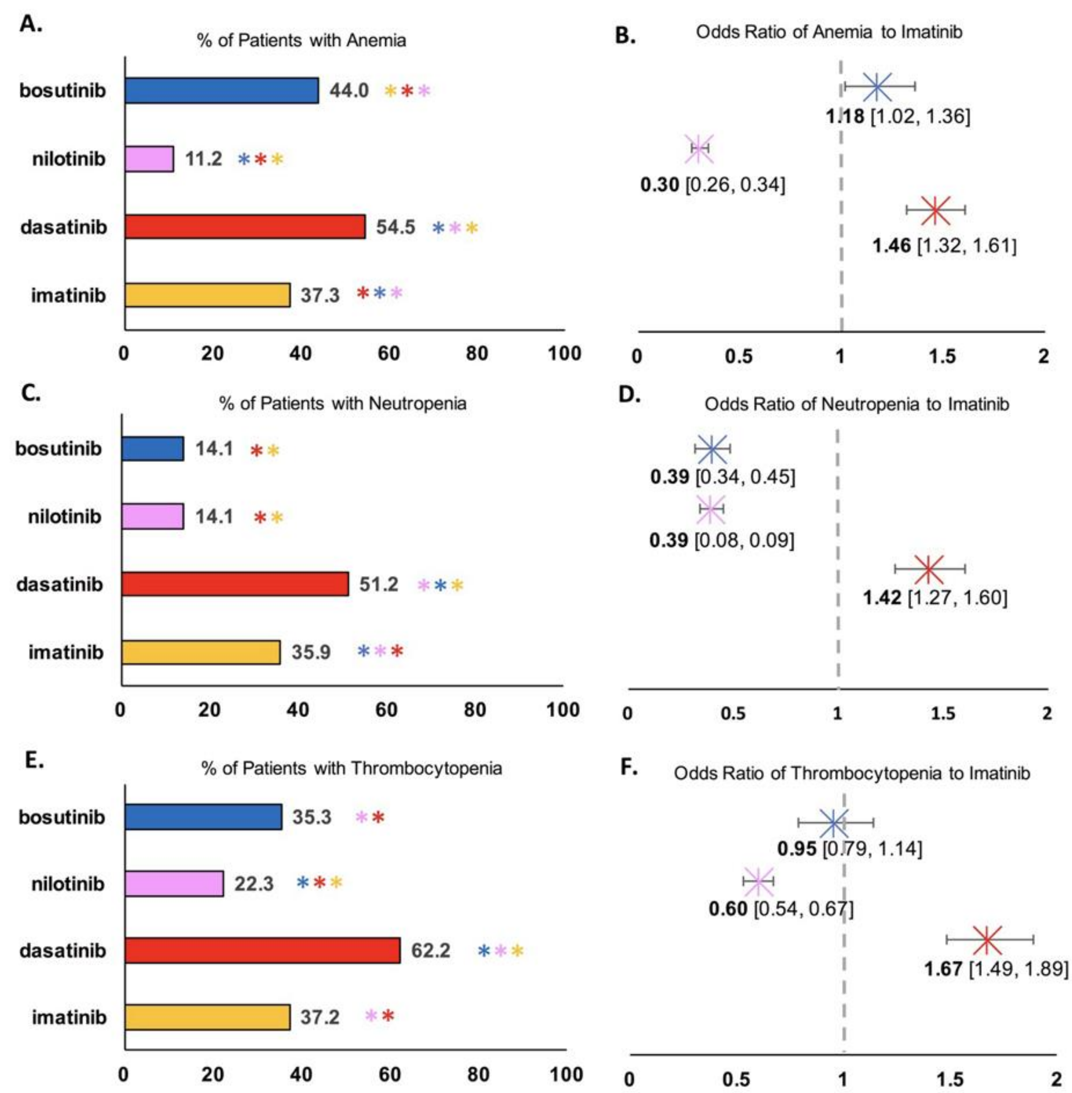

Figure 9. Statistical comparison of hematological adverse events (all severities, all CML phases) in CML patients treated with different TKIs. Percentage is aggregate incidence of the sample. For pairwise analysis at an alpha of 0.05 (Panels A., C., E.), significance less than the Bonferroni-corrected p-value of $p < 0.005$ is illustrated with an asterisk (*). The * are colored-coded to illustrate with which other TKI there is a significant pairwise connection. Multiple asterisks (**) of different colors indicate multiple significant pairwise relationships. For odds ratio analysis (Panels B., D. F.), odds of getting the hematological AE are calculated relative to imatinib. Error bars illustrate the 95%% confidence interval (CI) and the exact CI is shown inside brackets. An OR > 1 illustrates increased odds of AE compared to imatinib, and an OR <1 illustrates decreased odds compared to imatinib**. A.** Pairwise analysis of percentage of patients with anemia based on TKI therapy usage**. B.** Odds ratio of anemia for each TKI compared to imatinib. **C.** Pairwise analysis of percentage of patients with neutropenia based on TKI therapy usage. **D.** Odds ratio of neutropenia for each TKI compared to imatinib. E. Pairwise analysis of percentage of patients with thrombocytopenia based on TKI therapy usage. F. Odds ratio of thrombocytopenia for each TKI compared to imatinib.

CHAPTER 5. DISCUSSION

The present study utilized multiple types of innovative text mining to assess which types of adverse events or side effects might accompany long-term TKI therapy. While many of the predictions align with CML literature, there are also some highly ranked predictions that are less represented in the CML literature but could be pivotal to optimal patient management. The cross-domain text mining analysis included nearly 30 million PubMed articles in a knowledge graph, which enabled relationships to be identified that may not be explicitly present in the CML TKI literature. Three tiers of highly ranked foci were used to describe the potential relevance of the conditions to CML TKI patient management. Tier foci were conditions in the top 1%, tier 2 in the top 5% and tier 3 in the top 10%. In the Discussion, clinical evidence from the literature is used to validate and/or support the text mining model predictions and suggest optimal patient surveillance for each adverse event type.

5.1 Predicted Tier 1 Adverse Events Likely Tied to TKI Therapy

As noted in Figure 6, the tier 1 foci comprised the top 1% of predicted nodes relevant to tyrosine kinase and the CML hub nodes. The tier 1 foci were: 1. hematological conditions (many which were also hub nodes for the cross-domain neighborhood analysis, as illustrated in Figure 2); 2. glucose; 3. iron homeostasis; 4. cardiovascular conditions; 5. thyroid disorders. Of these conditions, hematological and cardiovascular disorders were expected based on their term frequency in the k-means clustering in the CML TKI literature. However, glucose, iron homeostasis, and thyroid disorders are less discussed in

CML literature. Thus, in particular, more awareness of these conditions as possible clinical or subclinical adverse events is necessary for optimal long-term TKI therapy patient care.

