

Semester 1, 2023

**E-COMpliance:**  
**Detecting Non-Compliant Therapeutic Goods Selling on**  
**E-commerce Websites**

9785: ITS Capstone Project

Project ID: 2023-S1-34  
Group ID: 9785-23-06

Table of Contents

[Table of Contents 1](#_Toc134047852)

[Abstract 4](#_Toc134047853)

[Introduction 4](#_Toc134047854)

[Background Information: 4](#_Toc134047855)

[Therapeutic Goods: 4](#_Toc134047856)

[Non-compliant products: 5](#_Toc134047857)

[Poisonous Ingredients 5](#_Toc134047858)

[False Therapeutic Claims: 5](#_Toc134047859)

[Importance of Detection: 5](#_Toc134047860)

[Maintain Customer trust: 5](#_Toc134047861)

[Protect public health: 5](#_Toc134047862)

[Prevent dishonest behaviour: 5](#_Toc134047863)

[Maintain regulatory compliance: 6](#_Toc134047864)

[Scope: 6](#_Toc134047865)

[Original Scope: 6](#_Toc134047866)

[Revised Scope: 6](#_Toc134047867)

[Purpose of the project: 6](#_Toc134047868)

[Create: 6](#_Toc134047869)

[Protect: 6](#_Toc134047870)

[Develop: 6](#_Toc134047871)

[Literature Review 7](#_Toc134047872)

[Methodology 8](#_Toc134047873)

[Resources Needed: 8](#_Toc134047874)

[System Overview: 9](#_Toc134047875)

[External Setup: 10](#_Toc134047876)

[System Setup: 10](#_Toc134047877)

[Web Scraper: 10](#_Toc134047878)

[Compliance Checks: 14](#_Toc134047879)

[Approved ARTG Check: 15](#_Toc134047880)

[Poisonous Ingredient Check: 15](#_Toc134047881)

[Therapeutic Claim Checker: 15](#_Toc134047882)

[Final Compliance Determination: 15](#_Toc134047883)

[Output: 15](#_Toc134047884)

[Results: 16](#_Toc134047885)

[Compliance: 16](#_Toc134047886)

[Compliance Issues: 17](#_Toc134047887)

[Key Performance Indicators (KPI): 18](#_Toc134047888)

[Collection Rate: 18](#_Toc134047889)

[System Accuracy: 19](#_Toc134047890)

[Overall: 19](#_Toc134047891)

[Time Allocation 19](#_Toc134047892)

[Discussion: 20](#_Toc134047893)

[System Limitations: 20](#_Toc134047894)

[Accuracy: 20](#_Toc134047895)

[Functional: 21](#_Toc134047896)

[Project Takeaways: 21](#_Toc134047897)

[Insights Gained: 21](#_Toc134047898)

[Lessons Learnt: 22](#_Toc134047899)

[Outstanding Issues: 22](#_Toc134047900)

[ARTG Check Issues: 22](#_Toc134047901)

[Incomplete Product Infromation: 22](#_Toc134047902)

[Manufacturer Name Recognition: 22](#_Toc134047903)

[Risks Mitigated: 22](#_Toc134047904)

[Poor Communication: 22](#_Toc134047905)

[Missed Deadlines/Deliverables: 23](#_Toc134047906)

[Handover Materials: 23](#_Toc134047907)

[Future Improvements: 23](#_Toc134047908)

[Web Scraper Optimisation: 23](#_Toc134047909)

[OCR Pre-Processing: 23](#_Toc134047910)

[Alternate ARTG Number Allocation: 23](#_Toc134047911)

[Displays: 24](#_Toc134047912)

[Recommendations: 24](#_Toc134047913)

[Sponsor: 24](#_Toc134047914)

[External: 24](#_Toc134047915)

[Education and Awareness Campaigns: 24](#_Toc134047916)

[Regular Monitoring and Auditing: 24](#_Toc134047917)

[Collaboration between TGA and E-commerce platforms: 25](#_Toc134047918)

[Conclusion 25](#_Toc134047919)

[References: 26](#_Toc134047920)

# **Abstract**

The sale of non-compliant therapeutic products is a global issue. This project aimed to develop a system to detect non-compliant therapeutic goods being sold on E-commerce websites in Australia. The system will achieve this by utilising web scraping techniques to collect advertisement and product data from the Australian Amazon marketplace. This data would then be analysed for compliance with the Therapeutic Goods Administration (TGA) regulations, including the detection of poisonous ingredients and unsubstantiated therapeutic claims. The goal for this system was to have a less than 5% false positive rate. Of the 8,241unique products that were collected, the system identified only 17 that were TGA compliant. This result was due to a lack of Australian Register of Therapeutic Goods (ARTG) numbers being connected to products. Although the system that we developed was not able to achieve this goal, it can act as a base for further development and refinement.

*Keywords:* non-compliant, Amazon, web scrape, e-commerce, therapeutic goods

# Introduction

The detection of non-compliance in the online sales of therapeutic products is a significant concern for governing agencies worldwide. While many E-commerce websites operate lawfully and in accordance with the rules established by their individual nations, there are still instances where non-compliant goods can slip through the cracks. This can occur because certain products may be approved for use in countries like the United States of America but may contain ingredients that are still deemed poisonous or in quantities that are not approved by the Therapeutic Goods Administration (TGA) of the Australian government.

