IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received June 23, 2019, accepted July 6, 2019, date of publication July 10, 2019, date of current version July 29, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2928122

Forecasting Hospital Emergency Department Patient Volume Using Internet Search Data

ANDREW FU WAH HO[1], **BRYAN ZHAN YUAN SE TO**[2,3], **JIN MING KOH**[4], **AND KANG HAO CHEONG**[4], **(Member, IEEE)**

[1] SingHealth Duke-NUS Emergency Medicine Academic Clinical Programme, Department of Emergency Medicine, Singapore General Hospital, Singapore S169608
[2] School of Computing Science, University of Glasgow Singapore, Singapore S737729
[3] Infocomm Technology Cluster, Singapore Institute of Technology, Singapore S138683
[4] Science and Math Cluster, Singapore University of Technology and Design, Singapore S487372

Corresponding author: Kang Hao Cheong (kanghao_cheong@sutd.edu.sg)

This work was supported by the Singapore University of Technology and Design Start-up Research Grant under Grant SRG SCI 2019 142.

ABSTRACT We present an efficient and scalable system to predict emergency department (ED) patient volume in hospitals using publicly available Google Trends search data. Search volume data are retrieved for a selected set of context-relevant query keywords with refinements, on which a series of correlation analyses are performed, and a multiple regression predictive model is constructed. We also develop a software suite to enable convenient access to data visualization and prediction capabilities by medical and administrative staff. A preliminary demonstration of the method and software is presented with data from a large public hospital as a form of validation. This paper enables informed resource and manpower allocation in hospitals and thus improved ability to respond to patient influx surges, and importantly, can serve as a key mitigation measure against worsening ED congestion problems that plague hospitals.

INDEX TERMS Data analytics, data-driven, predictive model, multiple regression, Google Trends, medical, emergency department, hospitals, healthcare, health services.

I. INTRODUCTION

The reputation of Google as a leader amongst search engines and web service providers is notable. A comparison by *Comscore* in April 2017 has revealed a large disparity between the annual 9.6 billion queries managed by Google and the 1.8 billion managed by competitor Yahoo, and that Google has achieved more than 63% and 94% market penetration for desktop and mobile search respectively. These figures suggest massive amounts of search data that Google possesses, possibly exploitable in solving real-world problems. In working towards making such data accessible, Google Trends (GT) offers the capability to track popular search terms in numerous countries, and comparisons between search volumes of different queries. This tool was created primarily for marketing analytics, but is available in the public domain, and can thus be used for a diverse range of purposes.

The popularity of Google search across platforms, and its penetration into web services of diverse types, including mapping, blogging, video streaming, and social media, allows the collation of user-generated search data of great comprehensiveness and breadth. Such data can possibly be used as a proxy indicator of real-time trends in various domains and regions. Indeed, various studies have explored the applicability of Google Trends search data in different industries. A notable example is an analysis of Google Flu Trends (GFT), revealing a consistent relationship between trends data and the number of flu reports collected by the Centers for Disease Control and Prevention (CDC) in the United States [1]. The statistical association can allow the identification of increases in flu cases weeks in advance of CDC records. The study suggests that search data can carry significant value in providing insights into quantitative trends, and when applied in the medical industry, can enable early intervention in service demand surges or declines.

Medical treatment in hospitals is limited not only by the available medical technology and expertise of medical professionals, but also by manpower and the ability to cope with patient influx surges. Fluctuations in patient volumes hinder effective distribution of resources and therefore diminishes treatment quality. In particular, emergency department (ED) overcrowding has become increasingly major in

The associate editor coordinating the review of this manuscript and approving it for publication was Jing Bi.

recent times [2]–[4], and currently stands as the greatest threat to emergency medical care with associations to increased amounts of errors from medical professionals, late time-critical care, and extra deaths [5]–[14]. In the local context of the state of Singapore [15], currently experiencing an ageing population and a consequent strain on medical infrastructure [16]–[18], bed shortage risks compromise the ability to cope with patient influx. The ability to accurately forecast expected patient volume in advance will allow hospitals to make better-informed manpower and resource allocations, potentially improving operational efficiency, reducing operational costs, and alleviating overcrowding problems.

In the present study, we utilize Internet search data made available by GT to forecast ED patient volume for the Singapore General Hospital (SGH), the largest public hospital in Singapore. A statistical analysis methodology to assess correlation traits between a selected set of medically-relevant query keywords and institutional hospital visit records is developed, and multiple regression predictive models are constructed on the basis of the correlational analyses. We also develop an interactive software that enables medical and administrative staff to access the data and predictive model conveniently. The primary motivation of our work is to mitigate worsening ED congestion problems that have not seen effective resolution with existing measures.

The structure of this paper is as follows—we first provide a technical overview, discussing current efforts in mitigating ED congestion in Section II-A, and existing studies in utilizing Internet search data for medical applications in Section II-B. We then discuss the analytical methods for correlation assessment and regression, and the functionalities of the developed *Dashboard* software platform, in Section III. Lastly, we detail key results, including validation of the accuracy of patient volume predictions, in Section IV.

II. TECHNICAL OVERVIEW

A. ED CONGESTION MITIGATION

Congestion in the ED leads to delayed medical treatment, subjecting patients to prolonged pain and decreased recovery rates [13], [14], [19]–[22]. Other than the direct health ramifications, related ethical issues also arise, in particular from the suppression of quality of care available to patients, and the need to distribute scarce resources and manpower over a large patient volume [23], [24]. Various solutions have been proposed to mitigate ED overcrowding—for instance, Salway *et al.* suggests that the boarding of ED patients can be lowered through internal ED-specific and external hospital-wide systemic adjustments [25]. Congestion in the ED occurs because of its limited ability to cope with surges in patient influx, and is also intricately linked to overcrowding in other departments, which diverts patient influx into the ED and hinders outflux; the study therefore recommends resource management and staff re-allocation as part of the mitigation strategy, with the goal of improving the flexibility of EDs in handling high-demand periods [25]. The smoothing of elective cases is also presented as a measure to free up resources during high-demand periods, and patient discharge schedules are also suggested to be optimized to reduce volatility in bed availability.

Yet ED overcrowding continues to be a worsening issue, with little practical progress in management and recovery. The degree and consequences of congestion problems are likely not entirely recognized in public medical facilities; and in hospitals that have implemented proposed measures, their effectiveness was not observed to be satisfactory. An analysis on the ‘four-hour target’ of the England National Health Service (NHS) had indicated an improvement of the availability of beds through system process revamps at approximately 5–8% of total capacity, but these process adjustments cannot be the only mitigation measure [26]. A growing consensus within the medical community is a need for greater automation in medical facilities, and more effective exploitation of massive data analytics to improve medical care [27]–[30]. Such programmes may encompass medical data collection on nationwide scales to better inform the adjustment of present procedures and policies. The current study indeed concerns massive data analytics—but on publicly available Internet search data, rather than patient-level medical data.