5.1.1 Hematological conditions as adverse events from TKI therapy

The innovative text mining analyses illustrated that myelosuppressive hematological conditions (namely anemia, neutropenia, thrombocytopenia, and pancytopenia) were the most connected to TKI usage for CML. This is not unexpected given the mechanism of action of TKIs and their intended target of stopping proliferation of Philadelphia chromosome positive white blood cells. However, the significant differences in quantified prevalence of these conditions, as shown in the statistical meta-analysis, illustrate their pervasive nature and the variations in myelosuppression across specific TKIs. As shown in the statistical meta-analysis (Figure 8, 9), a large portion of patients taking TKIs do present with one or more blood cell type deficiencies. Based on the present study's statistical meta-analysis results, approximately one-third of chronic phase CML patients or CML patients in remission who remained on TKI therapies have either anemia, neutropenia, or thrombocytopenia. Dasatinib had the highest prevalence of hematological adverse events (anemia, neutropenia, thrombocytopenia) of the TKIs, and nilotinib the lowest prevalence. For this reason, monitoring of blood cell counts even after patients have reached a deep molecular response is important given these adverse events could present even later in therapy. Anemia can be worsened or complicated by iron deficiency, which is the number two predicted adverse event. Persistence of a hematological adverse event or the deficiency of more than one blood cell type could suggest hypocellular bone marrow, which happens as a result of prolonged myelosuppression. Hypocellular bone marrow was one of the nodes that was highly ranked in the hematological condition foci. In summary, the mechanism by which TKIs work mean that myelosuppression is a real and somewhat expected possibility. Thus, the goal is adjusting the dose of TKI to minimize myelosuppression that causes anemia, neutropenia, or thrombocytopenia, or for patients who more susceptible to myelosuppression due to other co-morbidities or antecedent disease, select a TKI where myelosuppressive adverse events are not as pervasive. Due to the high prevalence of hematological adverse events, the results of the present study suggest regular monitoring of all CML TKI patients, and minimally a regularly performed complete blood count test. Where symptoms are present, more specific testing is warranted to determine if the hematological events are being worsened by other co-morbidities or TKI-related AEs, like iron or B12 deficiency, kidney dysfunction, thyroid dysfunction, etc.

5.1.2 Glucose-related adverse events from TKI therapy

Glucose-related conditions were predicted to be the second most relevant adverse event with TKI therapy. In the case of glucose, the impact appears to depend on habituation to the drug. Early on, patients on TKIs tend to have lower blood sugar levels or A1c. In fact, imatinib was used recently in a study examining it has a possible therapeutic for Type I diabetes [39]. While the drug lowered blood sugar initially in the first 12 months, the effect did not last when measured through 24-months [39]. Another published review also discussed the use of TKIs for type 2 diabetes treatment [40]. According to the text mining simulations here, insulin resistance and metabolic syndromes are predicted as highly ranked possibilities of long-term TKI therapy. Thus, it is possible that patients initially see lowering of glucose levels and insulin resistance. However, TKI with long-term therapy, habituation, the innate physiology overcompensates, potentially resulting in insulin resistance, pre-diabetes, metabolic disorder, or eventually, even Type 2 diabetes. Depending on age, lifestyle, and other co-morbidities this condition may not be recognized as attributed to TKIs, even though results presented here predict a strong association. Nilotinib was associated with diabetes in the k-means clustering text mining performed in this study. In fact, one clinical panel has recommended that nilotinib not be the primary therapy for CML patients who also have diabetes [41] as many studies have tied it to diabetes or minimally impaired glucose metabolism [42] [43]. Higher glucose is also associated with iron deficiency anemia; as such these two conditions are intertwined. Insulin resistance can also compound cardiovascular co-morbidities or TKI-related AEs. As such, regular monitoring of glucose and A1c [44] is recommended based on the results of this study.

5.1.3 Iron homeostasis adverse events with TKI therapy

The iron foci ranked number 3 in the cross-domain text mining analysis. Once again, opposing conditions are identified: iron deficiency (too little iron) and iron overload (too much iron). While it is possible both are relevant to TKIs, it is also plausible that iron overload is an artifact in this analysis due to its strong relationship with blood transfusions, a TOPP node which was strongly connected with the hub nodes used to generate the neighborhood rankings for determining the foci. Both iron deficiency and iron overload can impact the thyroid, which is the fifth ranked foci. A recent study showed than iron chelator, deferasirox, improved imatinib resistance and increased the clearance of BCR ABL cells [45]. Most of the iron connections in the graph were related to iron deficiency anemia. Given the importance of iron to red blood cell function, preventing anemia, and maintaining proper thyroid function, regular monitoring with an iron panel (peripheral blood levels of iron, ferritin, iron binding capacity, etc.) is recommended based on the results of this study.

5.1.4 Cardiovascular adverse events with TKI therapy

The cardiovascular foci ranked number 4 in the cross-domain text mining analysis. However, cardiovascular event terms were even more frequent in the k-means clustering of CML literature. Hypertension adverse events are approximately 10% for those treated with imatinib, dasatinib, bosutinib, nilotinib [46] but are as high as 30% of patients for patients treated with ponatinib [47]. Heart failure, hypertension, Q-T prolongation, and vascular events are cardiovascular adverse events that are most monitored among CML patients treated with TKIs. Heart failure ranges from <1% for imatinib or dasatinib to as high as 4% for patients on bosutinib or ponatinib [46]. Q-T prolongation and the associated sudden cardiac death in 0.3% of nilotinib patients prompted the black-box warning by the Food and Drug Administration; approximately 1% of dasatinib or imatinib patients had clinically significant Q-T prolongation but much less associated sudden cardiac death [46]. Arterial vascular events are approximately 2% for imatinib, 5% for dasatinib, and 11% for nilotinib, although the nilotinib number was based on relatively much smaller study [46]. Notably, ponatinib was most associated with "cardiovascular" and "arterial" in the k-means clustering of CML literature, followed by nilotinib; these results align well with the cited clinical meta-analyses results. Many clinicians agree that cardiovascular risk as one of the most important factors for selecting a long-term TKI regime [48]. Echocardiogram and electrocardiogram testing is currently recommended to monitor for cardiovascular AEs with TKI therapy [46] [48]. The results of the present study concur, given cardiovascular events were predicted as a tier 1 foci.