It is crucial for the health and safety of the public that only products meeting Australia's high standard for therapeutic goods be available for sale in the country, as non-compliant products can cause serious harm and endanger public health. To ensure compliance, products must carry an Australian Register of Therapeutic Goods (ARTG) product code on their label, indicating that they meet the necessary criteria.

As E-commerce sites are large entities containing hundreds of thousands of products an automated approach must be considered when trying to identify products that could be non-compliant by TGA standards. That is where this project comes in, the goal of the project is to develop a system to automatically collect and determine which products being sold are compliant or not.

# Background Information:

## Therapeutic Goods:

According to the TGA website a therapeutic good can be identified as the following.

Therapeutic goods can comprise a broad range of things, such as bandages, pregnancy testing kits, herbal remedies, tissue grafts and paracetamol. They generally fall under three main categories:

* **Medicines** - including prescription, over-the-counter and complementary medicines, such as paracetamol and echinacea
* **Biologicals** - something made from or containing human cells or tissues, such as human stem cells or skin
* **Medical devices** - including instruments, implants and appliances, such as pacemakers and sterile bandages [9]

## Non-compliant products:

Non-compliant products can vary in severity as misleading claims on a placebo product that wouldn’t harm a human is nowhere near as bad as something that contains poisonous ingredients. For this project, a products non-compliance status will fall into one of three categories.

Poisonous Ingredients:  
 Products that contain ingredients that are listed in the ARTG’s schedule 4 and 8 poisonous ingredient lists will be classified as non-compliant. A product’s ingredients must all be safe to consume for a product to be complaint by TGA standards.

False Therapeutic Claims:  
 Products should make truthful, scientifically proven claims when advertising their therapeutic status, and those that make misleading or false therapeutic claims will be classified as non-compliant by the system. Products claiming to be “100% natural” when they are in fact not 100% natural or products making claims to cure incurable diseases will fall into the non-compliant classification.

ARTG Issues: If a product does not have an approved ARTG number according to the ARTG database or the product has no ARTG number displayed, it will be classified as non-compliant. No ARTG number means it is not eligible to be sold in Australia. All international products will fall into this classification if they are not approved for sale in Australia.

## Importance of Detection:

Detecting non-compliant therapeutic goods on E-commerce websites is a paramount concern, as it is instrumental in upholding the highest standards of public health, ensuring regulatory compliance, safeguarding against dishonest practices, and maintaining the trust of customers in online marketplaces.

Maintain Customer trust:   
 Customers expect the products they purchase to be safe and effective, and non-compliant products can erode that trust. By ensuring that only compliant therapeutic products are sold on E-commerce websites, they can maintain or improve customer trust.

Protect public health:   
 Non-compliant therapeutic goods can pose a serious risk to public health. They may be ineffective, contaminated, or even dangerous. Detecting and removing these products from E-commerce websites can help protect the public from harm.

Prevent dishonest behaviour:   
 Some sellers may knowingly sell non-compliant therapeutic goods to make a quick profit. Detecting and removing these products from E-commerce websites can help prevent this type of dishonest behaviour.

Maintain regulatory compliance:   
 Regulatory authorities have specific requirements for the sale of therapeutic goods, and E-commerce websites must comply with these requirements too. By detecting and removing non-compliant products, E-commerce websites can maintain regulatory compliance and avoid penalties or legal action.

# Scope:

Original Scope:  
 The original scope of the project was to create a system to detect non-compliant therapeutic goods being sold on 2 different Australian E-commerce websites. The system had to collect product and advertising information from the product pages and check its compliancy in terms of ARTG number, poisonous ingredients check and a therapeutic claim check. The system then needed to output all non-compliant products to a .csv file.

Revised Scope:  
 Due to time constraints and issues with the web scraping component of the project, our team and the sponsors decided it was best to reduce the project scope. Our initial plan was to analyze products sold on two E-commerce websites, but this was changed to just focus on one E-commerce website Amazon, which the team had already started on. The rest of the previous scope has remained the same.

# Purpose of the project:

Create:  
 The project aims to develop an automated system that can detect non-compliant therapeutic goods sold on E-commerce websites. This system will use web scrapping as the method of data collection, it will also analyse the data it collects, and identify non-compliant therapeutic products. This system will be faster and more efficient than manual detection methods.

Protect:  
 We aim to protect the health and safety of the public from non-compliant goods. Being able to identify non-compliant products means we can help prevent consumers from purchasing those products that may be ineffective, contain poisonous ingredients, or even dangerous.

Develop:   
 We wanted to develop awareness of the global issue of non-compliant therapeutic goods. The average consumer is unaware of the risks associated with unregulated products. This would involve educating consumers about the importance of purchasing therapeutic goods from reputable sources, understanding regulatory compliance.

# Literature Review

The legalisation of cannabis in Canada has led to the availability of numerous therapeutic products, which are exclusively used for medical or therapeutic purposes. While evidence supporting the medical efficacy of these products is limited, anecdotal evidence suggests that sellers are representing their products for medical use through direct medical claims or advice, which is not allowed under the Cannabis Act without Health Canada approval. This has raised concerns about the safety and efficacy of therapeutic products sold online.