B. WEB DATA ANALYTICS

There have been a number of prior studies investigating the plausibility of using Internet search data for medical applications [31]–[33]. For instance, Gluskin *et al.* had successfully used Google Dengue Trends (GDT) alongside various other data modalities to construct descriptive models for dengue cases in Mexican states, with excellent model accuracy against real-world records [34]. A similar study using web search data had been conducted by Althouse *et al.* on dengue occurrence in Singapore and Bangkok [35], with step-down linear regression, generalized boosted regression, and negative binomial regression explored as plausible predictive models. The linear model with Akaike information criterion step-down was found to be the most suitable.

Further studies by Carneiro and Mylonakis had demonstrated a web-based tool for real-time surveillance of disease outbreaks, utilizing GT as a basis for monitoring medical care demand worldwide [36]. Medical communities around the world may devise disease control solutions that suit their particular needs more effectively through such a tool. Yang *et al.* has also introduced a framework, Autoregression with Google search data (ARGO), that utilizes GT data to estimate influenza-like illness activity levels in the United States [37]. The successful usage of Twitter activity data, rather than Google search data, as a proxy for monitoring and predicting flu trends have also been reported [38], [39]. The current state of research strongly indicates that web data holds immense potential in powering descriptive and predictive capabilities. The potential benefit of such technologies in the medical field heavily motivates the development of new application platforms, including that of the current study.

III. DESIGN AND IMPLEMENTATION

In this section, we detail the development of our software platform and data analysis methods, beginning with a general overview of architecture and framework (Section III-A), data collection (Section III-B), correlation analysis methods (Section III-C), the multiple regression model (Section III-D), and lastly the *Dashboard* software platform (Section III-E).

A. SOFTWARE FRAMEWORK

The software package developed in this study was implemented in *R* and *Visual Basic for Applications* (*VBA*). *R* is a programming language widely used in data analytics and statistical computation—we use the *gtrendsR* package to retrieve GT data, such as the search volume for specific search terms, hit locations, and query dates. Statistical analyses on the GT data and institutional medical records are also performed on *R*.

A multiple regression predictive model based on the results of the statistical analysis is implemented, alongside a graphical front-end in *VBA* for users to interface with. The front-end is called the *Dashboard*, and allows non-technical personnel to conveniently access and visualize the processed data, and to utilize the predictive model.

TABLE 1. **List of selected query keywords related to the Singapore General Hospital (SGH). The Mass Rapid Transit (MRT) in Singapore is the primary public transport metro system serving commuters, with a station in close proximity to SGH.**

Query Keywords	
SGH A&E	SGH appointment
SGH pharmacy	SGH contact
SGH parking	SGH food
SGH map	SGH
Singapore General Hospital address	Singapore General Hospital MRT
Singapore General Hospital map	Singapore General Hospital

B. DATA COLLECTION

1) NON-REALTIME DATA

Retrieved non-realtime GT search data is restricted to originate from the locale of Singapore and to be in the English language, the primary language across the local demographic. Datasets comprising non-realtime GT data span from 2006 to 2016. The query keywords for which data is retrieved are selected to be greatly relevant to SGH (see Table 1), under the assumption that patients might search these keywords online in some prior time period before visiting the hospital. The majority of these keywords concern locations and maps of SGH, supplemented by appointment-related keywords. Location-based keywords are logically expected to be greatly relevant to our prediction results, as patients or accompanying persons often need to determine how to navigate to the hospital; appointment searches are also expected to occur as patients confirm their appointment details. The selected set of keywords is supplemented by additional refinements from Google *Autocomplete*, which suggests searches to users as they type, presumably also frequently used by the patient demographic. In this manner, a wide range of keywords characteristic of searches by potential patients are captured. The retrieved data is condensed to count weekly search hits for each keyword in the selected set. A plot of retrieved non-realtime data is presented in Figure 1.

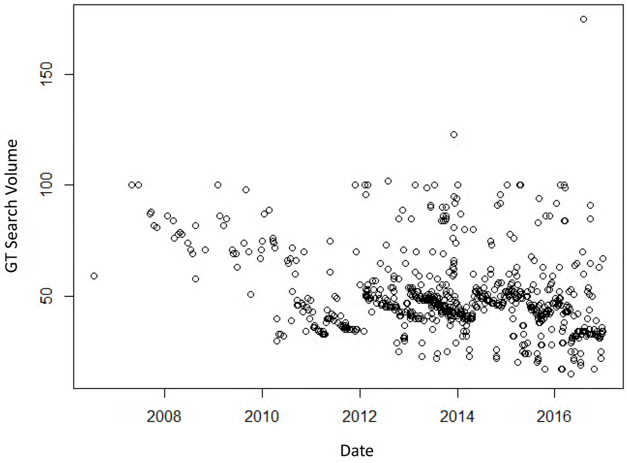


FIGURE 1. **Plot of retrieved non-realtime GT data with search volume index (SVI) > 1 over the 2006–2016 time period. The density of available data noticeably increases, especially in 2013 and beyond, attributable to the increase in Google search usage as mobile technology proliferates.**

TABLE 2. **Sample realtime GT search volume data for various query keywords, and official hospital visit records for the same dates.**

Dates	Actual ED	Query Keyword Search Volumes			
		"SGH Map"	"SGH"	"SGH Map" (Quartic)	"SGH" (Quartic)
1/07/2019	362	149	822	4.93×10^{8}	4.57×10^{11}
1/08/2019	351	210	1120	1.94×10^{9}	1.57×10^{12}
1/09/2019	344	322	1098	1.08×10^{10}	1.45×10^{12}
1/10/2019	352	385	1109	2.20×10^{10}	1.51×10^{12}
1/11/2019	335	239	1109	3.26×10^{9}	1.51×10^{12}
1/12/2019	311	126	871	2.52×10^{8}	5.76×10^{11}
1/13/2019	291	124	967	2.36×10^{8}	8.74×10^{11}
1/14/2019	381	438	1089	3.68×10^{10}	1.41×10^{12}
1/15/2019	354	377	1117	2.02×10^{10}	1.56×10^{12}

2) REALTIME DATA

Throughout the development period of the current study in January to February 2019, realtime GT search data was retrieved on a daily basis, for the same set of keywords as detailed in Table 1. Realtime data in this time range is expected to be of greater volume and quality than the historical 2006–2016 non-realtime data, due to the increase in Internet usage over the past decade. A sample of the retrieved realtime GT data is shown in Table 2.