5.1.5 *Thyroid disorders as adverse events with TKI therapy*

The thyroid foci was ranked number five in the cross-domain text mining analysis. Again, opposing conditions are identified hypothyroidism (too little thyroid hormone) and hyperthyroidism (too much thyroid hormone). In the general [non-CML] population, hyperthyroidism is typically attributed to Grave's disease, an autoimmune disorder. However, neutropenia is also associated with hyperthyroidism. Certain types of TKIs can cause hyperthyroidism due to their mechanism of action in the down-regulation tyrosine kinase, especially in patients who have a history of thyroid problems [49], with hyperthyroidism most likely during the early stages of TKI therapy. However, a recent 2020 study found that hypothyroidism is more prominent than hyperthyroidism in CML patients treated with TKIs [50]. Of 326 patients who were on TKI therapy and had a thyroid disorder as an adverse event, 74% had hypothyroidism and 20% had hyperthyroidism [50]. Of the 326 CML TKI therapy patients with a thyroid AE, hypothyroidism accounted for 73% of cases with imatinib, 75% with dasatinib, 48% with nilotinib, and 87% with bosutinib [50]. Thus, while hyperthyroidism may be more likely in early TKI therapy, long-term TKI therapy can result in physiological overcompensation that leads to hypothyroidism.

The parathyroid function is also of potential importance, although with a lesser predicted connection than the thyroid, itself, for TKI adverse events. Secondary hyperparathyroidism has been associated with bone marrow fibrosis (hypocellular bone marrow), which can cause pancytopenia, anemia, thrombocytopenia, or neutropenia. Patients with hyperparathyroidism may not exhibit overt symptoms early in the disease course, making monitoring all the more important. Changes in blood calcium (namely hypercalcemia) or phosphatase (namely hypophosphatemia) due to hyperparathyroidism can also impact cardiovascular adverse events [51]. In one study, up to 30% of patients taking nilotinib have reported hypophosphatemia [52].

Finally, there are key relationships between anemia, iron homeostasis, and thyroid function which complicates this triad of tier 1 foci of TKI-related adverse events. Anemia caused by iron deficiency (iron-deficiency anemia, or IDA) impairs thyroid metabolism function. Both overt and subclinical hypothyroidism are associated with anemia. Previous studies have shown improvement in both conditions with the addition of iron to thyroxine therapy [53]. In conclusion, regular peripheral blood testing with a thyroid panel or minimally TSH (thyroid stimulating hormone) and T4 (thyroxine) levels is recommended based on the results of the present study. Also, if regular peripheral blood chemistry testing reveals changes in calcium or phosphatase, or if a follow-up bone marrow biopsy reveals hypocellular bone marrow, parathyroid testing should be considered. Prior work suggests testing is recommended for patients.

5.2 Predicted Tier 2 Adverse Events Likely Tied to TKI Therapy

As noted in Figure 6, the tier 2 foci comprised approximately the top 5% of predicted nodes relevant to tyrosine kinase and the CML hub nodes in the cross-domain simulations. The tier 2 foci were: 6. Inflammation; 7. kidney function. Inflammation was one of the adverse events that occurred frequently in the k-means clustering of CML literature and was especially tied to dasatinib. "Kidney" occurred at the highest frequency in the nilotinib cluster.

5.2.1 Kidney adverse events with TKI therapy

Cluster 16 from the clustering methods, which was dominated by nilotinib, had the term “kidney” occur at high frequency within the cluster. All text mining analyses revealed a connection between kidney conditions and hematological condition, namely anemia). SemNet simulations predicted glomerulonephritis, glomerulopathy, toxic nephropathy, and chronic kidney disease as kidney conditions relevant to tyrosine kinase.

Chronic kidney disease (CKD) is characterized by excessive fibroblast activation, and it is hypothesized that tyrosine kinase can increase fibroblast activation and contribute to CKD development as a result. Therefore, it may be valuable to consider the impact of tyrosine kinase inhibitors on fibroblast activation by reversing the effect of tyrosine kinases, and in turn evaluate the impact of TKI treatment on the kidney health of patients. However, most TKIs inhibit more than one kinase, making their exact effect unclear and their individual impact on kidney health difficult to determine [54]. Some studies show imatinib, which inhibits BCR-ABL, c-Kit, and PDGF receptors, to improve the chronic kidney disease of animal models and, nilotinib, which inhibits BCR-ABL and PDGF receptors, to improve the thioacetamide based conditions of animal models [55]. However, other studies showed CML patients with pre-existing renal dysfunctions who took dasatinib and nilotinib experienced increased risk for acute kidney injury [56]. Another study showed imatinib to decrease renal function in patients with pre-existing renal dysfunction [57], and there also case reports of dasatinib-induced nephrotic syndrome [58]. Kidney conditions, including chronic kidney insufficiency and general kidney diseases, are cited heavily in literature to contribute to the development of anemia even in non-CML populations.

From the neighborhood analysis for the anemia hub, 5 nodes were identified as relevant to kidney function: iron deficiency, PTH , urea, uric acid, and C-reactive protein. While these nodes have clear and heavily weighted ties to the hematological hub nodes, it is unclear whether they are truly important to CML patients on TKI therapy who do not have a history of pre-existing kidney disease. It is possible that the hub node neighborhood methodology resulted in the over-weighting of the relevance of kidney disease due to its strong connections in each hub. However, the well-known impact of fluid retention [59] and fluid-related pleural effusion [60], which are both discussed in Section 2.4.3.3, amplifies the importance of the necessary good kidney function for patients on TKIs. In summary, based on the novel predictions of the present study and corroborating manual literature validation, it is recommended that patients only undergo periodic testing, with the periodicity based on the presence of relevant antecedent disease or other relevant risk factors that impact susceptibility. Metrics such as serum creatinine, glomerular filtration rate, and blood urea nitrogen are suitable candidates to help test for proper kidney function.