In Canada, therapeutic products are regulated under the Food and Drugs Act, which governs the production, distribution, and sale of all drugs, including natural health products and prescription drugs. Under the Act, therapeutic products must meet certain requirements to be sold legally, including proper labelling, packaging, and dosage information, these are the same requirements that Australia and the TGA enforce too. However, some sellers are making unsubstantiated health claims to promote the medical usage of their products, which is not allowed.

Several studies have examined the detection of non-compliant therapeutic goods from E-commerce websites. These studies have used a variety of methods to detect non-compliant products, including web scraping, machine learning, and manual review. One study used web scraping to collect data on 1,161 CBD products from 88 E-commerce websites in the United States and found that 26% of the products made health claims that were not supported by scientific evidence (Bonaccorsi et al., 2020)

Another study used machine learning algorithms to analyse product descriptions of 2,165 CBD products from 70 E-commerce websites in Canada and found that 53.3% of products contained at least one medical or therapeutic claim (Chen et al., 2021). The most prevalent claims were for the treatment or management of pain, anxiety, and inflammation, with claims also found for serious and life-threatening illnesses such as multiple sclerosis, arthritis, cancer, Crohn’s disease, Parkinson’s disease, and HIV. The study also found that CBD products were framed to prospective customers through three distinct frames: a specific cure or treatment, a natural health product, and a product used in a certain way to achieve results.

A third study used manual review to assess the compliance of 1,317 CBD products sold online in the United States and found that 21.4% of products made unauthorized health claims (Bonnet et al., 2021). The study also found that many products contained inaccurate or incomplete information about dosage, ingredients, and labelling.

The final study we looked at was conducted in 2019 by researchers at the University of Sydney, who analysed the claims made by 157 CBD products sold online in Australia and found that over 40% of the products made therapeutic claims that were not supported by scientific evidence (Lau et al., 2019). The most common claims were for the treatment or management of pain, anxiety, and sleep disorders.

The findings of these studies suggest that there is a need for increased regulation and monitoring of therapeutic products sold online in Australia and around the globe.

The studies reviewed in this report highlight the unwavering prevalence of unauthorized health claims and inaccurate information about dosage and ingredients on E-commerce platforms selling therapeutic products. This underscores the need for increased regulation and monitoring of these products to ensure compliance with the law and protect consumers.

E-commerce platforms should be taking steps to prevent the sale of non-compliant therapeutic goods on their platform. Steps like increased internal monitoring and customer-facing reporting mechanisms for non-compliant products could be introduced.

The growing popularity of E-commerce in the past decades has been exponential and law-making and compliance efforts are lacking in the space. While these governing bodies catch up, a solution still needs to be developed to ensure the public’s safety in the interim and that where we feel a web-scraping therapeutic compliance system can fit in.

# Methodology

|  |
| --- |
|  |
| *Fig. 1. Iterative process diagram [4]* |

We decided on an iterative methodology to our system development. This decision was made due to the need for continuous improvement and refinement of individual components of the project as new information becomes available, or problems arise.

## Resources Needed:

|  |  |
| --- | --- |
| **Data Resources** | **Location** |
| ARTG number csv file | ARTG interactive database |
| Poison directory | TGA website |

|  |  |  |
| --- | --- | --- |
| **Python Package Name** | **Use in the project** | **Project Section** |
| Tkinter | Open the file directory for users to select their save location. | Setup |
| Os | Used to manipulate the files and create folders in setup as well as deleting product images once OCR was completed. | Setup and Web Scraper |
| RegEx | Used to compile regular expressions for names in data resources. | Setup and Compliance Checking |
| Selenium | Used for the web scraping and website interactions. | Web Scraper |
| Chrome Web driver | Used in conjunction with Selenium (this can be replaced with Firefox or other browsers). | Web Scraper |
| PyTesseract | Used for OCR to extract text from the product images that are collected. | Web Scraper |
| Pillow | Used to import and manipulate image files during the pre-processing phase of OCR. | Web Scraper |
| Time | I added wait times for the web scraper, so the page had time to load before collecting data and before moving onto the next product page. | Web Scraper |
| Random | Used for picking a random number from a range to variability to the wait times. | Web Scraper |
| Datetime | Used to collect the system date and time for accurate data collection and for the name generation of outputs. | Web Scraper and Output |
| Action Chains | An extension to the Selenium package that allowed for the mouse interaction to be used on the websites for opening and closing windows and hovering over product images. | Web Scraper |
| Math | Used to get the ceiling of a float number so we can accurately iterate through several pages without running into server errors form Amazon. | Web Scraper |
| Pandas | Reading and creating .csv files. | Setup, Compliance Checking and Output |

## System Overview:

|  |
| --- |
|  |
| *Fig. 2. System overview* |

The system can be broken down into three core components, web scraping, compliance checks, and output as seen in Fig. 2.

### External Setup:

There are a few packages that need extra installation to be conducted to install the foundations for the python packages to use, these are OpenCV and the Google Chrome web driver. This install process will need system admin approval to be installed which could be an issue in certain instances. This could also lead to security issues down the line if vulnerabilities are discovered.

### System Setup:

The system will ask the user to navigate to a file location for the system to create folders at. These folders will be used to save and delete files in as well as where all the collected web scrape .csv files, proof images and results will be saved to.

