Drug Safety (2021) 44:939–948
https://doi.org/10.1007/s40264-021-01084-w

ORIGINAL RESEARCH ARTICLE

Augmenting Product Defect Surveillance Through Web Crawling and Machine Learning in Singapore

Pei San Ang[1] · Desmond Chun Hwee Teo[1] · Sreemanee Raaj Dorajoo[1] · Mukundaram Prem Kumar[1] · Yi Hao Chan[1] · Chih Tzer Choong[1] · Doris Sock Tin Phuah[1] · Dorothy Hooi Myn Tan[1] · Filina Meixuan Tan[1] · Huilin Huang[1] · Maggie Siok Hwee Tan[1] · Michelle Sau Yuen Ng[1] · Jalene Wang Woon Poh[1]

Accepted: 26 May 2021 / Published online: 19 June 2021

Abstract

Introduction Substandard medicines are medicines that fail to meet their quality standards and/or specifications. Substandard medicines can lead to serious safety issues affecting public health. With the increasing number of pharmaceuticals and the complexity of the pharmaceutical manufacturing supply chain, monitoring for substandard medicines via manual environmental scanning can be laborious and time consuming.

Methods A web crawler was developed to automatically detect and extract alerts on substandard medicines published on the Internet by regulatory agencies. The crawled data were labelled as related to substandard medicines or not. An expert-derived keyword-based classification algorithm was compared against machine learning algorithms to identify substandard medicine alerts on two validation datasets ($n = 4920$ and $n = 2458$) from a later time period than training data. Models were comparatively assessed for recall, precision and their $F1$ scores (harmonic mean of precision and recall).

Results The web crawler routinely extracted alerts from the 46 web pages belonging to nine regulatory agencies. From October 2019 to May 2020, 12,156 unique alerts were crawled of which 7378 (60.7%) alerts were set aside for validation and contained 1160 substandard medicine alerts (15.7%). An ensemble approach of combining machine learning and keywords achieved the best recall (94% and 97%), precision (85% and 80%) and $F1$ scores (89% and 88%) on temporal validation.

Conclusions Combining robust web crawler programmes with rigorously tested filtering algorithms based on machine learning and keyword models can automate and expand horizon scanning capabilities for issues relating to substandard medicines.

1 Introduction

Post-marketing drug safety surveillance has been predominantly concerned with the inherent adverse events of the active pharmaceutical ingredients. Despite drug manufacturing having the most stringent industrial manufacturing and production standards, critical quality issues relating to the final product can still occur. Yet, there remains a lack of published studies investigating methods to enhance the detection, assessment and prevention strategies relating to product quality defects (i.e. substandard medicines).

Key Points

With increasing globalisation of the health product supply chain, international collaboration to ensure regulatory compliance and public health safety is critical. Environmental scanning of the Internet for relevant issues on a regular basis aids in policing potential defective products affecting the local market.

Process automation with a web crawler tool for environmental scanning enables prompt and timely capturing of relevant alerts. Using a combined machine learning and keyword-based approach to identify substandard medicine-related alerts help to reduce manual labour and the time required to filter such alerts.

Automated processes to identify relevant product defects provide precise oversight over such issues as part of the active surveillance of product quality defects.

✉ Pei San Ang
ang_pei_san@hsa.gov.sg

[1] Vigilance and Compliance Branch, Health Products Regulation Group, Health Sciences Authority, 11 Biopolis Way, #11-01 Helios, Singapore 138667, Singapore

The World Health Organization defines substandard medicines as authorised medical products that fail to meet either their quality standards and/or their specifications and can induce harm [1, 2]. The presence of contaminants and prohibited substances, errors in packaging and labelling, deviations to approved storage conditions and distribution non-compliances are some of the problems leading to substandard medicines and subsequently cause serious adverse events. In 2018, several commonly prescribed medicines were found to be contaminated with excessive levels of *N*-nitrosamines, a class of probable human carcinogens. This resulted in large-scale recalls and subsequent shortages of medicines including angiotensin-II receptor blockers, metformin and ranitidine [3–6]. This contamination issue has become the focus of international regulatory agencies and concerted efforts are being made to improve the quality and safety of medicines. More recently, amidst the COVID-19 pandemic, at the time of writing, a race has begun to find and manufacture large quantities of vaccines to address the global demand and the risk of quality issues arising with accelerated vaccine manufacturing is a real possibility. Table 1 shows examples of recalls of defective products initiated by various drug regulators.

Monitoring for substandard medicines by regulatory agencies is therefore a key component in ensuring that medicines continue to remain safe for use post-marketing authorisation. There are about 5400 pharmaceutical products, 182,000 cosmetics and 11,000 Chinese Proprietary Medicines registered with, notified to and listed with the Health Sciences Authority, Singapore (HSA), respectively. In Singapore, a health product is considered to have a defect if it has been adulterated or tampered with; it is or is possibly a counterfeit or unwholesome; it is or is possibly of inadequate quality or unsafe or inefficacious for its intended purpose; or it fails or could possibly fail to satisfy such other standards or requirements as may be prescribed [7]. Although companies are legally required to report defective health products to regulatory agencies, inevitable delays in reporting may hinder timely regulatory actions needed to minimise harm and safeguard public health. As an additional measure, the HSA proactively conducts environmental scanning activities, which involve reviewing web pages of other regulatory agencies for alerts relating to substandard medicines. Given the multitude of marketed health products marketed worldwide, the numerous regulatory agencies and web pages within each agency's website, manual scanning is undoubtedly a laborious and time-consuming task.