5.2.2 Inflammation adverse events with TKI therapy

Inflammation was most represented in cluster 10 (dasatinib) but was also represented in cluster 7 (imatinib) and cluster 16 (nilotinib). "Inflammation" could represent many different types of symptoms. The three highest ranked types of inflammatory disorders in the cross-domain text mining analysis were chronic inflammatory disorder (CID), autoimmune diseases, and periodontitis. Autoimmune disease has been associated been identified as predisposing risk factor for getting CML;

one study found the odds ratio of getting CML was 1.55 (compared to 1) for patients that had an autoimmune disease prior to being diagnosed with CML [61]. There have also been ties to neurological autoimmune diseases which is discussed in Section 4.3.2. It was hypothesized that inflammation is an indicator of long-term success in CML. Yet, a more recent study found that biomarkers for inflammation could not predict whether CML patients were more likely to be successful or if they were more likely to relapse when attempting treatment-free remission (TFR), which is a planned cessation of TKI therapy [62]. There have been case reports of colitis followed by dasatinib TKI therapy [63]. Another study has reported that predisposition to pro-inflammation or pro-oxidative stress increases atherothrombotic adverse events with TKI therapy, particularly nilotinib [64]. Similarly, dasatinib has been connected to immune-mediated thrombotic thrombocytopenic purpura [65]. In summary, it appears that inflammation has connections to CML predisposition and is most likely a relevant adverse risk factor especially to patients who have pre-existing inflammatory disorders. Based on the predictions of this study and supporting direct literature evidence, it is recommended that inflammatory markers be periodically monitored with the periodicity based on the presence of relevant antecedent disease or other relevant risk factors, that impact susceptibility. In particular, patients more at risk for clotting should have inflammation markers monitored.

5.3 Predicted Tier 3 Adverse Events Likely Tied to TKI Therapy

As noted in Figure 6, the tier 3 foci comprised approximately the top 10% of predicted nodes relevant to tyrosine kinase and the CML hub nodes in the cross-domain simulations. The tier 3 foci were: 8. gastrointestinal; 9. neuromuscular; 10. Other. Gastrointestinal events were also frequent in the k-means clustering and accompanied by nearly every TKI. Neuromuscular or neurological were not as frequent in the k-means clustering. “Other” represented conditions that were in the top 10% but could not be categorized into an overlapping functional foci.

5.3.1 Gastrointestinal adverse events with TKI therapy

For the cross-domain text mining analysis, bowel regulation (diarrhea or constipation), liver dysfunction, and pancreatitis were among the most highly ranked gastrointestinal events. K-means clustering of CML literature found that gastrointestinal was among the most frequent terms and appeared with all TKIs. Gastrointestinal adverse events are some of the most common symptoms reported by CML patients. Many patients report nausea, vomiting, and diarrhea. A recent large cross-cohort meta-analysis extensively examined gastrointestinal adverse events with TKI therapy [28]. Briefly, results illustrated bosutinib had the most prevalent and severe gastrointestinal AEs, with dasatinib and imatinib being very similar and nilotinib having the least number of reported gastrointestinal events [28]. Diarrhea is most common early in treatment with some patients having more constipation with habituation to the drug with long-term treatment, perhaps another over-compensation of innate physiology to the TKI. In general, the most common gastrointestinal events tend to abate over time. However, attention must be paid to liver and pancreatic function given overt symptoms may not be present until more acute problems arise. A recent large-scale clinical meta-analysis reported the prevalence of toxic hepatopathology [66]. Statistically significant increases in toxic hepatopathology were reported for every TKI except dasatinib [66]. Some patients have been switched to dasatinib following imatinib hepatotoxicity [67]. A recent review article discusses the pharmacokinetics of TKI that lead to specific gastrointestinal adverse events [68]. Increase in amylase has been seen in some CML patients treated with TKIs [69], and while pancreatitis is a less frequently reported AE even in meta-analyses [28], the severity of the AE warrants some prudence in monitoring, particularly when possible symptoms appear. Current data suggest that ponatinib appears to have the most frequent incidence of pancreatitis [70]. In summary, the gastrointestinal foci were a tier 3 foci based on cross-domain text mining predictions. Based on the results of this study and supporting clinical evidence, gastrointestinal adverse events should be monitored based primarily on the presence of symptoms. Basic liver function can be assessed with a simply peripheral blood chemistry test, which should be performed regularly based on the tier 1 foci discussed above where blood chemistry testing provides helpful surveillance. If symptoms warrant, a peripheral blood test of amylase can provide evidence to pancreatitis. Of course, if a patient has known susceptibility or other complicating factors, periodic testing is warranted

5.3.2 Neuromuscular adverse events with TKI therapy

Based on cross-domain text mining analysis, the most highly ranked neuromuscular adverse events predicted to be related to tyrosine kinase included autonomia, neuropathy, muscle pain, and cognitive dysfunction. Neurological disorders did not show up as frequent term in the k-means clustering of CML literature, although musculoskeletal was appeared relatively often even though the term did not make the top terms list shown in Table 3.

There are mixed clinical findings on the impact of pre-existing neurologic autoimmune disease versus neurologic autoimmune adverse events caused by TKI therapy. The association of autoimmune neurological diseases like myasthenia gravis has a strong enough association with CML that some have suggested that they should be classified as paraneoplastic neurologic syndrome [71]. New demyelinating adverse events with TKI therapy have also been reported with dasatinib [72], especially demyelinating peripheral neuropathy. Peripheral neuropathy has also been reported with imatinib [73]. Some patients with this adverse event with dasatinib or imatinib were successfully transitioned to nilotinib [74]. However, nilotinib is not free of neurological adverse events either. There are reports of dystonia and cognitive deficits with nilotinib [75]. There have also been recent reports of neurocognitive deficits with dasatinib [76]. By far the most common neuromuscular adverse event is musculoskeletal pain, which can occur with initial treatment with TKIs [77] [78] or as part of withdrawal to the TKI during a planned TFR attempt to stop TKI therapy [79]. In summary, based on the predictions of the present study and corroborating clinical evidence, most neurological or neuromuscular adverse events require monitoring based only on the presence of overt symptoms.

5.3.3 Other adverse events with TKI therapy

In the present study, the "other" foci were used to categorize conditions that were found among the top 10% of predicted relevant conditions to tyrosine kinase. Secondary cancers were frequently occurring node. Secondary cancers have been associated with CML both before and after CML diagnosis [61] [80] [81] [82] [83]. Essentially every cited study found statistically significant increases in secondary malignancy with CML and TKI therapy. This is not unexpected given most forms of cytotoxic chemotherapy or immune modulating therapy increases risk of secondary malignancies.

Other conditions in the foci that were predicted as relevant to TKIs were various forms of vitamin deficiency such vitamin B6, B12, vitamin D, vitamin C, and of course, iron, which was ranked enough to be its own tier 1 foci. The majority of these vitamins are tied to hematopoiesis, which explains why they are likely present in these tier 3 foci. Beyond ties to hematopoiesis, there are presently no widespread hypotheses as to how or why vitamin deficiency may occur with TKI therapy.

Edema was another condition ranked in the top 10% in the cross-domain text mining analysis, and it also was a frequent term in the k-means clustering of CML literature, with imatinib being a strong association, as well as dasatinib [59]. Moreover, edema is noted as a patient-reported side effects that increases desire to try TFR [84]. Skin conditions are another adverse even type that belonged to the other tier 3 foci.

