Deliverable 4 – Health Analytics Project Submission

Hui Xia
hxia40@gatech.edu

CS-6440 – Introduction to Health Informatics

Georgia Institute of Technology

Mar 7th, 2020

INTRODUCTION

Thrombocythaemia represent in various myeloproliferative disorders, including chronic myeloid leukaemia, agnogenic myeloid metaplasia, and essential thrombocythaemia (ET) (Sanchez & Ewton, 2006), which is characterized by platelet production increase, and elevated platelet counts (Spencer & Brogden, 1994). Thrombocythaemia is usually seen in elderly patients, and could be dangerous, as the increased number of platelets might cause dangerous complications such as systemic thrombosis (Xiong et al., 2017). Thus, there is an essential need for drug development to limit the patient's platelet count in normal range, in order to minimize the risk of the potential cardiovascular adverse effects.

Anagrelide is an orally active quinazolin, which is known for causing thrombocytopenia, and has thus been evaluated for treating thrombocythaemia (Spencer & Brogden, 1994). Anagrelide is approved by FDA in 1997 in order to treat essential thrombocythemia, under then commercial name of AGRYLIN by Roberts Pharmaceutical. In the associated clinical trial, among a total of 551 patients, the most frequently reported adverse reactions were headache, palpitations, diarrhea, and abdominal pain (Solberg Jr et al., 1997). Patient thrombosis data was not included in this clinical trial. Hydroxyurea is a time-tested older drug, which has been widely used to treat sickle cell disease in 1980s (although FDA has not approved its usage until 1998) (Okam, Shaykevich, Ebert, Zaslavsky, & Ayanian, 2014). Compared with newer drugs, hydroxyurea is also cost-effective. The cost for hydroxyurea oral capsule 500 mg is around $78 for a supply of 100 capsules, while the cost for Anagrelide oral capsule 0.5 mg is around $273 for a supply of 100 capsules (Drugs.com, 2019).

One of the driving forces for the development of Anagrelide is that Hydroxyurea has been reported for being related with cancer risk. For example, Nand et. al. reported that hydroxyurea has 1–5.9% risk of causing leukemic transformation (Nand, Stock, Godwin, & Fisher, 1996). Hanft et. al. further suggested that in vivo hydroxyurea exposure could cause acquired DNA mutations (Hanft et al., 2000). Thus, whether

such suggestions against hydroxyurea is reasonable, and whether the alternatives thereof, such as Anagrelide, could perform better in cancer risk, will need to be investigated.

In this report, we investigate such clinical question: Do patients that have been treated with Anagrelide have a higher risk of cancer, compared with a more conventional alternative, Hydroxyurea? We conduct population-based cohort studies with claims and datasets using both Atlas and R package.

METHODS

Data Access

The codes for Anagrelide is found from the ICD-10-CM code, which used to specify a diagnosis of long term (current) use of antithrombotics/antiplatelets, which is Z79.02 ("2020 ICD-10-CM Diagnosis Code Z79.02," 2020). Using this ICD code, we found the concept ID for the corresponding drug, Anagrelide, on Athena. The code in anatomical therapeutic chemical (ATC) drug classification system for hydroxyurea is L01XX05, which can be found is this reference ("DRUG: Hydroxyurea,"). Using the ATC code, we then found related drug in Athena, followed by choosing related concepts in Atlas. As it has been reported that hydroxyurea could cause acquired DNA mutations (Hanft et al., 2000), to investigate on how much cancerous risk is associated with hydroxyurea, all cancer-related concepts on Atlas have been selected.

Cohort Definition

This cohort study was designed to learn the relationship between long-term usage of Anagrelide or Hydroxyurea, against the occurrence of cancer. Although it has been reported that patients are diagnosed with cancer after being exposed to hydroxyurea from 0 to 21 months (Hanft et al., 2000), it is premature to assume exposure to hydroxyurea could immediately cause cancer – after all, the mutation risk (if significant) of hydroxyurea exposure will take time to be effective. Thus, for the outcome cohort, we defined the observation window to be 3 months – 3 years before index start date, to exclude the patients that are diagnosed with cancer after hydroxyurea exposure.