In this study, we explore the possibility of automatic detection and relevant assessment of information published on the web pages of international regulatory agencies via a web crawler programme and a text classification algorithm, to strengthen the surveillance efforts relating to substandard medicines. This study was carried out in two phases: the first phase involved the development of a web crawler while the second phase involved identifying the relevant substandard medicine alerts, based on the information acquired by the crawler.

2 Methods

2.1 Overview of Environmental Scanning and Prioritisation of Information

The environmental scanning process involves several steps, as follows:

1. Extract and collate all relevant information from the predefined web pages.
2. Classify the information as "substandard" and "non-substandard" related using a combined machine learning and keyword-based prediction model.
3. Manual review of the "substandard" alerts by HSA officers, which includes the following actions:
    a. Triage to prioritise based on local impact, i.e. availability of the product and severity of the defect.
    b. Determine nature of issue and severity of defect.
    c. Gather further information from companies.

In the following sections, Steps 1 and 2 of the environmental scanning process are further elaborated, including details of how the extracted data are pre-processed, and how the performance of the prediction model was evaluated.

2.2 Web Crawler Development

There are in total 46 web pages across nine different drug regulatory authorities that the HSA actively monitors for alerts relating to product defect issues. These agencies are: (1) Australian Therapeutic Goods Administration; (2) European Commission; (3) European Medicines Agency; (4) US Food and Drug Administration; (5) Health Canada; (6) Hong Kong Department of Health; (7) Malaysia Ministry of Health; (8) Malaysia National Pharmaceutical Regulatory Agency; and (9) Medicines and Healthcare products Regulatory Agency of the UK.

These web pages contain recall notices and quality or safety alerts for various health and non-health products. The type of health products that are of interest in this study are pharmaceutical products, Chinese Proprietary Medicines, cosmetics, health supplements and traditional medicines. These products of interest were categorised under

Table 1 Examples of product defect issues

Drug(s)	Issue
Agents acting on the renin-angiotensin system - ACE inhibitors Antidiabetic agents - Metformin Drugs for acid-related disorders - Ranitidine	Contamination with nitrosamine impurities
Pituitary and hypothalamic hormones and analogues - Desmopressin	Out of specification results with higher concentrations of the active substance
Ophthalmologicals - Ciclosporin	Out of specification results with lower amount of the active ingredient
Immunosuppresants - Antithymocyte immunoglobulin (rabbit)	Adverse trends in the molecular size distribution test during stability studies
Antineoplastic agents - Trastuzumab	Affected batches of solvent vials might contain glass particulates
Drugs for functional gastrointestinal disorders - Trimebutine	Possibility of foreign material in bottle
Blood substitutes and perfusion solutions - Albumin	Contamination with trace amounts of ethylene glycol
Blood coagulation factor - Eptacog alfa	Compromise of sterility due to cracks in vials
Antivirals for systemic use - Ribavirin	Hairline cracks occurring during the vial filling process
Angiotensin II receptor blockers - Lovastatin	Product mix-up with amlodipine
Cardiac therapy - Epinephrine	Injection device failed to activate correctly
Ophthalmologicals - Timolol	Defective dropper bottle pump, leading to unpredictable delivery of drop volume

ACE angiotensin-converting enzyme

the class of substandard medicine-related issues ("substandard"). In this study, substandard issues occurring in medical devices, food, blood components, veterinary medicines, pesticides, tobacco products and household items were not considered, and hence were classified under non-substandard medicine-related issues ("non-substandard"). Web pages containing media releases and announcements, and other non-product defect information were classified as non-substandard. Safety issues such as adverse drug reactions boxed warnings and the addition of contraindications were also classified as "non-substandard".

2.3 Web Crawler Design and Structure

The web crawler tool developed uses Google Chrome Driver and Python packages such as 'Selenium' and 'Requests' to extract information from the 46 web pages. The information on every web page is usually presented in two layers. The first layer contains a list of alerts in the form of a table or list or in a downloadable flat file, along with a few other details such as date posted, title and agency name. A link will then redirect to the second layer containing detailed descriptions about the alert. At this layer, the web page structure can vary significantly across the 46 web pages and data can be presented by the website in a static or dynamic manner. Most information can be found in the form of text, but some could be embedded within images or within portable document format (PDF) files. The PDF documents and images, however, were not included in the training and evaluation of the machine learning algorithm. The PDF may contain other non-defect-related information not directly relevant to the substandard case description while some PDF documents were scanned images that require image reader tools to recognise the text. These data can be extracted into PDF documents via libraries such as 'Pytesseract' and 'Tika' for documentation purposes. Some agencies might also upload information in other languages, but these were translated via libraries such as 'Googletrans' and 'Translate'.

The design of the crawler tool considered the variabilities among the different web pages. In terms of code structure, the crawler contains base code files ("utils.py") that comprise common functions executed across all web pages (Electronic Supplementary Material [ESM]). Each specific web page also has its individual code file to address specific

structural and functional characteristics of the web page. These code files are executed on a regular daily basis at specified intervals with the use of the Windows Task Scheduler. During the second and subsequent crawls of the day, the crawler is designed to identify the same alerts that were crawled previously. These duplicate alerts will be removed and hence only append the new data onto the consolidated data file. The crawler algorithm employs the standard extract-transform-load process to load the data onto a shared repository accessible to HSA officers. This extracted information is stored in the form of a comma separated values file. Figure 1 provides an overview of how the relevant data are extracted and handled throughout the automated process.