3) OFFICIAL DATA

Official hospital visit records, which count the number of actual ED visits in SGH from January 2006 to February 2019, were extracted from the local health institutional data. These records are anonymized to preserve confidentiality, and are condensed to count daily ED visits.

C. STATISTICAL ANALYSIS

A number of correlation analysis methods is used to explore association between patient volume and various predictors derived from search data. An overview of these methods is given in the following subsections. The statistics yielded enable the identification of key correlated factors in patient volume, which informs the construction of the multiple regression predictive model.

1) PEARSON CORRELATION COEFFICIENT

The Pearson product-moment coefficient computes the robustness and direction of association between two variables, reflecting the extent of linear correlation. The coefficient is given by

$$r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2}}, \qquad (1)$$

where x_i and y_i are the values of independent and dependent variables respectively, $\bar{x}$ and $\bar{y}$ denote the means of x and y respectively, and n value pairs are considered. The coefficient r satisfies $-1 \leq r \leq 1$, with $r = 1$ and $r = -1$ indicating a perfect positive and negative correlation respectively, and $r = 0$ indicating no correlation between the variables. In the current context, x_i are the daily search volume of the selected keywords, and y_i represent the actual ED patient volume.

2) KENDALL RANK CORRELATION COEFFICIENT

The Kendall τ coefficient is a non-parametric measure that computes the degree of correlation between non-interval variables. The coefficient is computed as

$$\tau = \frac{2(n_c - n_d)}{n(n-1)}, \qquad (2)$$

where n_c is the total number of concordant pairs, n_d is the total number of discordant pairs, and n is the number of value pairs considered. Two value pairs (x_i, y_i) and (x_j, y_j) are concordant if $x_i < x_j$ and $y_i < y_j$, or $x_i > x_j$ and $y_i > y_j$; they are discordant if $x_i < x_j$ and $y_i > y_j$, or $x_i > x_j$ and $y_i < y_j$; and are neither if $x_i = x_j$ or $y_i = y_j$.

3) SPEARMAN RANK CORRELATION COEFFICIENT

The Spearman coefficient is also a non-parametric measure of association between rank variables. While the Pearson correlation coefficient assesses linear relationships between two variables, the Spearman coefficient assesses monotonic relationships. The Spearman coefficient is computed as [40]

$$\rho = \frac{\sum_{i=1}^{n}(x_i' - \bar{x'})(y_i' - \bar{y'})}{\sqrt{\sum_{i=1}^{n}(x_i' - \bar{x'})^2}\sqrt{\sum_{i=1}^{n}(y_i' - \bar{y'})^2}}, \qquad (3)$$

where x_i' is the rank of x_i, y_i' is the rank of y_i, and $\bar{x'}$ and $\bar{y'}$ are the means of x' and y' respectively.

D. MULTIPLE REGRESSION

We utilize a multiple regression model to enable predictions on expected ED patient volume, based on multiple indicators derived from search query volume metrics. In general, a linear multiple regression for k independent variables can be written [41] as

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + \epsilon, \qquad (4)$$

where y is the response (dependent) variable, x_i are the various predictor (independent) variables, β_i are weighting coefficients, and ϵ is an error or noise term. In the context of our study, the predictors are search volumes for various query keywords, and the response variable of interest is the ED patient volume.

The training of the model is performed using an iterative elimination method, where the entirety of available training search volume data is used at first as predictors, and variables are then removed if doing so improves the model, as measured by multiple R and multiple R^2 metrics. In this process, linear to quartic powers of the variables are trialed to improve the regression, enabling a flexibility beyond linear regression. This training procedure is continued until no significant improvement can be achieved, as reflected when all remaining predictors are of p-value ≤ 0.05. Intrinsically, such a regression procedure assumes that the correlation between search data and actual patient volume can be approximately captured through a linear combination of polynomials in each of the basis keyword volumes, hence, as with all series-based methods, increasing the number of terms allowed in the regression will in general aid accuracy in exchange for increased computation complexity, with diminishing returns.

Constructing a regression using time-synchronized search volume data and official records will produce a model suited for real-time estimation and prediction. However, we also wish to be able to make in-advance predictions of patient volume. This is enabled by introducing a time retardation offset to the search volume data, such that regression is performed on official records against search data that is of some duration prior. For instance, to construct a model for next-day prediction, a time offset of one day is used (denoted $T - 1$), and for next-week prediction, a time offset of seven days is used (denoted $T - 7$).

E. ANALYTICS DASHBOARD

The *Dashboard* software was developed using *VBA* to present a convenient interface for the access and visualization of processed data, and for the computational prediction of ED patient volumes. The software package is intended for use by medical staff to manage manpower allocation in an informed manner, as well as administrators and authorities overseeing the functioning conditions of hospitals. Two tabs are presented to the user—dashboard and data—whose functionalities are described in the following subsections.

1) DASHBOARD TAB

Key data visualizations and summaries are presented on the main *dashboard* tab (Figure 2). The default graphical configuration comprises three charts, plotting the predicted ED patient volume trend for a specified time range, a comparison

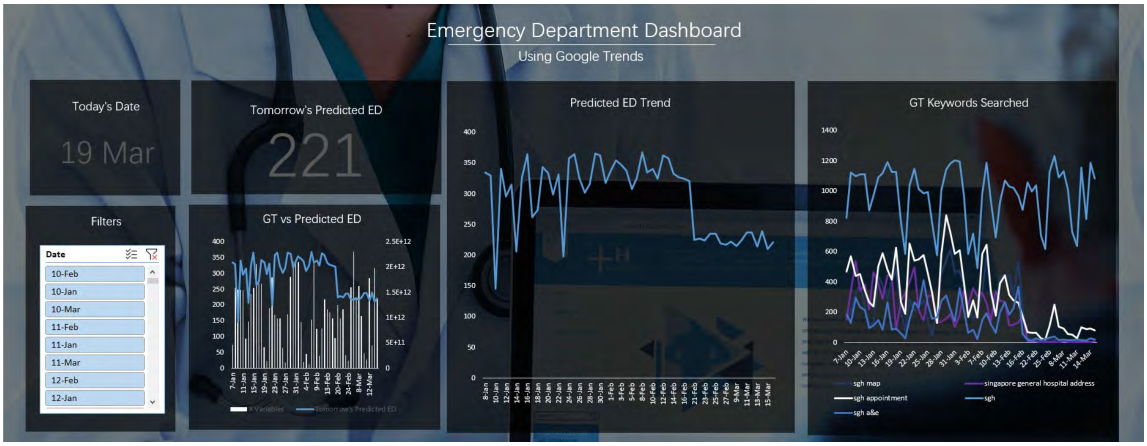
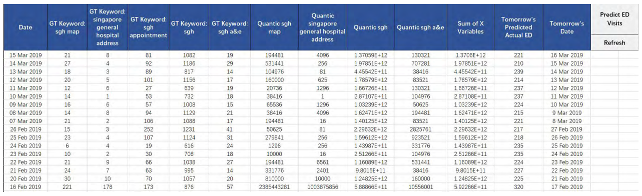


FIGURE 2. The *Dashboard* software main graphical interface. A series of charts visualize processed search volume data and the predicted ED patient volume, and a filter control panel is on the left. All graphical elements are interactive and are automatically updated as new data or predictions become available.