Links to the said cohorts are:

- [hxia40] patients taking Anagrelide

  http://gt-health-analytics-1.us-east-1.elasticbeanstalk.com/#/cohortdefinition/446
- [hxia40] patients taking Hydroxyurea

  http://gt-health-analytics-1.us-east-1.elasticbeanstalk.com/#/cohortdefinition/449

– [hxia40] cancer patients exposed to either drug

  http://gt-health-analytics-1.us-east-1.elasticbeanstalk.com/#/cohortdefinition/452

Characterization and Incidence Rates

To access the incidence rates for the above-defined cohorts, characterization analyzes are performed at the characterizations section of Atlas. Feature analyses were performed on "condition eras of any time prior", "demographic age group", "demographics gender", and "drug era any time prior". The characterization is performed on both the CMSDESynPUF100k dataset and the CMSDESynPUF23m dataset. The number of records in the 100k and 23m dataset are listed in **Table 1.** Incidence rate for a given cohort is calculated as "the number of patients in the target cohort who experienced the outcome cohort during the time at risk period" divided by "the number of patients in the comparator cohort who experienced the outcome cohort during the time at risk period". The "start of the time at risk" was defined as one day after cohort start, and the "end of the time at risk "was defined as the cohort end date.

R analysis

The cohorts are analyzed using an open-source R package: OHDSI CohortMethod ("OHDSI/CohortMethod:New-user cohort method with large scale propensity and outcome models," 2020) on the CMSDESynPUF100k dataset. Propensity score was used to balance between the target and comparator cohorts. An expansive propensity score model, including all available covariates, was used. Propensity score adjustment was performed.

RESULTS AND DISCUSSION

Population Characteristics

The number of records found from Atlas is listed in **Table 1**. Compared with the number of records from the 23m dataset, the number of records from the 100k dataset is limited. To evaluate if the records from the 100k dataset could serve as a reasonable representative for the records from the 23m dataset, the characteristics of the target and the comparator cohorts in CMSDESynPUF100k and CMSDESynPUF23m datasets are compared. As shown in **Tables 2-3**, the composition for both the target (Anagrelide) and the comparator (Hydroxyurea) cohorts are of comparable deposition among both datasets. Thus, we will use the 100k dataset to perform R analysis, as the cohorts built using the 100k dataset are representative for a broader population.

Table 1. number of records for concepts of Anagrelide, Hydroxyurea, and cancer within their respective cohorts.

Cohort	Patients taking Anagrelide	Patients taking Hydroxyurea	Cancer patients exposed to either drug
CMSDESynPUF100k	638	699	172
CMSDESynPUF23m	20,392	15,774	4,801

Table 2. Cohort characteristics (a) target cohort and (b) comparator cohort in the CMSDESynPUF100k dataset (partial).

a

Covariate	Explore	Concept ID	[hxia40] patients taking Anagrelide	
			Count	Pct
anagrelide	Explore	1381253	638	100.00%
Acetaminophen	Explore	1125315	427	66.93%
Hydrochlorothiazide	Explore	974166	415	65.05%
Type 2 diabetes mellitus	Explore	201826	402	63.01%
FEMALE	N/A	8532	390	61.13%
Simvastatin	Explore	1539403	368	57.68%
levothyroxine	Explore	1501700	361	56.58%
Lisinopril	Explore	1308216	313	49.06%
Pure hypercholesterolemia	Explore	437827	310	48.59%
Oxygen	Explore	19025274	300	47.02%

b

Covariate	Explore	Concept ID	[hxia40] patients taking Hydroxyurea	
			Count	Pct
hydroxyurea	Explore	1377141	699	100.00%
Acetaminophen	Explore	1125315	469	67.10%
FEMALE	N/A	8532	455	65.09%
Hydrochlorothiazide	Explore	974166	442	63.23%
Type 2 diabetes mellitus	Explore	201826	414	59.23%
levothyroxine	Explore	1501700	389	55.65%
Simvastatin	Explore	1539403	366	52.36%
Lisinopril	Explore	1308216	327	46.78%
Oxygen	Explore	19025274	325	46.49%
Lovastatin	Explore	1592085	321	45.92%

Table 3. Cohort characteristics (a) target cohort and (b) comparator cohort in the CMSDESynPUF23m dataset (partial).