The crawler was developed iteratively over several weeks. Preliminary assessments were carried out to evaluate whether the crawler acquired all the relevant information from a given web page. Where there was missed information, amendments were made to the code to address these errors. Subsequent verification checks were then performed by the officers to assess if the code amendments were sufficient.

2.4 Model to Identify Relevant Alerts Associated with Health Products

The main purpose of the modelling algorithm was to correctly detect if the content of a web page was related to a substandard medicine. To develop the model, data between October 2019 and May 2020 (a total of 8 months of data) were extracted from the web crawler and annotated by pharmacists in the HSA with at least 5 years of clinical experience. These officers reviewed the alerts collated by the web crawler and manually classified the alerts as relating to "substandard" or "non-substandard" medicines. The data were divided into three segments: (i) data from October to December 2019 were divided in an 80-20 split for training and testing, respectively, (ii) January to March 2020 for the first set of validation data, and (iii) April to May 2020 for the second set of validation data. The 2020 records (i.e. validation set 1 and 2) were not used for training of the model.

The text data captured via the web crawler were pre-processed with standard text processing methods before being used for modelling and prediction. The following data pre-processing methods were applied on the raw data: tokenisation, special character removal, uppercase to lowercase, removal of punctuations, possessive pronouns and stop words (frequent words such as "the" and "is" that do not have specific semantic significance), stemming and lemmatisation (reducing words to a common base form or root word). The processed text was then converted into numeric representations using the term frequency-inverse document frequency technique, and machine learning algorithms could then be applied on the transformed data [8].

Attempts were made to remove duplicates. These were duplicates consisting of alerts on the same medical product for the same substandard-related issue, published by the same agency, but were captured by the crawler more than once because of reasons such as minor editorial changes to the web page text.

After pre-processing, the text data were fed into a binary classification model for classification into substandard or non-substandard medicine alerts. The data were modelled using various machine learning algorithms such as logistic regression, random forest, gradient boosting and support vector machine. Hyperparameter optimisation was used to

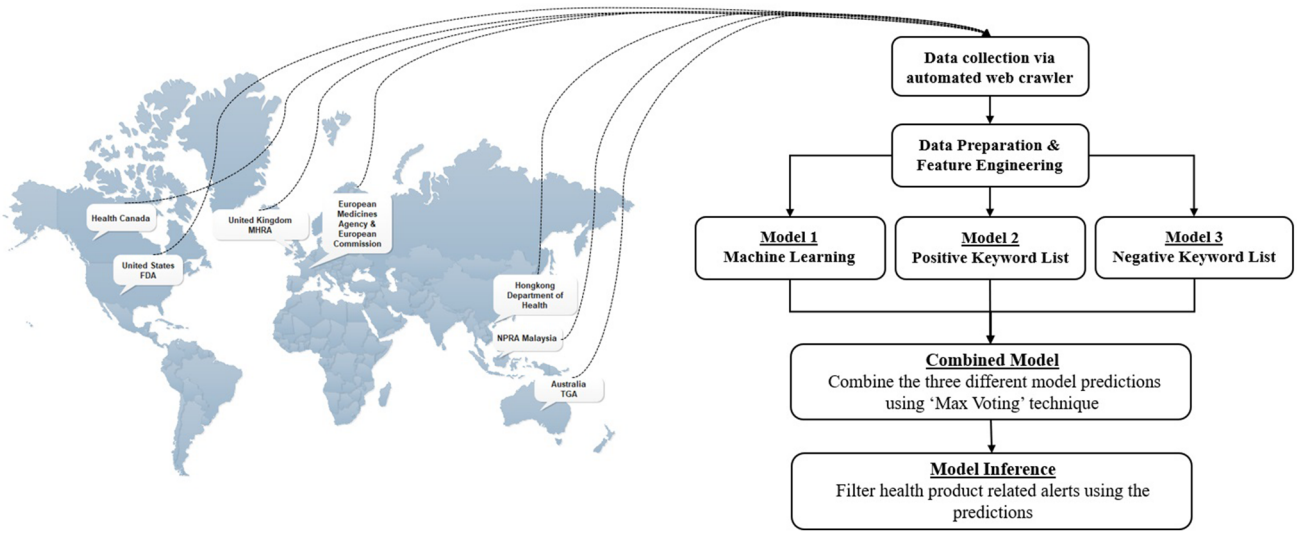

Fig. 1 Overview of the web crawler and machine learning algorithm. *FDA* Food and Drug Administration, *MHRA* Medicines and Healthcare products Regulatory Agency, *NPRA* National Pharmaceutical Regulatory Agency, *TGA* Therapeutic Goods Administration

optimise the performance of the different machine learning models. As the data sets were relatively small, on top of parameter tuning, cross-validation was also performed to further evaluate each of the machine learning models used.