Date	GT Keyword: sgh map	GT Keyword: singapore general hospital address	GT Keyword: sgh appointment	GT Keyword: sgh	GT Keyword: sgh a&e	Quantic sgh map	Quantic singapore general hospital address	Quantic sgh	Quantic sgh a&e	Sum of X Variables	Tomorrow's Predicted Actual ED	Tomorrow's Date	Predict ED Visits Refresh
15 Mar 2019	21	8	81	1082	19	194481	4096	1.37059E+12	130321	1.3706E+12	221	16 Mar 2019	
14 Mar 2019	27	4	92	1186	29	531441	256	1.97851E+12	707281	1.97851E+12	210	15 Mar 2019	
13 Mar 2019	18	3	89	817	14	104976	81	4.45542E+11	38416	4.45542E+11	239	14 Mar 2019	
12 Mar 2019	20	5	101	1156	17	160000	625	1.78579E+12	83521	1.78579E+12	214	13 Mar 2019	
11 Mar 2019	12	6	27	639	19	20736	1296	1.66726E+11	130321	1.66726E+11	237	12 Mar 2019	
10 Mar 2019	14	1	53	732	18	38416	1	2.87107E+11	104976	2.87108E+11	237	11 Mar 2019	
09 Mar 2019	16	6	57	1008	15	65536	1296	1.03239E+12	50625	1.03239E+12	224	10 Mar 2019	
08 Mar 2019	14	8	94	1129	21	38416	4096	1.62471E+12	194481	1.62471E+12	215	9 Mar 2019	
07 Mar 2019	21	2	106	1088	17	194481	16	1.40125E+12	83521	1.40125E+12	221	8 Mar 2019	
26 Feb 2019	15	3	252	1231	41	50625	81	2.29632E+12	2825761	2.29632E+12	217	27 Feb 2019	
25 Feb 2019	23	4	107	1124	31	279841	256	1.59612E+12	923521	1.59612E+12	218	26 Feb 2019	
24 Feb 2019	6	4	19	616	24	1296	256	1.43987E+11	331776	1.43987E+11	235	25 Feb 2019	
23 Feb 2019	10	2	30	708	18	10000	16	2.51266E+11	104976	2.51266E+11	235	24 Feb 2019	
22 Feb 2019	21	9	66	1038	27	194481	6561	1.16089E+12	531441	1.16089E+12	224	23 Feb 2019	
21 Feb 2019	24	7	63	995	14	331776	2401	9.8015E+11	38416	9.8015E+11	227	22 Feb 2019	
20 Feb 2019	30	10	70	1057	20	810000	10000	1.24825E+12	160000	1.24825E+12	225	21 Feb 2019	
16 Feb 2019	221	178	173	876	57	2385443281	1003875856	5.88866E+11	10556001	5.92266E+11	320	17 Feb 2019	

FIGURE 3. Sample of the data tab of the *Dashboard* software, displaying an expanded spreadsheet of the search data used as a basis for the predictive model. The tab allows users to manipulate the underlying data directly.

of the basis search data against the predicted trend to verify consistency, and a keyword break-down of search data, along with various controls for each graph. Hovering the cursor over the chart highlights the exact values of data points, and panning and zooming functionalities are enabled. On the left, a filter panel allows users to include or exclude search data of certain date ranges from the model. With appropriate training of the model, the software can allow prediction of patient volumes for the next 1–3 days, and even up to the next year, but accuracy beyond two months cannot be expected due to the inherent unpredictability in disease outbreaks.

2) DATA TAB

The *data* tab brings users to an expanded spreadsheet detailing the retrieved GT data (Figure 3), which underlies any patient volume prediction that the software makes. Further development in the software will allow users to manually force a data refresh, upon which the software will attempt to retrieve updated search data into the database, or force a model refresh, in which the software will re-analyze all collated data to construct a new multiple regression model.

IV. RESULTS & DISCUSSION

A. INTRICACIES IN DATA UTILIZATION

A problem is that the available search volume index (SVI) from GT is not absolute but relative—Google adjusts the volume of query keywords searches to be within a scale of 0–100, relative to the volume of all searches. This normalization process presents a loss of information on the true magnitude of search volume when comparing across different time periods. For instance, an SVI of 50 on some day may represent a true count of, say, 10^6 hits, but on another day might represent a different volume of 10^9 hits, if the total search volume on the second day is much greater than that on the first.

This implies that an increase in SVI may not necessarily represent a true surge in search volume, and similarly a decrease in GT data may not necessarily reflect a true dip—these fluctuations could be attributed to changes in total search volume. To mitigate this, a denormalization by collating data can be carried out over weeks, and scaling all daily SVI by weekly SVI ratios. We note a second problem, in that the adjustment by Google of SVI data to fit within the standard scale of 0–100 also involves a filter, through which volumes deemed too insignificant are suppressed to 0. This presents a considerable information limitation in our context, as only ~18% of our set of relevant query keywords has SVI $>$ 0 over the 2006–2016 period. The denormalization procedure does not aid in this, and the amount of search data useful for statistical analysis and the construction of the regression model is therefore restricted by the SVI adjustment. For the purpose of this preliminary study, we have managed to collect data on a daily basis for the period stipulated in our model.

B. CORRELATION ANALYSES

A comparison of the processed GT search volume data and the actual recorded ED patient volumes in SGH from 07 January to 16 February 2019 is presented in Figure 4. A correspondence between the two can indeed be observed.