ARTG and Poison schedule lists are read into dataframes and lists as required. Lists are initialised that will collect the web scaped information, these lists include all the required product data collection points and as well as those required for the compliance checks (ARTG, poison and therapeutic claims) and system guesses.

Web Scraper:   
 The system can be broken down into 2 forms, keyword searches or blanket search. The system starts by checking if the search .txt file is empty or contains keyword searches, if it contains searches, it will do those and not do the blanket search in the ‘Vitamins and Supplements’ department on Amazon. Both these forms work functionally the same but vary wildly in time to complete. The keyword search can be completed in a matter of a couple hours while the blanket search will take many hours to complete.

#### Collecting Results Pages:

Once the system has chosen if tis using keywords or blanket searching, it input the keyword (or start in the ‘vitamins and supplements’ department) and it will display a ‘results’ page, see Fig. below.

|  |
| --- |
|  |
| *Fig. 3. Amazon product page example. [7]* |

On this ‘results’ page, the system will find and collect hyperlinks to the products using XPATH. As Amazon is a rather complex website, this method was the only real option to guarantee we collect just the specific information we needed.

Once the hyperlinks for the page are collected, they are output to a csv file for later use. This step is import as Amazon doesn’t allow a continuous stream of hyperlink visits back-to-back and are limited to around 64-80 (based on testing) until Amazon blocks the system and all future attempts are forcefully redirected to an Amazon Prime Video page.

After all results pages are collected and output to .csv files, the system will navigate to the next results page for the search. Finally, once the system has cycled through all the results pages it will move on to collecting the product information.

#### Collecting Product Information:

An Amazon product page looks like Fig. Below.

|  |
| --- |
|  |
| *Fig. 4. Amazon product page example [5]* |

The data collection from the Amazon product pages is comprised of many components each working in tandem to collect and process information. To save time, this section will be covered by collecting up similar processes and offering one explanation to cover the ones that work functionally the same.

##### Fixed Position:

The product page screenshot, date collected, keyword and name of the product are all simple and always in the same position, so they all can be collected easily with XPATH or based on operating system information / functionality. The product description is also always in the same spot but there is an optional section down the page that can be filled out too, so the description also checks for that extra section.

##### Variable Position:

There are a few pieces of information that can be in alternative positions in their list on the product pages depending on what information is relevant to the product. The Amazon Standard Identification Number (ASIN) and Manufacturer can be found in two different positions on their list while the ingredients section (if even listed at all) can be in one of three different positions.

Collecting the seller information is a similar problem to above but the location can be one of two different positions or can be multiple sellers hidden behind a window that requires being clicked on to open and load the information into the page. Although it was an easily implementable solution, it made it different from all the previously similar solutions to this problem.

#### Image Collection and Processing:

The image collection and processing were completed using a combination of packages (see resources table) and is an integral part to the project as without this ARTG numbers and therapeutic claims could not be extracted from the product images.

##### Collect Product Images:

Collecting the images used to advertise the products gave a nice insight into some web development practices used at Amazon as the high-resolution images were not immediately loaded when you load a product page, only low-definition thumbnails are loaded. To retrieve the high-resolution images required for OCR, the thumbnails on the product pages had to be hovered over so the information could be loaded into the browser on our (client) side. This was obviously due to the pages being able to load quicker and with less resources required on Amazon’s side. Once the hyperlinks to the high-resolution images were loaded into the page, it was as simple as collecting them using XPATH and following the links to download them.

##### Optical Character Recognition (OCR):

Once the product images were collected, they were all put the OCR to extract any text-based content out of them. The primary reason for this process was to extract an ARTG number from the image of the product and to ascertain whether any false or misleading claims were made using text that would fall through the cracks of normal text-processing methods.

Processing the images was as simple as making a calling a single method from the PyTesseract package. The results of each image’s OCR process were collected into a big paragraph for processing by other aspects of the system.

##### Searching OCR Results:

After the results are compiled from the OCR, the paragraph that it creates is searched for key identifiers like “AUST L” and “AUST R” which act as prefixes for the ARTG numbers for the products.

|  |
| --- |
|  |
| *Fig. 5. Berocca product image [5]* |

E.g., The image in Fig. 5. could have an OCR output of;

“ORANGE FLAVOUR 30 Effervescent Tablets Berocca ENERGY 12 essential vitamins and minerals Mental Energy Physical Energy\* NEW FORMULA AUST L 358753 Dietary Supplement”

AUST L 358753 could be extracted during OCR and would be assigned as the ARTG number for that product. After the OCR process is completed on all a products advertising images, they are deleted as to save space on the user’s system.

The iterative methodology lent itself well to web scraping as the process is something you want to fully run as few times as possible but testing it on as many different products with different webpage configurations meant that situational exceptions could be made easily to ensure accuracy in collecting data from the product pages.

#### Compiling Data:

Once all the product data has been collected it is stitched together into a single dataframe for compliance checking.

# Compliance Checks:

Running the compliance checks is the second major part of the system and is broken down into three important parts; ARTG check, poisonous ingredient check and a therapeutic claims check.

## Approved ARTG Check:

The check looks at each product in the dataframe of web scraped products and checks if it has a valid ARTG number. If the product information the system collected does not have an ARTG number assigned to it, it is marked as non-compliant. If the product has an ARTG number, the code checks its validity against the list of approved ARTG numbers. The only way to achieve a compliant ruling is if the product information contains a valid ARTG number and it appears on the approved ARTG list.