Keyword-based models were concurrently explored and evaluated. A list of keywords commonly associated with health products, termed the "positive keyword list", and another list of negative terms to represent non-substandard-related issues, termed the "negative keyword list", were used. Both lists were curated by domain experts from the HSA with experience in pharmacovigilance and handling quality-related issues. Any disagreements in the choice of terms were settled by a majority vote. The positive keyword list included product types (e.g. cosmetic, dietary supplement, health supplement, herbal medicine, radiopharmaceutical), dosage forms (e.g. ampoule, auto injector, emulsion, injection), issues (e.g. affect delivery, breach of sterility, crack vial, defective cap, dosage uniformity, exceed acceptable limit, fail dissolution, failure in Good Manufacturing Practice, falsified medicine, foreign matter, glass particles, incorrect strength, label error, lack of sterility, mix-up, out of specification, seal defect, undeclared substance) and actions (e.g. drug recall). The negative keyword list included terms related to product types (e.g. agricultural, animal and veterinary, electronic, food product, medical device, toy) and issues (e.g. address medicine shortage, advertising approval, approve new therapy, draft guidance, food safety, industry forum, medical device hazard, organisational changes, public consultation, regulatory workshop). To further filter out the drug safety issues, keywords such as "boxed warning", "contraindication", "periodic safety update", "safety advisory", "safety review of" and "scientific evidence" were included in the negative keyword list. Refer to the ESM for the decision algorithm, and the positive and negative keyword lists. The predictive performance (i.e. precision, recall, *F*1) of distinguishing between substandard and non-substandard medicines, using each of the positive and negative keyword lists alone, was evaluated.

The top-performing machine learning algorithm was selected for combined modelling with the keyword-based model and the performance metrics were evaluated using validation sets (January to March 2020 and April to May 2020). In this study, we deemed recall (proportion of relevant alerts correctly identified) as the metric of highest priority because the main objective was to develop a model that could identify as many relevant substandard medicine alerts as possible. Hence, the top-performing algorithm would be selected primarily based on the recall performance. After which, the combined model comprising the selected machine learning technique together with the keyword-based model would be evaluated for its performance in both validation sets.

To better understand the limitations of the combined model, and to further improve its recall, an error analysis of the false negatives arising from both validation sets was undertaken. Code examples for crawling of static and dynamic web pages, as well as the positive and negative keyword lists are available at https://github.com/hsaproductdefect/webcrawl-substandardmeds/.

3 Results

Since October 2019, the completed automated web crawler had been crawling 46 web pages across nine different drug regulatory agencies on a daily basis with an average of 20 minutes. The crawler was programmed to run up to four times per day, at pre-specified time intervals starting from 12 midnight (GMT+8). This was to ensure that all alerts were comprehensively captured, after accounting for time zone differences across all agencies.

There were several data formats that were crawled and captured. All text-based information on web pages was extracted and neatly compiled in comma separated values format. The compiled files contained data fields such as 'Name of Agency', 'Webpage Title', 'Date of Publication', 'Description of Issue' and the Uniform Resource Locator of the web page crawled. All the information was arranged in reverse chronological order and compiled daily across all the agencies for easy review. For web pages presented in PDF or image formats, the crawler was designed to take this into account and extract these print-ready file formats into a shared repository.

Over the period of October 2019 to May 2020, a total of 12,238 alerts were crawled, translating to an average of 1530 alerts per month. The majority (80.9%) of the crawled information was from drug regulators in the USA and European countries. The number of alerts crawled from various web sources are illustrated in Figs. 2 and 3.

After removing duplicates, a total of 12,156 (99.3%) alerts were used for constructing the model. The first dataset of 4678 records (October to December 2019) was divided into two parts: 80% (3742) for training and 20% (936) for testing, respectively. A total of 7378 records were set aside for validation, 4920 records (January to March 2020) for the first validation set, and another 2458 records (April to May 2020) for the second validation set. There was a total of 2091 substandard medicine-related alerts for analysis (i.e. 745 from training set, 186 from testing set, 733 from first validation set and 427 from second validation set).

Several classical machine learning models were applied. Five-fold cross validation of the training data was performed across four different machine learning models, namely Gradient Boosting Classifier, Logistic Regression Classifier, Random Forest Classifier and Support Vector Classifier

(SVC). Overall, all the models yielded good accuracy and recall values with both the training and testing sets. The performance of SVC had both the highest accuracy and recall among all the models with 99% accuracy and 96% recall in the testing set (Table 2). Therefore, SVC was selected as the machine learning algorithm to be used for combined modelling.

Table 3 shows the performance comparison of the SVC model against a combination of SVC with a keyword-based model. Across the testing and validation sets, we applied both the SVC algorithm alone as well as the combined SVC with a keyword-based model. For the testing set, the recall was 96% for both SVC alone and SVC with the keyword-based model. For the first validation set, the recall was 86%