a

Covariate	Explore	Concept ID	[hxia40] patients taking Anagrelide	
			Count	Pct
anagrelide	Explore	1381253	15,942	78.18%
Type 2 diabetes mellitus	Explore	201826	13,222	64.84%
FEMALE	N/A	8532	12,400	60.81%
Acetaminophen	Explore	1125315	12,170	59.68%
Hydrochlorothiazide	Explore	974166	12,066	59.17%
levothyroxine	Explore	1501700	10,211	50.07%
Pure hypercholesterolemia	Explore	437827	9,903	48.56%
Simvastatin	Explore	1539403	9,867	48.39%
Atrial fibrillation	Explore	313217	9,173	44.98%
Hypothyroidism	Explore	140673	8,989	44.08%

b

Covariate	Explore	Concept ID	[hxia40] patients taking Hydroxyurea	
			Count	Pct
hydroxyurea	Explore	1377141	15,774	100.00%
Acetaminophen	Explore	1125315	10,180	64.54%
FEMALE	N/A	8532	10,055	63.74%
Hydrochlorothiazide	Explore	974166	10,044	63.67%
Type 2 diabetes mellitus	Explore	201826	9,439	59.84%
levothyroxine	Explore	1501700	8,735	55.38%
Simvastatin	Explore	1539403	8,009	50.77%
Pure hypercholesterolemia	Explore	437827	7,173	45.47%
Lisinopril	Explore	1308216	7,165	45.42%
Lovastatin	Explore	1592085	7,023	44.52%

Incidence Rates with Atlas

Among patients taking Anagrelide, 125.39/1,000 patients in the 100k dataset and 129.18/1000 patients in the 23m dataset were diagnosed with cancer. The incidence rates per 1k years are 103.49 and 102.95, respectively in the 100k and 23m datasets. Among patients taking Hydroxyurea, 133.05/1000 in the 100k dataset and 138.92/1000 patients in the 23m dataset were diagnosed with cancer. The incidence rates per thousand persons are 103.79 and 108.68, respectively in the 100k and 23m datasets. Based on this data lone, compared with the patient cohort taking Hydroxyurea, the patient cohort taking Anagrelide has a significantly lower chance of acquiring cancer.

Population selection for the R analysis

Among patients taking Anagrelide, 125.39/1,000 patients in the 100k dataset and As we have discussed above, the analysis with R package was performed using the CMSDESynPUF100kn dataset. **Figure 1** demonstrates how the study subjects are selected. Among both of the cohorts, 10 patients in the target cohort and 30 patients in the comparator cohort have been diagnosed with hypertension. Note that the patient groups of the Anagrelide and Hydroxyurea are not perfectly matching: the former is used majorly on essential thrombocythaemia (ET) patients, while the latter are widely used for all kinds of thrombocythemia (including ET), as well as for sickle cell disease patients. In the U.S., nearly all sickle cell disease patients are African Americans (Hassell, 2010), which in general are easier to be affected by hypertension (Kramer et al., 2004). Thus, to maintain a fair comparation between the target and the comparator cohorts, these hypertension patients are removed from the comparison. After filtering, a total number of 3,358 patients from the target cohort and 2,748 patients from the comparator cohort are subjected to the comparison study.

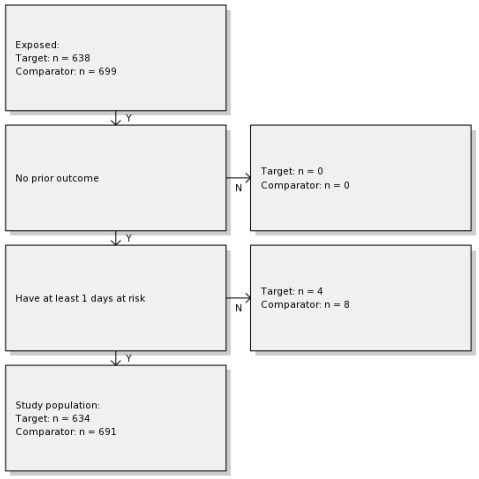

Figure 1. Attrition diagram demonstrating the number of subjects in the target and the comparator cohorts among the process of the analysis.

Table 4 presents the major population characteristics of the target and the comparator cohorts before and after propensity score adjustment, which suggest that the target and the comparator cohorts share similar composition on all age groups, gender, and race. **Figure 2** plots the preference score distributions for both the target ad the comparator cohorts. The preference score is a transformation of the propensity score that adjusts for differences in the sizes of the two cohorts. The high overlap between two populations indicates the high similarity of subjects in the two cohorts in terms of their predicted probability of taking one medication over the other. **Figure 3** demonstrates the standard difference for all input features of the propensity score model. The standard difference that are less than 0.1 is considered as well-adjusted. The number of features that have standard difference value that are higher than 0.1 are significantly reduced after the propensity score adjustment.