The increase in ED cases from 14 January to 18 January is clearly accompanied by a surge in Google queries, indicative of an association between patient influx and heightened public usage of Internet search. Additionally, a peak of 409 recorded ED cases on 08 February also corresponded

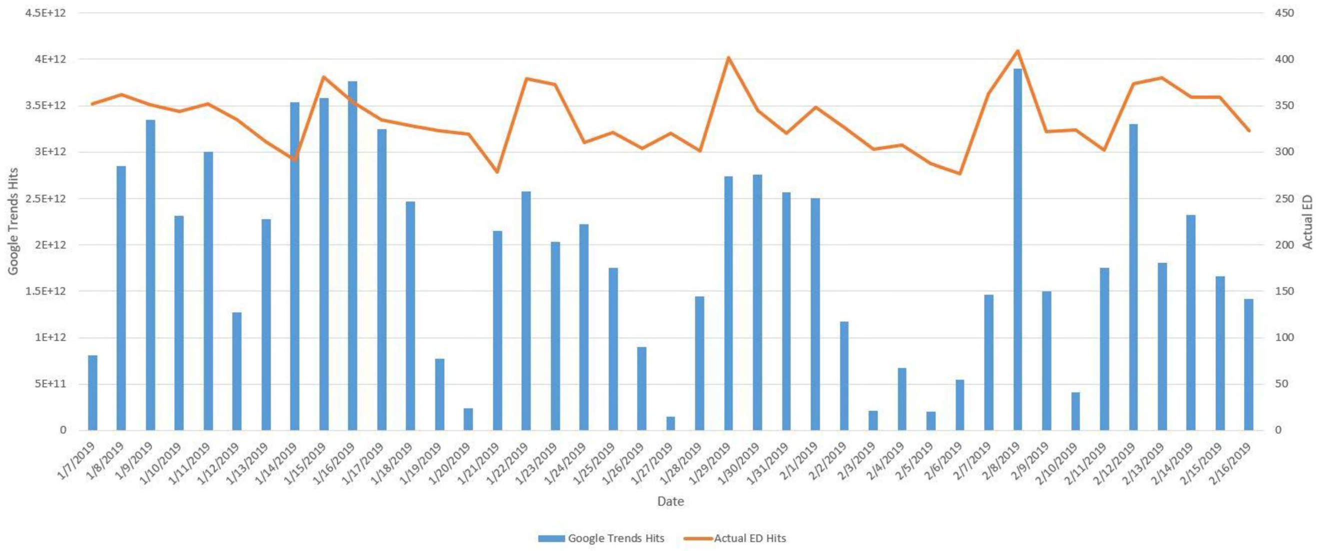


FIGURE 4. **Plot of daily recorded ED visits (orange line) and GT search volume (blue bars) from January 2019 to February 2019. A correlation between the two can be clearly observed.**

with the highest volume of GT hits. Certain patterns in patient volume can also be attributed to local holidays. For instance, the observed plummet from 03 February to 06 February is likely due to the Lunar New Year festive period, and the surge on 07 February and 08 February coincides with the resumption of work and school activities. It is notable that a significant portion of ED cases in SGH, at times accounting for more than half of total patient influx, concerns non-emergency conditions—for instance, gastrointestinal problems comprise $\sim$ 29% of ED cases. These non-emergency cases are expected to be responsible for the surge in ED demand associated with festive seasons and holidays.

To further illustrate the correlation properties, charts of daily patient volume-search volume pairs with Pearson, Kendall, and Spearman correlation analyses are presented in Figure 5. A positive correlation is observed with all three analysis methods, indeed reflecting the consistency in trends observed previously in Figure 4. Individual correlation analysis on query keywords in place of the total search volume was also performed, indicating moderate positive or negative correlation for 20 query keywords and weak or no correlation for remaining keywords.

C. MULTIPLE REGRESSION

Here we demonstrate the construction of the multiple regression predictive model for forecasting of ED patient volume one day in advance ($T - 1$). Multiple regression is performed on a total of 7 predictors—search volume for query keyword "SGH map" (x_1), quartic search volume for "SGH map" (x_2), quartic search volume for "Singapore General Hospital address" (x_3), search volume for "SGH appointment" (x_4), search volume for "Singapore General Hospital address" (x_5), quartic search volume for "SGH" (x_6), and quartic search volume for "SGH A&E" (x_7).

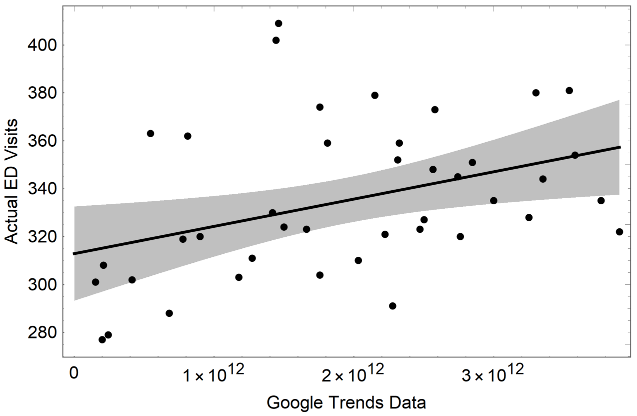


FIGURE 5. **Linear regression between recorded ED visits and GT search volume data. Gray region represents 95% confidence intervals. Pearson, Kendall and Spearman correlation coefficients are $r = 0.38$, $\tau = 0.31$, and $\rho = 0.45$ respectively.**

These predictors are chosen as they yielded the strongest association with actual ED patient volume, as indicated by correlation tests.

Figure 6 presents a multiple regression chart on these predictors. More than half of all data points lie within the 95% confidence intervals, demonstrating reasonable robustness of the regression. Regression results for the seven predictors are presented in Table 3, with overall regression statistics multiple $R = 0.7526$, $R^2 = 0.5664$, adjusted $R^2 = 0.4744$, and standard error of the mean SE $= 23.31$. These statistics indicate satisfactory consistency between the regression model and the recorded data. It is observed from Table 3 that five of the seven predictors (x_1–x_5) have $p \leq 0.05$, whereas the remaining two have large p-values of > 0.05, suggesting that they are statistically insignificant and can be discarded without noticeably affecting regression accuracy.

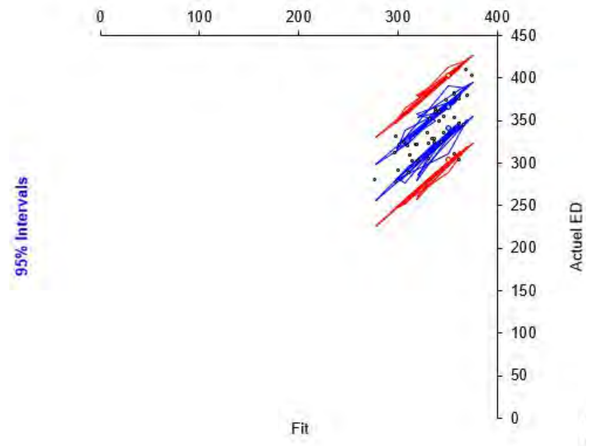


FIGURE 6. Final multiple regression model with data points drawn as black dots.