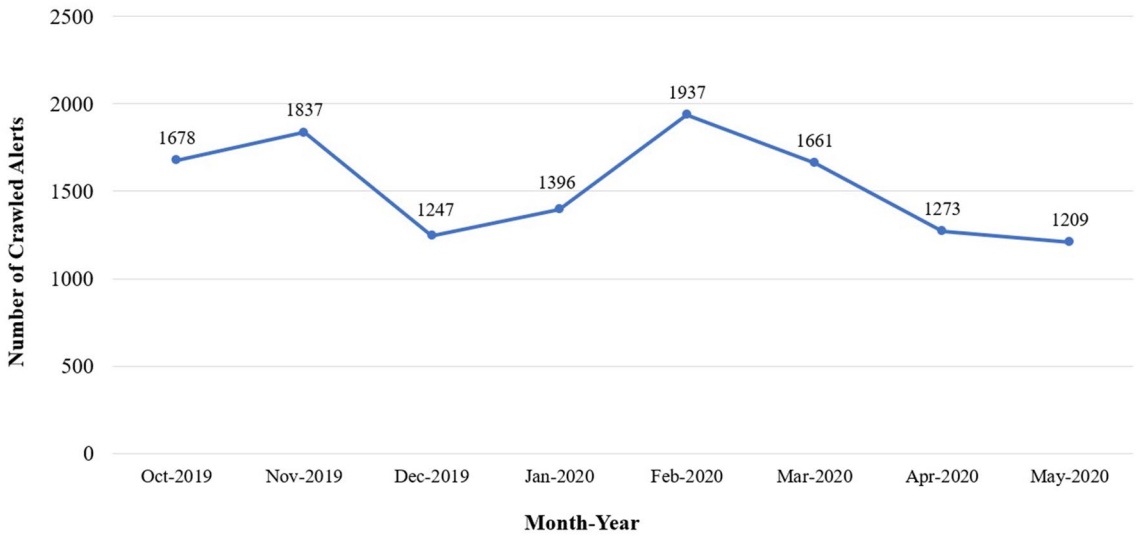

Fig. 2 Number of alerts crawled from various web sources ($n = 12{,}238$)

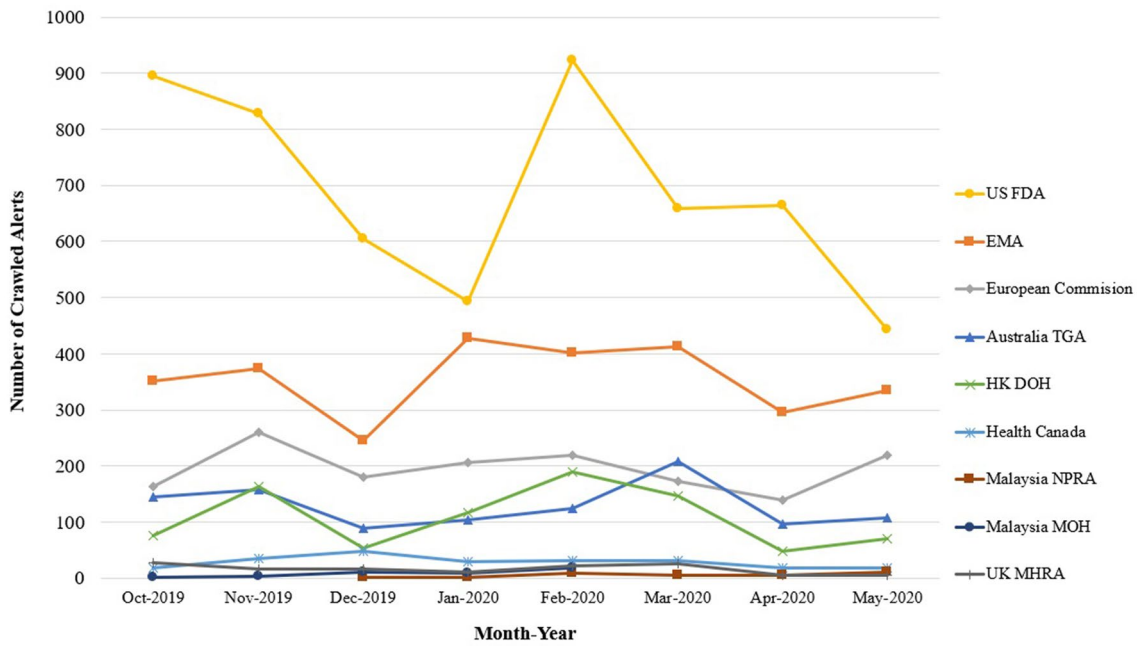

Fig. 3 Breakdown of the number of alerts crawled from various web sources ($n = 12{,}238$). *EMA* European Medicines Agency, *HK DOH* Hong Kong Department of Health, *MHRA* Medicines and Healthcare products Regulatory Agency, *MOH* Ministry of Health, *NPRA* National Pharmaceutical Regulatory Agency, *TGA* Therapeutic Goods Administration, *US FDA* US Food and Drug Administration

for SVC alone, which increased to 94% after the combined SVC with the keyword-based model was applied. Similarly, for the second validation set, the recall increased from 91 to 97% after the keyword-based models were added on top of SVC alone.

There were a total of 45 (0.9%) false negatives in validation set 1 ($n = 4920$). Of which, in 44 of these cases, the SVC predicted the alert to be "non-substandard". There was only one case in which SVC predicted correctly as a substandard medicine-related alert, though both positive and negative keyword lists voted the case as "non-substandard", hence giving an overall vote as a negative class. Among the 44 cases that SVC predicted as "non-substandard", 38 (86%) had either the positive or negative keyword list voting the case as "substandard". If these 38 cases were categorised by the SVC as "substandard" instead, the overall vote would have classified the alerts as substandard medicine-related thus reducing the number of false negatives. This demonstrated the potential value of keyword lists as a complementary mechanism to identify substandard medicine-related alerts, especially in scenarios where SVC was unable to accurately classify cases. For validation set 2 ($n = 2458$), there were 13 (5.3%) false negatives and similarly, 12 (92.3%) of these had either a positive or negative keyword list with a vote of "substandard". A continual revision and expansion of the keyword lists could likely further improve the recall of the combined model.