Table 4. Major population characteristics before and after propensity score adjustment (partial)

Characteristic	Before PS adjustment			After PS adjustment		
	Target	Comparator		Target	Comparator	
	%	%	Std. diff	%	%	Std. diff
Age group						
25-29	<0.8	0.7	-0.03	<0.8	0.8	-0.05
30-34	0.8	0.9	-0.01	<0.8	0.9	-0.01
35-39	2.4	1.0	0.10	2.3	1.0	0.10
40-44	0.9	1.6	-0.06	0.9	1.7	-0.07
45-49	3.6	2.7	0.05	4.3	2.1	0.12
55-59	4.2	2.9	0.07	4.5	2.9	0.09
60-64	5.2	4.7	0.02	5.8	4.8	0.05
65-69	16.5	15.3	0.03	15.7	15.6	0.00
70-74	16.6	19.7	-0.08	17.1	19.1	-0.05
75-79	15.8	16.9	-0.03	15.5	16.9	-0.04
80-84	13.2	14.4	-0.04	12.7	14.6	-0.05
85-89	9.2	9.2	0.00	9.2	9.7	-0.02
90-94	4.1	4.1	0.00	4.1	3.8	0.01
95-99	3.1	1.9	0.08	2.9	1.9	0.06
00-04	0.8	0.9	-0.01	<0.8	0.9	-0.03
Gender: female	61.1	65.1	-0.08	61.2	65.8	-0.10
Race						
race = Black or African American	10.7	9.7	0.03	10.4	9.7	0.02
race = White	81.0	83.4	-0.06	81.2	82.8	-0.04
Ethnicity						

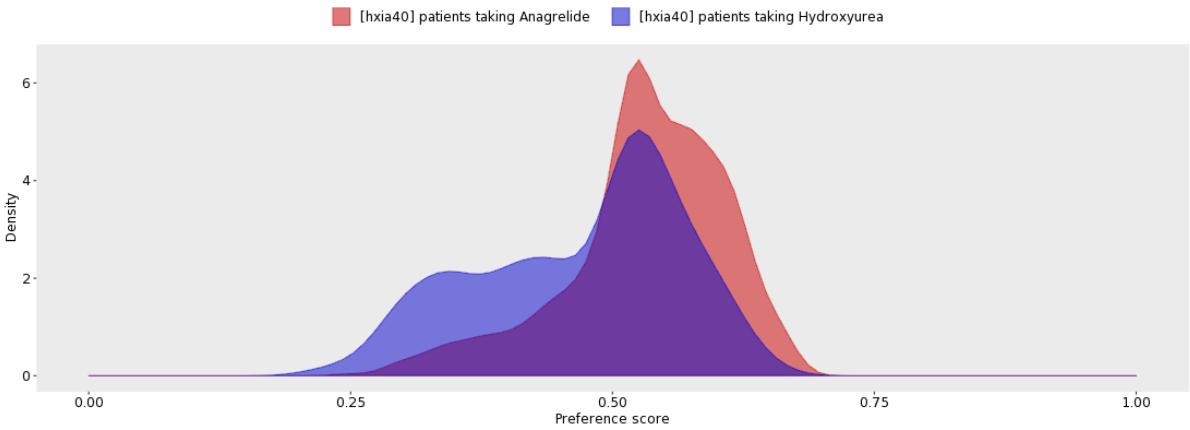

Figure 2. Preference score distribution for the target and the comparator cohorts.