TABLE 3. Multiple regression results, showing the best-fit values of coefficients β_0–β_7, the corresponding standard errors, and *p*-values.

Coefficient	Value	Standard Error	p-value
β_0	234.0	16.10	6.736×10^{-16}
β_1	0.2263	0.05956	0.0005931
β_2	-8.656×10^{-10}	2.377×10^{-10}	0.0009204
β_3	-3.338×10^{-9}	1.185×10^{-9}	0.008144
β_4	0.09946	0.03538	0.008247
β_5	0.2038	0.07929	0.01486
β_6	-2.016×10^{-11}	1.111×10^{-11}	0.07860
β_7	2.375×10^{-10}	1.467×10^{-10}	0.1151

The predictive multiple regression model is therefore constructed as

$$y = \beta_0 + \sum_{i=1}^{5} \beta_i x_i, \tag{5}$$

where y is the predicted ED patient volume and β_i is the regression coefficient for predictor x_i.

Utilizing this regression model, predictions on next-day ED patient volume can be produced. From the perspective of the *Dashboard* user, the correlation analyses and the construction of the predictive regression model is automatic; though as detailed in Section III-E, the software also presents interfaces for the user to directly observe and manipulate predictor data and the regression model. We present in Figure 7 a comparison of the predicted next-day ED patient volume against actual recorded patient volume. It is clear that prediction accuracy is excellent, with good consistency between the predicted and observed trends over the plotted range of approximately five weeks. To illustrate, the actual and predicted ED patient volume for 13 January were 284 and 291 respectively, reflecting a relative error of 2.5%, and those for 04 February were 285 and 288 respectively, reflecting a 1.1% error.

It is therefore clear that GT search data can be used to effectively predict ED patient volumes. We note that the runtime of the current software solution is $<$ 10 seconds for data update and $<$ 1 minute to construct the statistical model, through a parallelized implementation on a quad-core administrative workstation. Such an overhead can be considered negligible in the daily workflow of hospitals, especially as the patient volume predictions are expected to be run in the background near the end of each working day, to plan manpower in the next. The resource cost of our solution is thus minimal, with other routine administrative processes being vastly more time-demanding in comparison.

While we have detailed next-day predictions, the analysis methodology and the *Dashboard* software are also able to make predictions in advance using appropriate data offsets, though of course with diminishing expected accuracy for farther predictions. It also noted that the software can be configured to lump groups of query keywords into a single predictor, as opposed to the one-to-one approach demonstrated here. Doing so can increase the quality of regression as noise in search data is suppressed, but may also incur a loss in regression robustness if predictors of opposing correlation are grouped together. In the interest of ease of use, the default method is set to treat single query keywords as

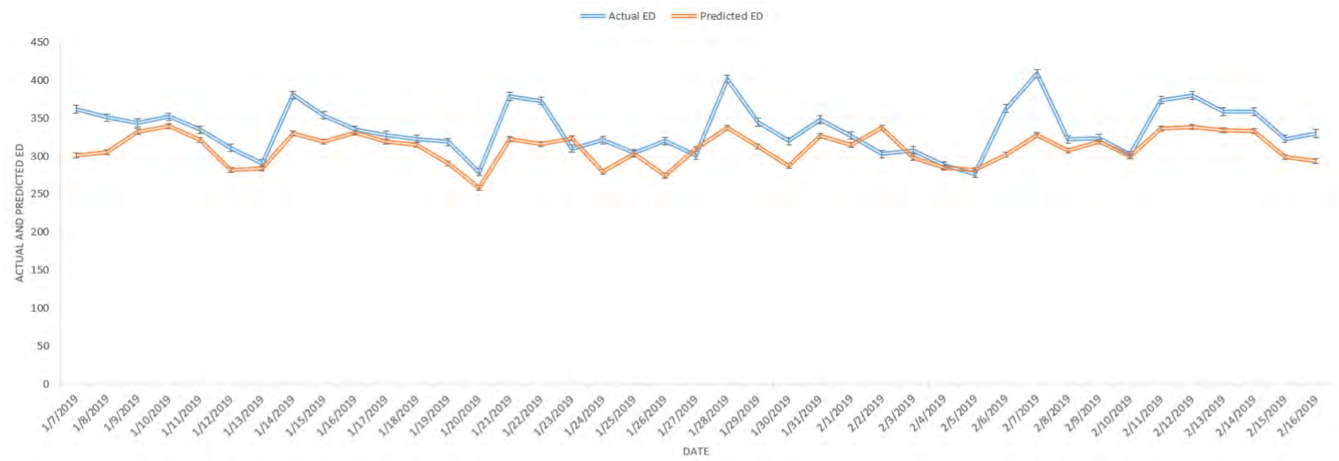


FIGURE 7. Comparison of the actual recorded ED patient volume to next-day predictions made by the regression model, constructed based on realtime GT data, for a time period of January 2019 to February 2019. Good consistency in actual and predicted values is observed.

predictors, unless specified otherwise by the user in the filter panel of *Dashboard*. Lastly, we note a possibility of patient influx variations due to festivities or holiday seasons—the regression model can be made to account for these by including week-wise, month-wise, and year-wise day counts as appropriate.

V. CONCLUSION

This study has demonstrated a feasible method of exploiting publicly available Internet search data to forecast emergency department (ED) patient volume in hospitals. Search volume data were retrieved from Google Trends (GT) and correlation analyses performed to determine key factors associated to patient volume, and regression-based predictive models have been constructed. A software suite has also been developed to make data visualization and forecasting capabilities easily available to hospital medical and administrative staff. The system has been demonstrated in the context of the Singapore General Hospital (SGH), with validation of prediction results against real-world records. The forecasting capability presented here enables informed resource and manpower allocation, and is applicable in hospitals and general medical facilities; in particular, it represents a potential effective measure against worsening ED congestion problems that remain unresolved to date.

Notably, the dynamic regression approach presented is relatively simplistic in comparison to more complex methods [42]–[44], such as neural network-based deep learning, but has been shown to yield sufficient performance, and has the advantage of great transparency and ease of management and adjustment in practical use. These aspects are important, especially for long-term operation in safety-critical medical settings. Our work can improve medical care quality, especially important in emergency cases; and we hope the strong evidence on the wealth of potential in search data analytics indicated here can motivate further research in medical applications.

COMPETING INTERESTS

The authors declare no competing interests.

ACKNOWLEDGMENT

The authors would like to thank Prof Ser Aik Quek for the in-house multiple regression software. They would also like to thank the handling editor and two anonymous reviewers for their constructive comments and suggestions.