Table 2 Performance comparison of various binary classification models for substandard medicine-related and non-substandard medicine-related alerts based on testing data

Model	Period of analysis: October 2019 to December 2019					
	Testing data ($n = 936$)					
	Information	No. of records	Accuracy	Precision	Recall	*F*1 score
LRC	Substandard	186	0.97	0.94	0.92	0.93
	Non-substandard	750		0.98	0.99	0.98
RFC	Substandard	186	0.98	0.96	0.94	0.95
	Non-substandard	750		0.98	0.99	0.99
GBC	Substandard	186	0.98	0.97	0.95	0.96
	Non-substandard	750		0.99	0.99	0.99
SVC	Substandard	186	0.99	0.96	0.96	0.96
	Non-substandard	750		0.99	0.99	0.99

Values correct to 2 decimal places

GBC Gradient Boosting Classifier, *LRC* Logistic Regression Classifier, *RFC* Random Forest Classifier, *SVC* Support Vector Classifier

Table 3 Performance comparison of the SVC alone with SVC + K, to classify between substandard medicine-related and non-substandard medicine-related alerts

Dataset	Period	Model	Information	No. of records	Precision	Recall	*F*1 score
Testing ($n = 936$)	October 2019 to December 2019	SVC	Substandard	186	0.96	0.96	0.96
			Non-substandard	750	0.99	0.99	0.99
		SVC + K	Substandard	186	0.92	0.96	0.93
			Non-substandard	750	0.99	0.98	0.98
Validation Set 1 ($n = 4920$)	January 2020 to March 2020	SVC	Substandard	733	0.93	0.86	0.89
			Non-substandard	4187	0.98	0.99	0.98
		SVC + K	Substandard	733	0.85	0.94	0.89
			Non-substandard	4187	0.99	0.97	0.98
Validation Set 2 ($n = 2458$)	April 2020 to May 2020	SVC	Substandard	427	0.96	0.91	0.93
			Non-substandard	2031	0.98	0.99	0.99
		SVC + K	Substandard	427	0.80	0.97	0.88
			Non-substandard	2031	0.99	0.95	0.97

Values correct to 2 decimal places

SVC Support Vector Classifier, *SVC + K* combined model of SVC with a keyword-based model

4 Discussion

To the best of our knowledge, this is the first attempt to develop an automated tool to conduct environmental scanning of websites for product defect alerts relating to health products. The use of a web crawler tool for environmental scanning, coupled with machine learning and a keyword-based model provides an oversight towards active surveillance of product quality defect issues.

Substandard medicines pose global public health concerns [9–12]. The supply chain of health products can be complex and intricate [13, 14]. Companies are operating increasingly sophisticated global supply chains by outsourcing and establishing various manufacturing sites, making comprehensive product traceability and timely recall of affected marketed products difficult.

It is a legal requirement for manufacturers, importers, suppliers and product registrants to report defective health products. Guidelines were published for pharmaceutical companies on reporting, handling and investigating suspected quality defects, with an emphasis on the legal requirements in managing defects and recalls. However, drug regulators may not always be notified of the product defects in a timely manner because of delays in reporting by the pharmaceutical companies or limited regulatory oversight of products, which are not subjected to pre-approval authorisation [15, 16]. Therefore, there is a need to leverage on other regulatory drug agencies to share information on product defect issues.

Environmental scanning of the Internet on a regular basis aids the HSA in minimising potential defective products from affecting the local market. This complements the other post-marketing surveillance strategies that the agency has, such as mandatory product defect reporting by companies, product quality surveillance programmes, good manufacturing practice inspections and information sharing via international working groups. However, environmental scanning becomes labour intensive when there is a plethora of information available on the Internet. The aim of the web crawling algorithm is to utilise the shortest amount of time to find all the relevant information over the Internet. Approximately 350 alerts are extracted from the 46 web pages of interest every week. The automated crawler resulted in a time saving of about 55%. The time taken on average for trained officers to review the extracted alerts from the automated crawler was 75 minutes per week as compared with manual scanning of the web pages and converting the web pages into PDF for follow-up which took the same officers 165 min per week. Overall, the automated crawler saves a considerable amount of time spent by officers and helps to automate time-consuming tasks related to scanning web page content. This allows officers to dedicate the time saved for other tasks. The web crawler is a scalable automation solution that allows the seamless addition and removal of web pages. This aids in monitoring the less frequented web pages whenever there is new information. Compared to manual scanning whereby irrelevant information will not be recorded, the web crawling algorithm is able to store all the information in a manageable format to facilitate information retrieval when needed.

However, there are challenges in developing and validating the web crawler algorithm. First, there was no automated method to determine the extraction efficiency of the crawler, i.e. whether the crawler captured all alerts on the target web page as intended. In the initial phases of the project, officers from the HSA manually verified the crawled content against the actual web page content. Over time, when there was little to no variability and the crawler was deemed to be sufficiently reliable, manual verification was not necessitated. Other issues of the crawler development included the reliability of network connectivity and the presence of information in languages other than English. For web pages in non-English languages, the use of the 'Googletrans' package helped to convert the content into English before extraction. However, there were additional troubleshooting steps involved, such as ensuring the right character encoding and verifying the accuracy of translation through officers conversant in both languages. Each website also has its unique format and structure, which makes it challenging to apply the same crawling framework across multiple websites. This format and structure are also subjected to changes over time. Additionally, some web pages restrict the amount of data that can be crawled, and the crawler could be blocked because of the overloading of requests (limit varies for each website). We had also encountered situations whereby the posted dates of the alerts were misspelled on the website and as a result the crawler was not able to pull out this newly posted information. Occasionally errors could happen during web crawling. Reviews of crawler log files are thus required to amend the web crawler algorithm where needed.