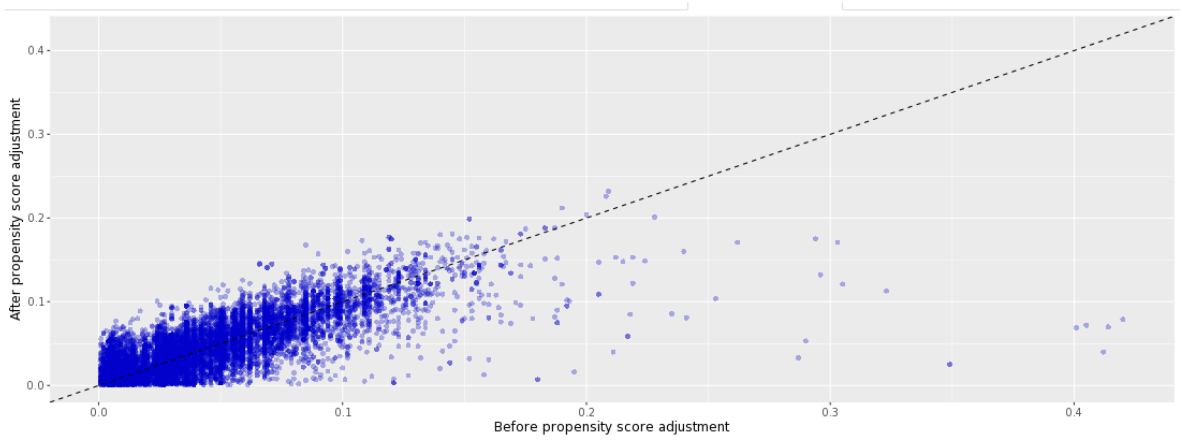

Figure 3. Covariate balance before and after propensity score adjustment.

Table 5 lists the numbers of subjects, follow-up time, numbers of outcome events and event incidence rates per 1,000 patient years. Compared with the comparator cohort, the target cohort has less follow-up time and numbers of outcome events, which is similar to the incidence rates derived from the Atlas, which we have shown above. **Table 6** demonstrates the time to be diagnosed as cancer for patients in both cohorts. No major difference between the target and the comparator. **Figure 4** demonstrates the Kaplan-Meier survival curves for both cohorts. While in **Table 5** and the results from Atlas that the comparator cohort is related with higher incidence rates, the survival probability of the target cohort and the comparator cohort is similar over time.

Table 5. Number of subjects, follow-up time (in years), number of outcome events, and event incidence rate per 1,000 patient years (PY) in the target and the comparator cohorts after the prosperity score adjustment, and the minimum detectable relative risk (MDRR).

Target subjects	Comparator subjects	Target years	Comparator years	Target events	Comparator events	Target IR (per 1,000 PY)	Comparator IR (per 1,000 PY)	MDRR
634	691	775	898	80	93	103.15	103.56	1.53

Table 6. Time (days) at risk distribution expressed as minimum (min), 25th percentile (P25), 75th percentile (P75), and maximum (Max) in the target and the comparator cohorts after propensity score adjustment.

Cohort	Min	P10	P25	Median	P75	P90	Max
Target	3	115	211	404	788	850	1,078
Comparator	4	124	230	469	799	874	1,089

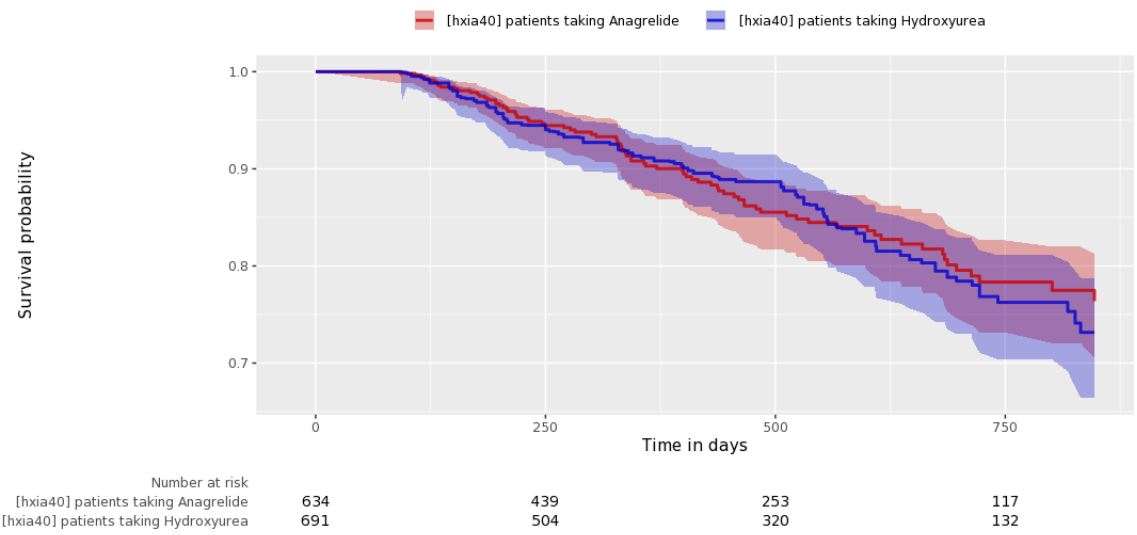

Figure 4. Kaplan Meier plot on the patient survival rate as function of time. The data used are adjusted using the propensity score. The target cohort curve is plotted using the actual observed survival. The comparator curve is plotted using the reweighted data, to approximate the counterfactual of what the target cohort survival would look lie had the target cohort been exposed to the comparator instead. The shaded area denotes 95 % confidence interval.