REFERENCES

[1] F. Pervaiz, M. Pervaiz, N. A. Rehman, and U. Saif, "Flubreaks: Early epidemic detection from Google flu trends," *J. Med. Internet Res.*, vol. 14, no. 5, Oct. 2015, Art. no. e125.

[2] K. H. Cheong, A. F. W. Ho, H. Zheng, A. Earnest, P. P. Pek, J. Y. Seok, S. K. T. Chan, N. Liu, J. W. C. Tan, T. H. Wong, J. M. Koh, D. Hausenloy, and M. E. H. Ong, "Significant association between transboundary air pollution and acute myocardial infarction," *J. Amer. College Cardiol.*, vol. 73, no. 9, p. 248, Mar. 2019. [Online]. Available: http://www.onlinejacc.org/content/73/9_Supplement_1/248

[3] A. F. W. Ho, H. Zheng, A. Earnest, K. H. Cheong, P. P. Pek, J. Y. Seok, N. Liu, Y. H. Kwan, J. W. C. Tan, T. H. Wong, D. J. Hausenloy, L. L. Foo, B. Y. Q. Tan, and M. E. H. Ong, "Time-stratified case crossover study of the association of outdoor ambient air pollution with the risk of acute myocardial infarction in the context of seasonal exposure to the southeast asian haze problem," *J. Amer. Heart Assoc.*, vol. 8, no. 6, Mar. 2019, Art. no. e011272.

[4] A. F. W. Ho, S. N. N. B. M. Yazid, H. Zheng, A. Earnest, Y. Y. Ng, N. Liu, S. S. W. Lam, M. E. H. Ong, J. M. Koh, and K. H. Cheong, "Abstract 12831: Effects of air pollution and other environmental parameters on all-cause mortality in Singapore," *Circulation*, vol. 138, Nov. 2018, Art. no. A12831.

[5] R. Forero, K. M. Hillman, S. McCarthy, D. M. Fatovich, A. P. Joseph, and D. B. Richardson, "Access block and ED overcrowding," *Emergency Med. Aust.*, vol. 22, no. 2, pp. 119–135, Apr. 2010.

[6] R. Forero, S. Mccarthy, and K. Hillman, "Access block and emergency department overcrowding," *Critical Care*, vol. 15, no. 2, p. 216, Mar. 2011.

[7] G. Braitberg, "Emergency department overcrowding: Dying to get in?" *Med. J. Aust.*, vol. 187, nos. 11–12, p. 624, Dec. 2007.

[8] D. M. Fatovich, G. Hughes, and S. M. McCarthy, "Access block: It's all about available beds," *Med. J. Aust.*, vol. 190, no. 7, p. 362, Apr. 2009.

[9] M. J. Schull, J.-P. Szalai, B. Schwartz, and D. A. Redelmeier, "Emergency department overcrowding following systematic hospital restructuring: Trends at twentyhospitals over ten years," *Acad. Emergency Med.*, vol. 8, no. 11, pp. 1037–1043, Nov. 2001.

[10] M. Mohsin, R. Forero, S. Ieraci, A. E. Bauman, L. Young, and N. Santiano, "A population follow-up study of patients who left an emergency department without being seen by a medical officer," *Emergency Med. J.*, vol. 24, no. 3, pp. 175–179, Mar. 2007.

[11] A. J. Forster, H. J. Murff, J. F. Peterson, T. K. Gandhi, and D. W. Bates, "The incidence and severity of adverse events affecting patients after discharge from the hospital," *Ann. Internal Med.*, vol. 138, no. 3, pp. 161–167, Feb. 2003.

[12] D. M. Fatovich, "Effect of ambulance diversion on patient mortality: How access block can save your life," *Med. J. Aust.*, vol. 183, no. 11, p. 672, Dec. 2005.

[13] D. B. Richardson, "Increase in patient mortality at 10?days associated with emergency department overcrowding," *Med. J. Australia*, vol. 184, no. 5, pp. 213–216, Mar. 2006.

[14] P. C. Sprivulis, J.-A. Da Silva, I. G. Jacobs, G. A. Jelinek, and A. R. Frazer, "The association between hospital overcrowding and mortality among patients admitted via western australian emergency departments," *Med. J. Australia*, vol. 184, no. 5, pp. 208–212, Mar. 2006.

[15] A. F. W. Ho, D. Chew, T. H. Wong, Y. Y. Ng, P. P. Pek, S. H. Lim, V. Anantharaman, and M. E. H. Ong, "Prehospital trauma care in singapore," *Prehospital Emergency Care*, vol. 19, no. 3, pp. 409–415, Dec. 2015. doi: 10.3109/10903127.2014.980477.

[16] M. D. Barr, "Singapore: The limits of a technocratic approach to health care," *J. Contemp. Asia*, vol. 38, no. 3, pp. 395–416, May 2008.

[17] J. P. Ansah, R. L. Eberlein, S. R. Love, M. A. Bautista, J. P. Thompson, R. Malhotra, and D. B. Matchar, "Implications of long-term care capacity response policies for an aging population: A simulation analysis," *Health Policy*, vol. 116, no. 1, pp. 105–113, May 2014.

[18] S. P. Low, S. Gao, and G. Q. E. Wong, "Resilience of hospital facilities in Singapore's healthcare industry: A pilot study," *Int. J. Disaster Resilience Built Environ.*, vol. 8, no. 5, pp. 537–554, Nov. 2017.

[19] J. M. Pines and J. E. Hollander, "Emergency department crowding is associated with poor care for patients with severe pain," *Ann. Emergency Med.*, vol. 51, no. 1, pp. 1–5, Jan. 2008.

[20] J. E. Wilson and J. M. Pendleton, "Oligoanalgesia in the emergency department," *Amer. J. Emergency Med.*, vol. 7, no. 6, pp. 620–623, Nov. 1989.

[21] R. Stalnikowicz, R. Mahamid, S. Kaspi, and M. Brezis, "Undertreatment of acute pain in the emergency department: A challenge," *Int. J. Qual. Health Care*, vol. 17, no. 2, pp. 173–176, Feb. 2005.

[22] T. Rupp and K. A. Delaney, "Inadequate analgesia in emergency medicine," *Ann. Emergency Med.*, vol. 43, no. 4, pp. 494–503, Aug. 2004.

[23] S. Agrawal, "Emergency department crowding: An ethical perspective.," *Acad. Emergency Med.*, vol. 14, no. 8, pp. 750–751, 2007.