The use of machine learning algorithms can make information retrieval easier and more efficient. SVC is a valuable tool for making classifications and generalises well to a wide range of applications. It is a supervised machine learning technique for binary classification and finds the optimal boundary that classifies information into two classes [17]. It works well on smaller cleaner datasets and is less effective on noisier datasets with overlapping classes. In linear SVC, we tried to find a line to separate the information into two classes. If the data are non-linearly separable, techniques such as the kernel method could be applied to project the data into a higher dimension feature space and hence make the data linearly separable [17].

We developed two sets of keyword-based models that could identify words or key phrases from the crawled text to describe substandard medicine alerts. The recall for the combined machine learning and keyword-based model was higher than the machine learning algorithm in both

validation sets (94% vs 86% in validation set 1 and 97% vs 91% in validation set 2). It was noted that the increase in recall had led to a considerable decrease in the precision. This hybrid model demonstrated an improvement in performance for the text classification model. Positive keywords that frequently appeared in alerts included “”pharmaceutical”, “lack of sterility”, “out of specification” and “certification”. Top negative keywords included “food”, “veterinary”, “orphan designation” and “food safety”. Generally, using the positive keyword model or negative keyword model alone produced an average recall of 80%, which was lower than the recall of the combined machine learning and the keyword-based model. Our intention was to improve the accuracy of detection and help reduce false-positive rates. Achieving a higher recall or sensitivity ensured relevant product defect alerts were not missed as the number of false negatives were minimised. There were challenges developing the keyword lists. For example, words such as apricot and strawberry could not be added to the negative keyword list as these appear in brand names, flavourings or ingredients of health supplements. The keyword list needs to be constantly reviewed to optimise the model. During the COVID-19 pandemic period, we had added COVID-19-related terms such as “COVID-19 test” and “hand sanitiser” to filter away noise for model training.

We also looked at the false-positive and false-negative cases to determine what affected the performance of the algorithm. In general, it was difficult to differentiate safety-related alerts from quality alerts using keyword lists as safety issues often contained the same positive keywords for quality alerts. Some examples included information about routine surveillance testing of the quality of vaccines to monitor their quality compliance and these alerts were not in response to any specific safety or quality concerns; general announcements on nitrosamine-related matters such as mentions of metformin products being tested and not exceeding the acceptable daily intake for N-nitrosodimethylamine; as well as COVID-19-related advisories that cautioned users about false or misleading claims to prevent, treat or cure COVID-19. Generally, these were not quality defect alerts but were picked up as false-positive cases.

There were also updates on the formation of advisory committee on vaccines and regulators’ announcements of guidelines that were not quality-related issues. There were news items about healthcare professionals being sentenced for supplying drugs on the black market. This could be tackled by adding new negative keywords such as “prison sentence”, “sentenced” and “trafficking”. There were letters to researchers stating non-compliance to research protocols and study drug administration. Similarly, relevant negative keywords could be added. There were news items about no shortages of medicines despite bushfires. We were unable to include the word “shortage” in the negative keyword list as some quality-related alerts may mention supply and shortage.

Some alerts were very brief with little details and there would be a need to rely on the product names as mentioned in the alerts. However, some product names for cosmetics and health supplements were not easily recognisable as health products. Similarly, some medical devices may not have the word ‘device’ in their product name. Theoretically, it is possible that over time, as more data are captured and the model retrained, the combined model would be better equipped to differentiate these ambiguous classifications. Alternatively, smaller models could be developed to handle these data subsets, and specifically target ambiguous classes for more definite predictions.

The main limitation of the study was the lack of detailed information on the affected products and case descriptions posted on the web pages. If there was insufficient information about the product, for example, with only a brand name provided and active ingredients not being stated, the machine learning algorithm may not be able to correctly classify the alerts into substandard medicine-related alerts. In some circumstances, machine learning algorithms may treat alerts of inherent safety issue as product quality defect issues. While a hybrid model is proposed, there may be other modelling methods (and a combination of models) that may be better at detecting relevant alerts. A thorough investigation on the best approach for addressing the prediction task at hand unfortunately is beyond the scope of this study. Nonetheless, the error analysis and the finding that the keyword-based model still retains some residual complementary value when ensembled with machine learning highlights the need for future work on optimal methods of developing models for detecting relevant alerts.

Despite not achieving 100% sensitivity, we still think there is a role for a web crawler tool for environmental scanning, aimed at quickening the rate of detecting alerts to substandard medicines. Given that companies are legally required to report defective health products, it is reasonable to assume that key issues will eventually be made known to regulators. However, to quicken the rate of detecting such information, we propose web crawling and the classification model as an additional measure. Manual checking of the output of the tool will still be required to ensure a complete capture of all cases but the frequency at which these need to be done may now be reduced.

The HSA receives product defect information from various sources daily. A defect can be classified as either critical or non-critical according to the potential impact to public health, the nature of the issue and the risks posed to the intended users of the therapeutic product. There is a need to distinguish critical defects from non-critical defects and prioritise the critical defects for review to ensure that prompt

actions are taken if required. By classifying the product defect based on the nature of the issue, it helps regulators to identify cases that are more crucial. Further work needs to be carried out to explore text classification of the nature of issues for prioritisation and impact assessment.