## Poisonous Ingredient Check:

This checks if any of the ingredients listed for a product are deemed poisonous by TGA standards. This is done by checking each ingredient to see if it matches with a poisonous ingredient from the list, we obtained from the TGA website. When the system finds there is a match, the product is assigned a non-compliant verdict. If a product does not have any ingredients listed, it will be assigned a non-compliant classification because it is better to air on the side of caution than approve a potentially dangerous product. A product is deemed compliant if there is a list of ingredients and they all don’t appear on the poisonous ingredient list.

## Therapeutic Claim Checker:

Finally, there is a check to identify any false or misleading therapeutic claims that may be present in any of the advertising material for the product. Using a dictionary of keywords and phrases that we developed,

This check can filter any therapeutic goods which contain false claims. The keywords and phrases are based off what the TGA believes constitutes a false therapeutic good claim, including curing incurable diseases and illnesses. Any products that have been filtered are classed as non-compliant therapeutic goods.

## Final Compliance Determination:

Based on the results of the previous steps, the products are categorized as compliant or non-compliant with regards to TGA regulations. A compliant verdict is given to a product that contains has the ‘compliant’ status given to the three of the ARTG, Poisonous ingredient and therapeutic claim checks. If even just one of those three checks are deemed ‘non-compliant’ then the product overall will be delivered the non-compliant verdict here.

# Output:

The output is the simplest aspect of the project as it is just compiling a .csv file of all the non-compliant products. Before exporting the .csv file, we decided to remove any duplicate product listings from the dataframe to ensure we were only outputting unique products in this step. An example of the output can be seen below in Fig. 6.

|  |
| --- |
|  |
| *Fig. 6. System output .csv file.* |

# Results:

Web scraper collected 11,186 products overall from the vitamins and supplements department on Amazon, of those, there was 8,241 unique products meaning we had 2,945 recurring products. The large number of recurring products was due to the ‘sponsored’ listings on the search. These sponsored results are paid for by the product manufacturers or reseller’s company to appear more frequently in the searches and at the top listing on every other page.

This results section will just focus on the 8,241 unique products that the system collected.

## Compliance:

|  |
| --- |
|  |
| *Fig. 7. Vitamins & supplements compliancy graph* |

As seen above in Fig. 7. the system found that almost all products on in our search area were non-compliant with TGA regulations. This is an impossible conclusion to come to as even a layperson could see that a lot of products on the site are ARTG approved products that you can see in your local grocery shop, so something was not right here.

## Compliance Issues:

The results show that there is a direct overlap with the ARTG compliancy outcome meaning that the ARTG collection or ARTG compliance check skewed the data. Manually looking at the data you can see that almost all of the 8,224 ‘non-compliant’ products contained no ARTG number at all meaning that the OCR could not accurately identify an ARTG on the product images or no such number existed in the first place. We expected less of an overlap with one single compliance check and to identify far more complaint products than the poultry seventeen we did find.

|  |
| --- |
|  |
| *Fig. 8. Cenovis product images that are both ARTG compliant products. [7]* |

The image in Fig. 8. illustrates the difference in product’s advertising images used. A couple observations can be made from Fig. One being that both products are made by the same company, Cenovis and secondly, that the one on the left is far simpler looking than the one on the right. To a keen observer, the image on the right has Australian compliant packaging whereas the one on the left is more generic and what we have deemed an ‘international’ advertising image as the same generic image can be used across the globe in different regions to sell a regionally compliant product. These ‘international’ kind of advertising images have certainly skewed the data collection of our system and contributed to the overall non-compliance of our results.

Another contributing factor to the system’s failure to collect ARTG numbers is in the form of ‘bulk packaging’ like those seen below in Fig. 9. The bulk package contains no visible ARTG number as the products inside the box contain individual ARTG numbers. The system that we developed can only identify what information is collected from the product page and has no capacity to reason that the bulk product comprises of two compliant products and therefore should be compliant itself.

|  |  |
| --- | --- |
|  |  |
|  |
| *Fig. 9. ‘Bulk pack’ (Left) [10] comprising of two of the Voost products (top right) [11] which are ARTG compliant products (bottom right) [11]* | |

# Key Performance Indicators (KPI):

Our project had two key KPIs to hit for it to be deemed a success.

## Collection Rate:

The outlined KPI from our project plan states, “The system must accurately determine what products are therapeutic goods from search results with a greater than 95% success rate.”

The idea behind this KPI was that it is crucial to ensure that the system can correctly identify relevant products that fall under the purview of TGA regulations and to ensure that the system is not wasting resources analysing irrelevant products. Due to time constraints, we did not collect information relating to this KPI. The system collects all the products from a search regardless of whether it is a therapeutic product or not, so we certainly are wasting resources collecting irrelevant information.

System Accuracy:  
 The second KPI states, “The system must correctly identify non-compliant products >90% with less than 5% false positive rate.”

This KPI focuses on the recall and precision of the system in detecting non-compliant products. The 5% false positive indicator was chosen to ensure the system was not erroneously flagging compliant products as non-compliant. The system we developed has a significantly higher than 5% false positive rate and thus fails to meet this KPI.