Imatinib has been documented to contribute to developments of rash and skin lesions, as is seen in cluster 7, where the terms "rash" and "lesion" occur at the highest frequency.

Pleural effusion is perhaps one of the most well-known acute adverse events with TKI therapy. Dasatinib is documented to induce pleural effusion in some CML patients at higher rates than other TKIs [60]; this is also seen in our findings for cluster 10, where the terms "pleural" "effusion", and "pulmonary" occur at relatively high frequencies. Interestingly, pleural effusion was not an event predicted in the top 10% for the cross-domain text mining simulations. Pleural effusion is tied to fluid retention and edema, which were predicted as very relevant connections to TKIs. Pleural effusion is an acute event, although the presence of edema certainly increases the risk of occurrence.

In summary, most other conditions should be monitored based on either present symptoms or according to general age and gender matched clinical guidelines for non-

CML patients. Given the higher incidence of secondary malignancies, patients should be encouraged to be proactive in standard screenings such as mammograms and pap smears (female) and prostrate screenings (male), colon cancers for those over the age of 45, etc.

CHAPTER 6. CONCLUSION

The need for long-term therapy with cytotoxic tyrosine kinase inhibitors to treat CML warrants careful evaluation and patient monitoring for undesired adverse events or side effects. CML is a relatively rare condition and TKIs are still relatively new, with only about 20 years since the first TKI, imatinib, came to market. Multi-factorial cross-cohort longitudinal studies of long-term adverse events and side effects are not readily available. Text mining provides an innovative means to assess potential long-term undesired effects of TKIs by examining cross-domain pathophysiological connections between TKIs and other potential related adverse events or side effects. Three types of text mining were performed: 1) bag-of-words text mining analysis followed by k-means clustering to identify frequent co-occurring terms that connect adverse event conditions with specific TKIs within the CML literature; 2) Cross-domain text mined relationships from all of PubMed within a heterogeneous information network (also known as a knowledge graph) using SemNet; 3) Link prediction of potential but yet to be explicitly published cross-domain relationships based on similar patterns of relationships in the knowledge graph and hub node neighborhood analysis.

All text mining analyses predicted that hematological adverse events were the most related to CML TKI therapy. A statistical meta-analysis to quantify the presence of anemia, neutropenia, and thrombocytopenia confirms the validity of this prediction. Approximately 1 of 3 TKI patients experienced a hematological AE and with significant differences in prevalence between TKI drug types. While there are variations based on specific AE type and severity, overall, dasatinib has the most hematological AEs.

The cross-domain and hub node neighborhood analysis predicted three tiers of adverse events as most relevant to tyrosine kinase. Tier 1 foci (top 1%) included hematological events, glucose, iron, cardiovascular, and thyroid conditions. Thus, based on the predicted results and supporting clinical evidence, tier 1 foci are recommended for regular monitoring of all TKI patients regardless of medical history. Tier 2 foci (top 5%) included inflammation and kidney conditions. The associations with these conditions are less clear than the tier 1 conditions but still reflect the need for periodic surveillance with periodicity determined by individual patient medical history, co-morbidities, or susceptibility. Tier 3 foci (top 10%) include gastrointestinal, neuromuscular, or "other" conditions. The tier 3 conditions warrant monitoring based primarily on the presence of overt symptoms or patient-specific risks. However, proactive monitoring and standard age and gender-matched cancer screenings recommendation are particularly relevant to CML patients given their greater risk of secondary malignancies. In summary, while TKIs have been pivotal in changing the outcome of CML, these medications are not benign. More careful and proactive monitoring of adverse events for patients on long-term TKI therapies is recommended. Finally, personalized medicine to optimally select TKIs based on medical history, lifestyle, and other individual factors can help mitigate some adverse event risks.

Funding Acknowledgement: This research was funded by the National Science Foundation CAREER grant 1944247 to C.M, National Institute of Health grant R21-CA232249 to C.M., and Children's Hospital of Atlanta Aflac Pilot Grant Award to C.M.

CHAPTER 7. FUTURE WORK

In this work, it was shown that text mining can be used to uncover semantically related concepts within the biomedical knowledge graph. In the presented application, text mining identified long-term adverse effects associated with chronic myeloid leukemia treatment. This is just one of many ways text mining can be used to improve outcomes in leukemia. Beyond ameliorating adverse leukemia treatments, text mining can also be applied in other aspects of disease management, such as further examining the causes of leukemia, identifying risk factors, and devising new therapies.

Text mining techniques have already been shown to be serviceable for disease etiology. A major challenge in systems biology is identifying the various relationships between the exorbitant amounts of entities within the literature. There are numerous protein-protein interactions and molecular pathways that can influence the progression of diseases. Natural language processing methods have been applied to help streamline the extraction of causal relationships between certain diseases and other genetic phenomena in the literature [85]. By conducting specified natural language processing towards key molecules in leukemia progression, such as B cell lymphocytes in chronic lymphocytic leukemia or BCR-ABL gene in chronic myeloid leukemia, the molecules' roles in the progression of their respective leukemia can be further elucidated, providing possible direction and reference for future experimental studies.

Researchers have also begun to use text mining on electronic patient records to help determine disease correlations and stratify patient cohorts [86]. Roque et al. state that clinics currently stratify their patients mainly from their primary diagnosis. They demonstrate that by clustering the patients through a more complete profile constructed by their text-mined patient history, they can stratify the patients more effectively as opposed to stratifying just from their diagnosis. Extending their findings to a leukemia cohort could prove to be instrumental towards improving future leukemia patient management. There is an increasing quantity of data on leukemia patient quality of life and therapeutic efficacy, but a paucity of the utilization of holistic patient profiling in chronic myeloid leukemia patients [87]. Beyond peer-reviewed journal articles, other textual data sources like electronic patient records and patient questionnaires, could also be mined to improve nuanced clinical decisions.

Many aspects of leukemia research, treatment, and patient management remain unsolved. Text mining can accelerate the knowledge discovery process with the ability to extract deeper or lesser-known relationships among leukemia concepts [88]. Even prior to leukemia diagnosis, there may be risk factors specific to a clinical cohort that are not entirely considered. Text mining could highlight these risk factors, which in turn would allow for a more prophylactic approach for screening. Should a patient become diagnosed, text mining can potentially postulate a more elaborate yet effective treatment option, such as combination therapy.