5 Conclusions

There is an increasing focus on post-marketing surveillance and higher public expectations of drug regulators to play a leading role in safeguarding public health. Process automation with a web crawler tool for environmental scanning, coupled with machine learning and a keyword-based model to identify relevant product defects, offers a promising approach for an effective post-marketing surveillance of health products. This aids the HSA in keeping abreast with issues that occur locally and internationally to ensure the safety, quality and efficacy of locally available health products. There is a potential for scalability to a broader geographic reach and this work can be adopted by other drug regulatory agencies.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s40264-021-01084-w.

Declarations

Funding This initiative received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

Conflicts of interest/Competing interests Pei San Ang, Desmond Chun Hwee Teo, Sreemanee Raaj Dorajoo, Mukundaram Prem Kumar, Yi Hao Chan, Chih Tzer Choong, Doris Sock Tin Phuah, Dorothy Hooi Myn Tan, Filina Meixuan Tan, Huilin Huang, Maggie Siok Hwee Tan, Michelle Sau Yuen Ng and Jalene Wang Woon Poh have no conflicts of interest that are directly relevant to the content of this article. The views expressed in this article are the authors' personal views and may not be understood or quoted as being made on behalf or reflect the position of the HSA.

Ethics approval Not applicable.

Consent to participate Not applicable.

Consent for publication Not applicable.

Availability of data and material The information used for this study is available on public domain.

Code availability All codes relevant to the study, i.e. to run the web crawler, and machine learning algorithm and keyword-based model, are included in the article as ESM.

Author contributions PSA proposed the research idea and design. DCHT, MPK and SRD designed the models and provided the computational framework to analyse the data. YHC and MPK wrote the codes for the web crawler. PSA, CTC, DSTP, DHMT, HH, MSHT and MSYN provided the domain expertise for manual annotation of the data and developed the list of positive and negative keywords. JWWP provided the thought leadership of the project. All authors read and approved the final version of the article.

References

1. World Health Organization. Substandard and falsified medical products. https://www.who.int/news-room/fact-sheets/detail/substandard-and-falsified-medical-products. Accessed 17 Oct 2020.
2. World Health Organization. A study on the public health and socioeconomic impact of substandard and falsified medical products. Geneva: World Health Organization; 2017.
3. World Health Organization. WHO Information Note: update on nitrosamine impurities. Apr 2020. https://www.who.int/docs/default-source/substandard-and-falsified/informationnote-nitrosamine-impurities-april2020.pdf?sfvrsn=4f06ed1b_2. Accessed 11 Oct 2020.
4. Health Sciences Authority. HSA updates on overseas recall of angiotensin II receptor blockers. 2019. https://www.hsa.gov.sg/announcements/news/hsa-updates-on-overseas-recall-of-angiotensin-ii-receptor-blockers-(mar-2019). Accessed 11 Oct 2020.
5. Health Sciences Authority. HSA stops supply of eight brands of ranitidine products in Singapore. 2019. https://www.hsa.gov.sg/announcements/news/hsa-stops-supply-of-eight-brands-of-ranitidine-products-in-singapore. Accessed 11 Oct 2020.
6. Health Sciences Authority. HSA recalls three out of 46 metformin medicines. 2019. https://www.hsa.gov.sg/announcements/news/hsa-recalls-three-out-of-46-metformin-medicines. Accessed 11 Oct 2020.
7. Health Products Act (Cap 122D, 2008 Rev Ed) s 42(6).
8. Wikipedia Wikimedia Foundation. tf-idf. 2020. https://en.wikipedia.org/wiki/tf-idf. Accessed 20 Oct 2020.
9. Almuzaini T, Sammons H, Choonara I. Substandard and falsified medicines in the UK: a retrospective review of drug alerts (2001–2011). BMJ Open. 2013;3(7):e002923. https://doi.org/10.1136/bmjopen-2013-002923.
10. Almuzaini T, Sammons H, Choonara I. Quality of medicines in Canada: a retrospective review of risk communication documents (2005–2013). BMJ Open. 2014;4(10):e006088. https://doi.org/10.1136/bmjopen-2014-006088.
11. Rasheed H, Höllein L, Holzgrabe U. Future information technology tools for fighting substandard and falsified medicines in low- and middle-income countries. Front Pharmacol. 2018;9:995. https://doi.org/10.3389/fphar.2018.00995.
12. Johnston A, Holt D. Substandard drugs: a potential crisis for public health. Br J Clin Pharmacol. 2014;78(2):218–43. https://doi.org/10.1111/bcp.12298.
13. Shah N. Pharmaceutical supply chains: key issues and strategies for optimisation. Comput Chem Eng. 2004;28(6–7):929–41.
14. Langer E. Biopharmaceutical industry outsourcing trends: China and India continuing full-court press. Pharm Outsourcing. 2012;13(5).
15. Cross C. Health Canada to improve drug recall process. CMAJ. 2013;185(18):E811–2. https://doi.org/10.1503/cmaj.109-4641.
16. Janetos T, Akintilo L, Xu S. Overview of high-risk Food and Drug Administration recalls for cosmetics and personal care products from 2002 to 2016. J Cosmet Dermatol. 2019;18(5):1361–5. https://doi.org/10.1111/jocd.12824.
17. Ben-Hur A, Weston J. A user's guide to support vector machines. Methods Mol Biol. 2010;609:223–39. https://doi.org/10.1007/978-1-60327-241-4_13.