## Overall:

Comparing our project to the KPIs we set out in the project plan, it can only be said that the project was a failure on this front. Although the system that was developed functionally works, a lot more works needs to be done to it to bring it up the standard set in the KPIs.

# Time Allocation

As part of our project, we tracked an estimated number of hours spent by each team member on various aspects of the project and its surrounding assignments. The following table outlines these estimates and provides insight into the time allocation and workload distribution for the project.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Task** | **Parbat** | **Swikriti** | **Matthew** | **Dylan** |
| **Project Management** | 10 | 8 | 30 | 5 |
| **Project Plan Document** | 35 | 36 | 45 | 35 |
| **Meetings (Internal, Sponsor and Mentor)** | 11 | 11 | 12 | 12 |
| **Programming (Web Scraper)** | 0 | 0 | 280 | 0 |
| **Programming (Compliance Checking)** | 60 | 50 | 15 | 50 |
| **Programming (other)** | 0 | 0 | 20 | 0 |
| **Testing and Debugging** | 20 | 10 | 20 | 5 |
| **Final Report Writing** | 50 | 5 | 35 | 15 |
| **Poster Creation** | 0 | 0 | 5 | 20 |
| **Presentation Planning** | 10 | 20 | 15 | 10 |

In terms of what could have been done differently, one area for improvement could be in the planning and execution of the web scraping setup. This process took almost double the amount of time we anticipated. Additional resources or expertise may have been helpful in streamlining this aspect of the project.

The combination of a large programming task in conjunction with managing the team was a mistake and should be avoided in the future. Shifting the project management role from Matthew to someone else could have had resulted in more time dedicated to getting the web scraper completed in a timely manner and then he could have tackled the ARTG compliance issue plaguing the results.

Finally, it may have been useful to allocate even more time for contingency planning, as unexpected issues or delays can arise in any project. Even with the reining in of the scope of the project from two E-commerce sites to just one was not nearly enough to properly complete this project.

# Discussion:

## System Limitations:

As with any project, there are limitations to the system which can be boiled down into two categories, accuracy and functional.

### Accuracy:

The system is dependent on the accuracy and completeness of data that is available on the E-commerce websites, which can vary from product to product and as such, it will flag some compliant products as non-compliant due to the lack of ARTG number associated with a product. It also currently produces a lot of false negatives due to the lack of proper OCR data collection or interpretation. The systems implementation of OCR also leaves a lot to be desired as it will not be able to pick up an ARTG number like the one shown in Fig.10. When analysing therapeutic claims, it can be difficult to accurately determine the intent behind a claim due to ambiguity of the sentence itself of the tone of the sentence itself, as such this will almost always remain a limitation of the system. Finally, the system does not update individual product information once a web scrape is complete, it requires redoing the entire web scrape to get any updated information.

|  |  |
| --- | --- |
|  |  |
| *Fig. 10. Centrum product, example of sideways ARTG number. [6]* | *Fig. 11. Example of non-English product from ‘Vitamins & Supplements’ Amazon search. [8]* |

Functional:  
 The system is unable to process non-English text such as that seen above in Fig. 11. It is unable to accurately parse ingredients that use acronyms in place of ingredients, such as ‘PCP’ instead of the formal ingredient name ‘phencyclidine’, the system functionally parses that ingredient as compliant because that acronym does not appear on the ARTG poisons schedule. OCR pre-processing is non-existent in the current form of the system. Finally, the web scraping component is the slow and can be tedious to wait for results to be collected.

## Project Takeaways:

### Insights Gained:

Throughout the course of this project, we have gained valuable knowledge and experience in the areas of web scraping, TGA regulations, and product compliance. We are grateful for the opportunity to work on a project that has allowed us to expand our skills and understanding of this industry.

An important ethical consideration we had in data collection was that we wanted the system involved in collecting data from E-commerce websites to not violate any terms of service for automated process using the Amazon website. The issue was around privacy and consent for automated system interactions from third party actors (us) on a website. We did cross the line of ethical bot use when our system got IP blocked by Amazon for maintaining a connection to their server for too long. This issue was later rectified. We learned the importance of developing clear ethical guidelines for data collection online.

We developed a moderate understanding of regulatory frameworks as the project required a reasonable understanding of TGA regulations to accurately analyse compliance. We learned the importance of investing time in understanding the regulatory frameworks that govern the industry we are working in to ensure that our analysis is accurate and effective.

Lessons Learnt:  
 Lessons learned from this project include the importance of thorough data validation and testing, developing a lot of different use cases to test for rather just a few obvious ones. An extensive testing document should have been developed for each component of the claim checking system to ensure all cases are covered.

One project management/programming issue arose that could have been avoided had the project manager clarified what inputs, outputs and possibly even what the process should look like when designating programming tasks to multiple programmers. This issue boils down to a lack of communication from the project manager.

Outstanding Issues:   
 We believe outstanding issues identified this section are things that could be addressed in the next project phase if we were to continue working on this:

ARTG Check Issues:  
 Currently the ARTG collection is lacklustre and in need of improvement, but the checker is currently approving ARTG numbers that are less than what's required to be an official, approved ARTG number. E.g., the ARTG number of ‘17’ would be approved as ‘17’ is included as a part of ‘171111’ and ‘171112’ etc. This issue means that we can have products that might stumble their way into being approved based on what their branding says located near numbers.