For example, it is hypothesized that some clinical cohorts are more prone to developing infections during their rigorous leukemia treatment. In small, heterogeneous cohorts where standard clinical studies have insufficient sample size, cross-domain text mining could potentially identify which factors or metrics are the best predictor(s) for infection.

In summary, while the case study presented in this thesis focused on predicting long-term adverse events in chronic myeloid leukemia therapy, the same general techniques could be used on other leukemia populations. Text mining is particularly valuable when patient populations are smaller, and thus, direct clinical investigation study with traditional statistics is more difficult. Cross-domain text mining overcomes the limitations of sample and patient heterogeneity to mine patterns that span multiple domains. Future work will focus on utilizing and enhancing the text mining techniques presented here to improve personalized medicine for both CML and other non-CML leukemias, including both adult and pediatric populations.

APPENDIX A.

A.1 Conference Proceedings

Varmeziar, A., Allegri, S., White, B., Al-Hussaini, I., Miller, T., Mitchell, C.S. (2021). SemNet- Text Mining to Identify Predictive Features of Infection Susceptibility. American Association of Physicists in Medicine. 2021 Conference on Data Science in Medical Physics. Virtual conference. 26-27 Sept. 2021.

REFERENCES

[1] Chan, I. S., & Ginsburg, G. S. (2011). Personalized medicine: progress and promise. Annual Review of Genomics and Human Genetics, 12, 217–244. https://doi.org/10.1146/annurev-genom-082410-101446

[2] Cohen, A. M., & Hersh, W. R. (2005). A survey of current work in biomedical text mining. *Briefings in Bioinformatics*, *6*(1), 57–71. https://doi.org/10.1093/bib/6.1.57

[3] Zhu, F., Patumcharoenpol, P., Zhang, C., Yang, Y., Chan, J., Meechai, A., Vongsangnak, W., & Shen, B. (2013). Biomedical text mining and its applications in cancer research. *Journal of Biomedical Informatics*, *46*(2), 200–211. https://doi.org/https://doi.org/10.1016/j.jbi.2012.10.007

[4] Kobayashi, V. B., Mol, S. T., Berkers, H. A., Kismihók, G., & Den Hartog, D. N. (2018). Text Mining in Organizational Research. Organizational Research Methods, 21(3), 733–765. https://doi.org/10.1177/1094428117722619

[5] Tan, A.-H., Ridge, K., Labs, D., & Terrace, H. (2000). *Text Mining: The state of the art and the challenges*.

[6] Cohen, K. B., & Hunter, L. (2008). Getting started in text mining. *PLoS Computational Biology*, *4*(1), e20. https://doi.org/10.1371/journal.pcbi.0040020

[7] Wang, J., Liu, P., She, M. F. H., Nahavandi, S., & Kouzani, A. (2013). Bag-of-words representation for biomedical time series classification. *Biomedical Signal Processing and Control*, *8*(6), 634–644. https://doi.org/https://doi.org/10.1016/j.bspc.2013.06.004

[8] Zhao, R., & Mao, K. (2018). Fuzzy Bag-of-Words Model for Document Representation. *IEEE Transactions on Fuzzy Systems*, *26*(2), 794–804. https://doi.org/10.1109/TFUZZ.2017.2690222

[9] Kastrin, A., Rindflesch, T. C., & Hristovski, D. (2016). Link Prediction on a Network of Co-occurring MeSH Terms: Towards Literature-based Discovery. *Methods of Information in Medicine*, *55*(4), 340–346. https://doi.org/10.3414/ME15-01-0108

[10] Manisha, P. (2012). Leukemia: a review article. *International Journal of Advanced Research in Pharmaceutical & Bio Sciences*, *1*, 397+. https://link.gale.com/apps/doc/A357968820/HRCA?u=gainstoftech&sid=googleScholar&xid=00308d6f

[11] “Chronic Myeloid Leukemia - Cancer Stat Facts.” SEER, https://seer.cancer.gov/statfacts/html/cmyl.html.

[12] Athale, U., et al. (2019). "Management of chronic myeloid leukemia in children and adolescents: Recommendations from the Children's Oncology Group CML Working Group." Pediatr Blood Cancer 66(9): e27827.

[13] Rea, D. and J. M. Cayuela (2018). "Treatment-free remission in patients with chronic myeloid leukemia." Int J Hematol 108(4): 355-364.

[14] Massimino, M., et al. (2020). "ABL1-Directed Inhibitors for CML: Efficacy, Resistance and Future Perspectives." Anticancer Res 40(5): 2457-2465.

[15] Atallah, E. and K. Sweet (2021). "Treatment-Free Remission: the New Goal in CML Therapy." Curr Hematol Malig Rep 16(5): 433-439.

[16] Mahon, F. X. and G. Etienne (2014). "Deep molecular response in chronic myeloid leukemia: the new goal of therapy?" Clin Cancer Res 20(2): 310-322.

[17] Cortes, J., et al. (2019). "Treatment-free remission with first- and second-generation tyrosine kinase inhibitors." Am J Hematol 94(3): 346-357.

[18] Rousselot, P., et al. (2007). "Imatinib mesylate discontinuation in patients with chronic myelogenous leukemia in complete molecular remission for more than 2 years." Blood 109(1): 58-60.

[19] Ross, D. M., et al. (2013). "Safety and efficacy of imatinib cessation for CML patients with stable undetectable minimal residual disease: results from the TWISTER study." Blood 122(4): 515-522.

[20] Saussele, S., et al. (2018). "Discontinuation of tyrosine kinase inhibitor therapy in chronic myeloid leukaemia (EURO-SKI): a prespecified interim analysis of a prospective, multicentre, non-randomised, trial." Lancet Oncol 19(6): 747-757.

[21] Ross, D. M., et al. (2018). "Durable treatment-free remission in patients with chronic myeloid leukemia in chronic phase following frontline nilotinib: 96-week update of the ENESTfreedom study." J Cancer Res Clin Oncol 144(5): 945-954.

[22] Hughes, T. P., et al. (2021). "Long-term treatment-free remission in patients with chronic myeloid leukemia after second-line nilotinib: ENESTop 5-year update." Leukemia 35(6): 1631-1642.

[23] Shah, N. P., et al. (2020). "Dasatinib discontinuation in patients with chronic-phase chronic myeloid leukemia and stable deep molecular response: the DASFREE study." Leuk Lymphoma 61(3): 650-659.

[24] Shih, Y. T., et al. (2019). "Treatment value of second-generation BCR-ABL1 tyrosine kinase inhibitors compared with imatinib to achieve treatment-free remission in patients with chronic myeloid leukaemia: a modelling study." Lancet Haematol 6(8): e398-e408.