[24] C. Van Der Linden, R. Reijnen, R. W. Derlet, R. Lindeboom, N. Van Der Linden, C. Lucas, and J. R. Richards, "Emergency department crowding in The Netherlands: Managers' experiences," *Int. J. emergency Med.*, vol. 6, no. 1, p. 41, Oct. 2013.

[25] R. J. Salway, R. Valenzuela, J. M. Shoenberger, W. K. Mallon, and A. Viccellio, "Emergency department (ED) overcrowding: Evidence-based answers to frequently asked questions," *Revista Médica Clínica Las Condes*, vol. 28, no. 2, pp. 213–219, Mar./Apr. 2017.

[26] J. Orr, "The good, the bad, and the four hour target," *BMJ*, vol. 337, p. a195, Jul. 2008.

[27] K. Liu, T. Wang, Z. Yang, X. Huang, G. J. Milinovich, Y. Lu, Q. Jing, Y. Xia, Z. Zhao, and Y. Yang, S. Tong, W. Hu, and J. Lu, "Using Baidu search index to predict dengue outbreak in China," *Scientific reports*, vol. 6, Dec. 2016, Art. no. 38040.

[28] G. He, Y. Chen, B. Chen, H. Wang, L. Shen, L. Liu, D. Suolang, B. Zhang, G. Ju, L. Zhang, S. Du, X. Jiang, Y. Pan, and Z. Min, "Using the Baidu search index to predict the incidence of HIV/AIDS in China," *Sci. Rep.*, vol. 8, no. 1, Jun. 2018, Art. no. 9038.

[29] X. Huang, L. Zhang, and Y. Ding, "The Baidu index: Uses in predicting tourism flows—A case study of the forbidden city," *Tourism Manage.*, vol. 58, pp. 301–306, Feb. 2017.

[30] L. Vaughan and Y. Chen, "Data mining from Web search queries: A comparison of google trends and baidu index," *J. Assoc. Inf. Sci. Technol.*, vol. 66, no. 1, pp. 13–22, Jan. 2015.

[31] A. L. Beam and I. S. Kohane, "Big data and machine learning in health care," *Jama*, vol. 319, no. 13, pp. 1317–1318, Apr. 2018.

[32] Z. Obermeyer and E. J. Emanuel, "Predicting the future—Big data, machine learning, and clinical medicine," *New England J. Med.*, vol. 375, no. 13, pp. 1216–1219, Sep. 2016.

[33] N. Jothi and W. Husain, "Data mining in healthcare—A review," *Procedia Comput. Sci.*, vol. 72, pp. 306–313, Jan. 2015.

[34] R. T. Gluskin, M. A. Johansson, M. Santillana, and J. S. Brownstein, "Evaluation of internet-based dengue query data: Google dengue trends," *PLoS ONE*, vol. 8, no. 2, Feb. 2014, Art. no. e2713.

[35] B. M. Althouse, Y. Y. Ng, and D. A. Cummings, "Prediction of dengue incidence using search query surveillance," *PLoS ONE*, vol. 5, no. 8, Aug. 2011, Art. no. e1258.

[36] H. A. Carneiro and E. Mylonakis, "Google trends: A Web-based tool for real-time surveillance of disease outbreaks," *Clin. Infectious Diseases*, vol. 49, no. 10, pp. 1557–1564, Nov. 2009.

[37] S. Yang, M. Santillana, and S. C. Kou, "Accurate estimation of influenza epidemics using google search data via argo," *Proc. Nat. Acad. Sci. USA*, vol. 112, no. 47, pp. 14473–14478, Nov. 2015.

[38] J. Bellika, A. Bravo-Salgado, M. Brezovan, D. D. Burdescu, J. Chartree, J. C. Denny, L. Ferreira, K. Ghazinour, L. Hanlen, and S. S. Imambi, *Text Mining of Web-Based Medical Content*, vol. 1. Berlin, Germany: Walter de Gruyter, 2014.

[39] V. Lampos and N. Cristianini, "Tracking the flu pandemic by monitoring the social Web," in *Proc. 2nd Int. Workshop Cognit. Inf. Process.*, Jun. 2010, pp. 411–416.

[40] S.-D. Bolboaca and L. Jäntschi, "Pearson versus Spearman, Kendall's tau correlation analysis on structure-activity relationships of biologic active compounds," *Leonardo J. Sci.*, vol. 5, no. 9, pp. 179–200, Jul. 2006.

[41] M. E. Mendenhall, T. M. Kuhlmann, G. K. Stahl, and J. S. Osland, "Employee development and expatriate assignments," in *Proc. Blackwell Handbook Cross-Cultural Manage.*, 2002, pp. 155–183.

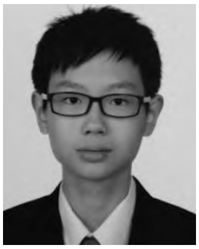
[42] R. Burbidge, M. Trotter, B. Buxton, and S. Holden, "Drug design by machine learning: Support vector machines for pharmaceutical data analysis," *Comput. Chem.*, vol. 26, no. 1, pp. 5–14, Dec. 2001.

[43] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, *Data Mining: Practical Machine Learning Tools and Techniques*. San Mateo, CA, USA: Morgan Kaufmann, 2016.

[44] C. H. Lee and H.-J. Yoon, "Medical big data: Promise and challenges," *Kidney Res. Clin. Pract.*, vol. 36, no. 1, pp. 3–11, Mar. 2017.

ANDREW FU WAH HO received the M.B.B.S., M.R.C.E.M., and M.M.ED. degrees. He is the Chief Resident and Clinician-Scientist Track Resident in the SingHealth Duke-NUS Emergency Medicine Academic Clinical Programme. He maintains an active clinical practice across the SingHealth group of hospitals. His research interests include resuscitation, pre-hospital emergency care, and health services research.

BRYAN ZHAN YUAN SE TO received the B.Sc. degree (Hons.) in computing science from the University of Glasgow (in partnership with the Singapore Institute of Technology), in 2019. His research interests include data analytics and machine learning in health services research.



JIN MING KOH received the NUS High School Diploma (High Distinction) in 2016. Since 2017, he has been undertaking research projects offered by K. H. Cheong.

KANG HAO CHEONG (M'18) received the B.Sc. degree (Hons.) from the Department of Mathematics and University Scholars Programme, National University of Singapore (NUS), in 2007, the Ph.D. degree from the Department of Electrical and Computer Engineering, NUS, in 2015, and the Postgraduate Diploma degree in education from the National Institute of Education, Singapore. He was an Assistant Professor of engineering with the Singapore Institute of Technology, from 2016 to 2018. He is currently an Assistant Professor in the Science and Math Cluster with the Singapore University of Technology and Design.