Incomplete Product Infromation:  
 Upon further inspection for the raw data, some products are still missing core pieces of data like ASIN and Manufacturer, so a deep dive into where this information is located on special, edge-case scenario products would rectify this issue.

Manufacturer Name Recognition:  
 The system guess system does a check against sponsors name in the ARTG list btu there are a lot of manufacturers that have ‘Pty Ltd’ at the end of their name in the ARTG list but do not have this suffix when using their name in product advertisements and product data, so a simple removal of these suffixes could help improve the guess system.

Risks Mitigated:   
 The project went through with only a few issues. The issues that arose were already outlined in the risk management section of the project plan document and thankfully no issues arose that weren't already included there. The issues that arose were the following.

Poor Communication:  
 There were times where we had poor communication between team members especially on tasks outcomes and processes used but due to our risk mitigation strategy of increasing the frequency of progress updates, we were able to nip misunderstandings in the bud before more anymore instances could come to light.

Missed Deadlines/Deliverables:  
 Due to the web scraper coming taking until midway through week 12 to full collect results we had to fall back on the mitigation strategy of pivoting the scope of the project from two E-commerce websites to just one. Although it was a difficult decision, it was not something that we could have completed in the time frame without burning out the team.

Handover Materials:  
 The materials related to the handover of the project to sponsors are all available in the link that the team sent them. It includes some how-to guides on installing the prerequisite packages so the python file can run, as well as some screenshots on what is to be expected when the file runs.

# Future Improvements:

These are aspects of the project we would like to improve on if we got to continue with the project after this semester or if this project was to be passed on to another team to further develop.

Web Scraper Optimisation:  
 Having seen the performance of the web scraper in the project, we know there are some easy optimisations that could be made to improve on the performance of this component. One being, reducing the time that the system waits on certain pages could be reduced significantly. Alternatively, there are also methods to automatically detect when a browser has completely loaded a webpage, this would significantly improve time as it currently waits a set amount of time before starting it collection processes on a product page.

OCR Pre-Processing:  
 The system doesn’t implement any pre-processing at all on images before running them through OCR to extract therapeutic claims or ARTG numbers. A simple pre-processing function to rotate product images could be implemented which would pick up ARTG like the one in Fig. 10. where the ARTG number would be missed in the current setup. Additional pre-processing to add more contrast between the text and background colour in the image would also greatly improve the results OCR can deliver.

Alternate ARTG Number Allocation:  
 Once the system has completed its data collection, a system could be implemented to look for products on the ARTG list and try and see if there are any products that closely match the products that were web scraped but don’t have an associated ARTG number. Like a reverse search for an ARTG number, rather than checking an ARTG number is valid, it checks to see if there is aa close neighbour that could be the product E.g. An orange flavoured Berocca product didn't get an ARTG number when scraped, the system could search for ‘Berocca’ and ‘orange’ to check if there are any product names that contain those keywords and assign an ARTG number to the product that way. This could help alleviate the ‘bulk packaging’ issue covered in the compliance issue of the results section.

### Displays:

Developing a user interface (UI) and a results dashboard will improve the user experience for our system. The UI could be used to segment the system and run individual modules without running through all the previous ones, like if you wanted the system to just load your data collection and run the compliance checks on the products, completely bypassing the web scraper that would previously run. The results dashboard could help users explore the data collected and display graphs relating to the compliancy results, this would offer significantly more insight than a plain .csv file output would.

# Recommendations:

Having worked on this project for the semester, we have a few recommendations to share based on our experiences.

## Sponsor:

This project is a quite a large endeavour and should be recommended that only strong programming teams in the future attempt it as building the web scraper for this project was a whole person’s job and even then, the web scraping results weren't able to be collected until the midway through week 12, leaving very little time for other aspects of the project to be modified to compensate for any shortcomings.

Secondly, if teams do undertake it in the future, let them know that they should aim to do a ‘dry run’ of just navigating between the numerous pages and multiple searches early on in developing the web scraper as we spent a lot of time collecting the data to then find out we hit redirects and IP blocks from Amazon when navigating multiple product pages and results pages.

The OCR problems we faced could also be a separate data science/machine learning project too. Developing a solution to extracting the text from an image when not all images are the similar. We experienced a variety of different aspect ratios, resolutions, colour palettes and text orientations during the project. Having a system in place to determine what the best pre-processing strategy could be for an image based on its pixel data, image resolution or colour information could be a fun challenge.

## External:

These are for the greater public, the TGA and E-commerce websites operating in Australia.

Education and Awareness Campaigns:   
 Education and awareness campaigns can be conducted to educate the public on TGA regulations and the importance of making sure the therapeutic products they choose to use are approved by the TGA and contain an ARTG number on their packaging.

### Regular Monitoring and Auditing:

Regular monitoring and auditing of E-commerce platforms can be carried out to ensure compliance with TGA regulations. This can involve a combination of automated web scraping tools, manual inspections and data analysis to identify non-compliant products. Fines can be issued to E-commerce sites that fail to act on supplying or facilitating non-compliant goods being sold on their platform.