One of the limitations of this study is that it would be more meaningful to compare the chance of leukemia, instead of all cancers, between the target and the comparator cohorts. After all, the hydroxyurea is believed to promote leukemia, rather than all cancers. Re-performing this study on a larger dataset that contains enough data on various kinds of leukemia patients will be needed. Furthermore, the possibility of other cohort-related factors for cancer development cannot be fully excluded. For example, hydroxyurea is a well-used drug for the treatment of sickle-cell disease, while Anagrelide is not used to treat sickle-cell disease. How sickle-cell disease gene affect the chance for a person to develop cancer or leukemia still remains uncertain.

In this study, the comparator cohort is defined as patients that are exposed to hydroxyurea. However, a more accurate method to investigate the risk of Anagrelide and hydroxyurea should also involve a cohort using placebo. While there are reports suggest hydroxyurea might be related with leukemia, here we see can neither find significantly higher chance of cancer, nor significantly higher survival rate in either of the cohort. With our comparing with a 'placebo cohort', no conclusion on whether the target or the comparator cohort can increase the risk of cancer (let alone leukemia) can be made.

Reference

2020 ICD-10-CM Diagnosis Code Z79.02. (2020). Retrieved from https://www.icd10data.com/ICD10CM/Codes/Z00-Z99/Z77-Z99/Z79-/Z79.02

DRUG: Hydroxyurea. Retrieved from https://www.genome.jp/dbget-bin/www_bget?dr:D00341

Drugs.com. (2019). Retrieved from https://www.drugs.com/price-guide/

Hanft, V. N., Fruchtman, S. R., Pickens, C. V., Rosse, W. F., Howard, T. A., & Ware, R. E. (2000). Acquired DNA mutations associated with in vivo hydroxyurea exposure. *Blood, 95*(11), 3589-3593.

Hassell, K. L. (2010). Population estimates of sickle cell disease in the US. *American journal of preventive medicine, 38*(4), S512-S521.

Kramer, H., Han, C., Post, W., Goff, D., Diez-Roux, A., Cooper, R., . . . Shea, S. (2004). Racial/ethnic differences in hypertension and hypertension treatment and control in the multi-ethnic study of atherosclerosis (MESA). *American journal of hypertension, 17*(10), 963-970.

Nand, S., Stock, W., Godwin, J., & Fisher, S. G. (1996). Leukemogenic risk of hydroxyurea therapy in polycythemia vera, essential thrombocythemia, and myeloid metaplasia with myelofibrosis. *American journal of hematology, 52*(1), 42-46.

OHDSI/CohortMethod:New-user cohort method with large scale propensity and outcome models. (2020). Retrieved from https://github.com/OHDSI/

Okam, M. M., Shaykevich, S., Ebert, B. L., Zaslavsky, A. M., & Ayanian, J. Z. (2014). National Trends in Hospitalizations for Sickle Cell Disease in the United States following the FDA Approval of Hydroxyurea, 1998 to 2008. *Medical care, 52*(7), 612.

Sanchez, S., & Ewton, A. (2006). Essential thrombocythemia: a review of diagnostic and pathologic features. *Archives of pathology laboratory medicine, 130*(8), 1144-1150.

Solberg Jr, L. A., Tefferi, A., Oles, K. J., Tarach, J. S., Petitt, R. M., Forstrom, L. A., & Silverstein, M. N. (1997). The effects of anagrelide on human megakaryocytopoiesis. *British journal of haematology, 99*(1), 174-180.

Spencer, C. M., & Brogden, R. N. (1994). Anagrelide. *Drugs, 47*(5), 809-822.

Xiong, N., Gao, W., Pan, J., Luo, X., Shi, H., & Li, J. (2017). Essential thrombocythemia presenting as acute coronary syndrome: case reports and literature review. *Journal of thrombosis thrombolysis, 44*(1), 57-62.