[25] Abruzzese, E., et al. (2020). "Tyrosine kinase inhibitors and pregnancy in chronic myeloid leukemia: opinion, evidence, and recommendations." Ther Adv Hematol 11: 2040620720966120.

[26] Reff, M. J., et al. (2020). "Front-line use of tyrosine kinase inhibitors in chronic phase chronic myeloid leukemia: Practice considerations." J Oncol Pharm Pract 26(1): 156-174.

[27] Chan, O., et al. (2020). "Side-effects profile and outcomes of ponatinib in the treatment of chronic myeloid leukemia." Blood Adv 4(3): 530-538.

[28] Mohanavelu, P., et al. (2021). "Meta-Analysis of Gastrointestinal Adverse Events from Tyrosine Kinase Inhibitors for Chronic Myeloid Leukemia." Cancers (Basel) 13(7).

[29] Kantarjian, H., et al. (2012). "Improved survival in chronic myeloid leukemia since the introduction of imatinib therapy: a single-institution historical experience." Blood 119(9): 1981-1987.

[30] Steegmann, J. L., et al. (2016). "European LeukemiaNet recommendations for the management and avoidance of adverse events of treatment in chronic myeloid leukaemia." Leukemia 30(8): 1648-1671.

[31] Hughes, T. and D. White (2013). "Which TKI? An embarrassment of riches for chronic myeloid leukemia patients." Hematology Am Soc Hematol Educ Program 2013: 168-175.

[32] Sedler, A. R. and C. S. Mitchell (2019). "SemNet: Using Local Features to Navigate the Biomedical Concept Graph." Front Bioeng Biotechnol 7: 156.

[33] Kirkpatrick, A., et al. (2022). "Optimizations for Computing Relatedness in Biomedical Heterogeneous Information Networks: SemNet 2.0." Big Data and Cognitive Computing 6(1): 27.

[34] Zhang, Y., et al. (2010). "Understanding bag-of-words model: A statistical framework." International Journal of Machine Learning and Cybernetics 1: 43-52.

[35] Macqueen, J. (1967). Some methods for classification and analysis of multivariate observations. In 5-th Berkeley Symposium on Mathematical Statistics and Probability: 281--297.

[36] Pedregosa, F. a. V., G. and Gramfort, A. and Michel, V., et al. (2011). "Scikit-learn: Machine Learning in {P}ython." Journal of Machine Learning Research 12: 2825--2830.

[37] McCoy, K., et al. (2021). "Biomedical Text Link Prediction for Drug Discovery: A Case Study with COVID-19." Pharmaceutics 13(6).

[38] Mitchell, C. S., et al. (2015). "Undergraduate Biocuration: Developing Tomorrow's Researchers While Mining Today's Data." J Undergrad Neurosci Educ 14(1): A56-65.

[39] Gitelman, S. E., et al. (2021). "Imatinib therapy for patients with recent-onset type 1 diabetes: a multicentre, randomised, double-blind, placebo-controlled, phase 2 trial." Lancet Diabetes Endocrinol 9(8): 502-514.

[40] Fountas, A., et al. (2015). "Tyrosine Kinase Inhibitors and Diabetes: A Novel Treatment Paradigm?" Trends Endocrinol Metab 26(11): 643-656.

[41] Valent, P., et al. (2012). "Nilotinib as frontline and second-line therapy in chronic myeloid leukemia: open questions." Crit Rev Oncol Hematol 82(3): 370-377.

[42] Samaha, M. M., et al. (2019). "Nilotinib enhances β-islets integrity and secretory functions in a rat model of STZ-induced diabetes mellitus." Eur J Pharmacol 860: 172569.

[43] Elsherbiny, N. M., et al. (2015). "Amelioration of experimentally induced diabetic nephropathy and renal damage by nilotinib." J Physiol Biochem 71(4): 635-648.

[44] Roden, M. (2016). "[Diabetes mellitus: definition, classification and diagnosis]." Wien Klin Wochenschr 128 Suppl 2: S37-40.

[45] Kim, D. S., et al. (2016). "Use of deferasirox, an iron chelator, to overcome imatinib resistance of chronic myeloid leukemia cells." Korean J Intern Med 31(2): 357-366.

[46] Medeiros, B. C., et al. (2018). "Cardiovascular, pulmonary, and metabolic toxicities complicating tyrosine kinase inhibitor therapy in chronic myeloid leukemia: Strategies for monitoring, detecting, and managing." Blood Rev 32(4): 289-299.

[47] Jain, P., et al. (2019). "Analysis of cardiovascular and arteriothrombotic adverse events in chronic-phase CML patients after frontline TKIs." Blood Adv 3(6): 851-861.

[48] Mauro, M. J. (2021). "Lifelong TKI therapy: how to manage cardiovascular and other risks." Hematology Am Soc Hematol Educ Program 2021(1): 113-121.

[49] Ahmadieh, H. and I. Salti (2013). "Tyrosine kinase inhibitors induced thyroid dysfunction: a review of its incidence, pathophysiology, clinical relevance, and treatment." Biomed Res Int 2013: 725410.

[50] Patel, S., et al. (2020). "BCR-ABL1 tyrosine kinase inhibitor-associated thyroid dysfunction: A review of cases reported to the FDA Adverse Event Reporting System and published in the literature." Am J Hematol 95(12): E332-e335.

[51] Lim, D. J., et al. (2007). "Pancytopenia and secondary myelofibrosis could be induced by primary hyperparathyroidism." Int J Lab Hematol 29(6): 464-468.

[52] O'Sullivan, S., et al. (2011). "The skeletal effects of the tyrosine kinase inhibitor nilotinib." Bone 49(2): 281-289.

[53] Soliman, A. T., et al. (2017). "Chronic anemia and thyroid function." Acta Biomed 88(1): 119-127.

[54] Wang, J. and S. Zhuang (2017). "Src family kinases in chronic kidney disease." Am J Physiol Renal Physiol 313(3): F721-f728.

[55] Iyoda, M., et al. (2011). "Nilotinib attenuates renal injury and prolongs survival in chronic kidney disease." J Am Soc Nephrol 22(8): 1486-1496.

[56] Sasaki, K., et al. (2016). "Clinical Safety and Efficacy of Nilotinib or Dasatinib in Patients With Newly Diagnosed Chronic-Phase Chronic Myelogenous Leukemia and Pre-Existing Liver and/or Renal Dysfunction." Clin Lymphoma Myeloma Leuk 16(3): 152-162.