Collaboration between TGA and E-commerce platforms:   
 We recommend the TGA collaborate with large E-commerce platforms to enhance the overall compliance with TGA regulations. This collaboration can include the development of a shared database of non-compliant products to aid in their removal from the platform, the establishment of joint inspection programs to ensure compliance, and the creation of guidelines for compliant product listings that align with TGA regulations.

# Conclusion

The purpose of the project was to effectively identify non-compliant therapeutic goods on E-commerce websites to which we did but to the detriment of all the compliant products.

We did however prove that the use of web scraping and data analysis is a promising approach to detection although the project's results leave a lot to be desired but the underlying system that was developed acts as a good base to build up from.

We feel that for the best results will come from a collaboration between regulatory bodies and the E-commerce sites themselves to enact restrictions on what can and cannot be sold on an Australia E-commerce marketplace to ensure the health and safety of the public.

# References:

[1] Kwok, S.Y.C., Lim, G.Y., Tam, W.W.S. et al. The Association between Social Media Use and Eating Disorder Symptoms among Adolescents: A Systematic Review and Meta-analysis. BMC Public Health 21, 1667 (2021). <https://doi.org/10.1186/s12889-021-11282-x>

[2] Lau, N., Sales, K., & Harris, N. (2019). A cross-sectional study of the quality of selected Australian medicinal cannabis products available via the internet. Medical Journal of Australia, 211(5), 218-225. <https://doi.org/10.5694/mja2.50222>.

[3] Australian Government Department of Health Therapeutic Goods Administration. (n.d.). Advertising, enforcement and outcomes. Retrieved April 26, 2023, from <https://www.tga.gov.au/advertising-enforcement-and-outcomes>.

[4] Radiant Digital. (2022). What is Iterative Design? Retrieved April 20, 2023, Available at: <https://radiant.digital/what-is-iterative-design/>

[5] Berocca. (n.d.). Berocca - With Essential Vitamins for Physical & Mental Sharpness - Effervescent Tablets. [Online image]. Amazon. Available at: <https://www.amazon.com.au/Berocca-Essential-Vitamins-Sharpness-Effervescent/dp/B0B5WG2KMG/ref=sr_1_1_sspa>

[6]. Centrum. (n.d.). Men 50+ Tab 60. Amazon. Retrieved April 30, 2023, Available at: <https://www.amazon.com.au/Centrum-Men-50-Tab-60/dp/B07CCH1QF7/ref=sr_1_4_sspa?crid=31BI6099KYTQ&keywords=vitamins+%26+supplements&qid=1682992796&sprefix=vitamins+%2Caps%2C527&sr=8-4-spons&psc=1&spLa=ZW5jcnlwdGVkUXVhbGlmaWVyPUExVUNITk42VEdTMUw3JmVuY3J5cHRlZElkPUEwNTU2NjA1MzlHRFdTTzNRMzRSMSZlbmNyeXB0ZWRBZElkPUFaTVRJVkdCQTNDRDAmd2lkZ2V0TmFtZT1zcF9hdGYmYWN0aW9uPWNsaWNrUmVkaXJlY3QmZG9Ob3RMb2dDbGljaz10cnVl>

[7]. Amazon.com.au. (n.d.). Vitamins & Supplements. Retrieved April 30, 2023, [online] Available at: <https://www.amazon.com.au/s?k=vitamins+%26+supplements&crid=31BI6099KYTQ&sprefix=vitamins+%2Caps%2C527&ref=nb_sb_noss_2>

[8]. Vitamin Supplement Capsules, 2023. Amazon.com.au. Retrieved April 30, 2023, [online] Available at: <https://www.amazon.com.au/Vitamin-Supplement-Capsules-Healthy-Product/dp/B0B4RP2B4F/ref=sr_1_216?crid=31BI6099KYTQ&keywords=vitamins+%26+supplements&qid=1682993267&sprefix=vitamins+%2Caps%2C527&sr=8-216>

[9]. Australian Government Department of Health. (2022). Therapeutic goods administration. Retrieved April 30, 2023, [online] Available at: <https://www.tga.gov.au/about-tga/what-we-do/what-are-therapeutic-goods>.

[10].VOOST.(n.d.).Multivitamin Orange Effervescent Tablets. Amazon.com.au. Retrieved April 30, 2023, [online] Available at: <https://www.amazon.com.au/V%C3%96OST-Multivitamin-Orange-Effervescent-Tablets/dp/B09S96LRBT/ref=sr_1_5?crid=2XZ4NHROUOUGP&keywords=voost&qid=1682576129&sprefix=voos%2Caps%2C513&sr=8-5>

[11]. Voost,"Voost Multivitamin Effervescent 20 Pack," Amazon.com.au. Retrieved April 30, 2023, [online] Available at: <https://www.amazon.com.au/Voost-Multivitamin-Effervescent-20-Pack/dp/B07G531J8X/ref=d_pd_sbs_sccl_3_1/356-0008053-7619460?pd_rd_w=L04Wq&content-id=amzn1.sym.48833f7a-0d5e-4044-852e-e6a9d7f4d9f6&pf_rd_p=48833f7a-0d5e-4044-852e-e6a9d7f4d9f6&pf_rd_r=M69S84BE11W7Y49NS6G5&pd_rd_wg=UxXjc&pd_rd_r=3b365d97-2fcb-4f37-958c-91605f4db163&pd_rd_i=B07G531J8X&psc=1>.