[57] Tong, W. G., et al. (2010). "Imatinib front-line therapy is safe and effective in patients with chronic myelogenous leukemia with pre-existing liver and/or renal dysfunction." Cancer 116(13): 3152-3159.

[58] ElShaer, A., et al. (2021). "Dasatinib-Induced Nephrotic Syndrome: A Case Report." Cureus 13(12): e20330.

[59] Masiello, D., et al. (2009). "The occurrence and management of fluid retention associated with TKI therapy in CML, with a focus on dasatinib." J Hematol Oncol 2: 46.

[60] Hailan, Y. M., et al. (2021). "Dasatinib-Induced Pleural and Pericardial Effusions." Cureus 13(10): e19024.

[61] Gunnarsson, N., et al. (2016). "Increased prevalence of prior malignancies and autoimmune diseases in patients diagnosed with chronic myeloid leukemia." Leukemia 30(7): 1562-1567.

[62] Söderlund, S., et al. (2020). "Plasma proteomics of biomarkers for inflammation or cancer cannot predict relapse in chronic myeloid leukaemia patients stopping tyrosine kinase inhibitor therapy." Leuk Res 90: 106310.

[63] Oshima, N., et al. (2021). "Differential gene expression analysis of dasatinib-induced colitis in a patient with chronic myeloid leukemia followed for 3 years: a case report." BMC Gastroenterol 21(1): 19.

[64] Bocchia, M., et al. (2016). "Genetic predisposition and induced pro-inflammatory/pro-oxidative status may play a role in increased atherothrombotic events in nilotinib treated chronic myeloid leukemia patients." Oncotarget 7(44): 72311-72321.

[65] Demirsoy, E. T., et al. (2018). "Dasatinib-induced immune mediated-thrombotic thrombocytopenic purpura." Transfus Apher Sci 57(2): 222-224.

[66] Wang, Z., et al. (2021). "Comparison of Hepatotoxicity Associated With New BCR-ABL Tyrosine Kinase Inhibitors vs Imatinib Among Patients With Chronic Myeloid Leukemia: A Systematic Review and Meta-analysis." JAMA Netw Open 4(7): e2120165.

[67] Harbaum, L., et al. (2014). "Treatment with dasatinib for chronic myeloid leukemia following imatinib-induced hepatotoxicity." Int J Hematol 99(1): 91-94.

[68] Kumar, V., et al. (2022). "Transport and metabolism of tyrosine kinase inhibitors associated with chronic myeloid leukemia therapy: a review." Mol Cell Biochem 477(4): 1261-1279.

[69] Nesr, G., et al. (2019). "The influence of salivary amylase on total amylase elevation in CML patients treated with TKI therapy: a case series of 3 patients." Leuk Lymphoma 60(13): 3333-3334.

[70] Shamroe, C. L. and J. M. Comeau (2013). "Ponatinib: A new tyrosine kinase inhibitor for the treatment of chronic myeloid leukemia and Philadelphia chromosome-positive acute lymphoblastic leukemia." Ann Pharmacother 47(11): 1540-1546.

[71] Sasi, S., et al. (2022). "Myasthenia Gravis and Myeloproliferative Neoplasms - Mere Association or Paraneoplastic Neurologic Syndrome: A Mini-Review." Acta Biomed 92(6): e2021437.

[72] Ishida, T., et al. (2018). "Dasatinib-associated reversible demyelinating peripheral polyneuropathy in a case of chronic myeloid leukemia." Int J Hematol 107(3): 373-377.

[73] Kavanagh, S., et al. (2018). "Peripheral neuropathy associated with imatinib therapy for chronic myeloid leukemia." Blood Res 53(2): 172-174.

[74] Inoue, H., et al. (2020). "Dasatinib-induced Reversible Demyelinating Peripheral Neuropathy and Successful Conversion to Nilotinib in Chronic Myelogenous Leukemia." Intern Med 59(19): 2419-2421.

[75] Chan, J., et al. (2019). "Nilotinib-Induced Dystonia and Cognitive Deficits in a Neurologically Normal Patient with Chronic Myeloid Leukemia." Case Rep Neurol Med 2019: 3679319.

[76] Chamoun, K., et al. (2020). "A case of neurocognitive deficit strongly related to dasatinib therapy." Hematol Transfus Cell Ther 42(1): 80-82.

[77] Chow, E. J., et al. (2021). "Adverse events among chronic myelogenous leukemia patients treated with tyrosine kinase inhibitors: a real-world analysis of health plan enrollees." Leuk Lymphoma 62(5): 1203-1210.

[78] Yu, L., et al. (2019). "Variables associated with patient-reported symptoms in persons with chronic phase chronic myeloid leukemia receiving tyrosine kinase inhibitor therapy." Medicine (Baltimore) 98(48): e18079.

[79] Breccia, M., et al. (2020). "Tyrosine kinase inhibitor discontinuation in the management of chronic myeloid leukemia: a critical review of the current practice." Expert Rev Hematol 13(12): 1311-1318.

[80] Duman, B. B., et al. (2012). "Secondary malignancy after imatinib therapy: eight cases and review of the literature." Leuk Lymphoma 53(9): 1706-1708.

[81] Lee, C. H., et al. (2020). "Occurrence of lymphoplasmacytic lymphoma in a chronic myeloid leukemia patient following long-term treatment with tyrosine kinase inhibitors: A case report." Medicine (Baltimore) 99(19): e19962.

[82] Miranda, M. B., et al. (2016). "Secondary malignancies in chronic myeloid leukemia patients after imatinib-based treatment: long-term observation in CML Study IV." Leukemia 30(6): 1255-1262.

[83] Kumar, V., et al. (2018). "An observational study on risk of secondary cancers in chronic myeloid leukemia patients in the TKI era in the United States." PeerJ 6: e4342.

[84] Narra, R. K., et al. (2017). "Chronic Myeloid Leukemia-the Promise of Tyrosine Kinase Inhibitor Discontinuation." Curr Hematol Malig Rep 12(5): 415-423.

[85] Rindflesch, T. C., et al. (2003). "Semantic relations asserting the etiology of genetic diseases." AMIA Annu Symp Proc 2003: 554-558.

[86] Roque, F. S., et al. (2011). "Using Electronic Patient Records to Discover Disease Correlations and Stratify Patient Cohorts." PLOS Computational Biology 7(8): e1002141.

[87] De Marchi, F., et al. (2017). "How could patient reported outcomes improve patient management in chronic myeloid leukemia?" Expert Review of Hematology 10(1): 9-14.

[88] Uramoto, N., et al. (2004). "A text-mining system for knowledge discovery from biomedical documents." IBM Systems Journal 43(3): 516-533.