2021 IEEE 7th International Conference on Big Data Intelligence and Computing (DataCom) | 978-1-6654-2071-6/21/$31.00 ©2021 IEEE | DOI: 10.1109/DataCom53700.2021.00009

Big Data Intelligence Solution for Health Analytics of COVID-19 Data with Spatial Hierarchy

Carson K. Leung (✉), Chenru Zhao
University of Manitoba, Winnipeg, MB, Canada
✉ kleung@cs.umanitoba.ca

***Abstract*—In the current era of big data, technological advancements have made it easy and quick to generate and collect huge volumes of varieties of data from wide ranges of rich data sources. These big data may be of different levels of veracity, including precise data and imprecise or uncertain data. Embedded in the data are valuable information and useful knowledge that can be discovered by big data intelligence and computing. In this paper, we propose a big data intelligence solution for health analytics with spatial hierarchy. In particular, we focus on analyzing coronavirus disease 2019 (COVID-19) epidemiological data at different spatial granularity levels. Since its outbreak, there have been cumulatively millions of COVID-19 cases observed in various spatial locations around the world. Our solution focuses on analyzing and mining valuable information and useful knowledge (e.g., distribution, frequency, patterns) of health-related states and characteristics in populations in various spatial locations in a top-down fashion along the spatial hierarchy. To reduce redundancy, our solution discovers and returns to users (e.g., researcher, civilian) new information and knowledge not found at previous spatial hierarchical levels. The discovered information and knowledge helps the users to understand the disease better, and thus take an active role to fight, control, and/or combat the disease. Evaluation of our big data intelligence solution on real-life COVID-19 data demonstrates its practicality in health analytics of the data with spatial hierarchy and in revealing new knowledge about COVID-19 cases at different spatial granularity levels. The solution is expected to be adaptable to health analytics of other diseases.**

Keywords*—*big data, big data intelligence, big data intelligence and computing, data science, data mining, coronavirus disease, COVID-19, spatial data

I. Introduction and Related Works

Big data [1-3] are everywhere nowadays. This is partially due to technological advancements, which in turn have led to easy production and collection of huge volumes of varieties of valuable data at high velocities. These big data have been produced and collected from wide ranges of rich data sources. The big data may also be of different levels of veracity, including precise data and imprecise or uncertain data. This explains why big data—as characterized by the 5Vs of the big data landscape (i.e., volume, variety, value, velocity, and veracity)—has been popular and become one of rapidly expanding research areas spanning the fields of computer science and information management. Moreover, the term "big data" has also become a ubiquitous term in understanding and solving complex problems in various disciplinary fields like applied mathematics, business, computational biology, education, engineering, finance, government, healthcare, medicine, social networks, telecommunications, and transportation.

Valuable information and useful knowledge embedded in the big data can be discovered by *big data intelligence and computing*. It can be run in conjunction with data science [4-6], data mining (e.g., clustering [7, 8], classification [9], graph mining [10, 11], pattern mining [12-15], sequential mining [16, 17], stream mining [18, 19]), machine learning [20-22], data analytics [23-26], visual analytics [27-30], social network analysis [31-35], mathematical and statistical modeling [36], etc., for social good.

Let us consider a few real-life examples in several real-life application areas. As a first example, conducting *big social data analytics* [37, 38]—such as census microdata on home languages—allow users (e.g., social scientists, decision makers) to get a better understanding of the data. This helps them in the study of social science related phenomena (e.g., successful detection of shifts in home languages) and inspire them to take appropriate actions to residents in various communities (e.g., by providing adequate support or services in their preferred languages).

As another example, conducting *transportation and/or urban data analytics* [39-43]—such as analyses and mining on data on commuting mode—helps users (e.g., social scientists, researchers, city planners, policy makers) to get a better understanding of the demand of commuters in different neighborhoods, which may inspire them to come up ways to fulfill demands of commuters, improve existing services, and/or add new services. It may also inspire users (e.g., residents) to consider active transportation modes (e.g., cycling, walking) and/or sustainable transportation modes (e.g., carpooling, public transit), and thus preserving our environments.

As a third example, conducting *business data mining and/or business data analytics* [44] allows users (e.g., business owners) to get a better understanding of their business data or transactions. This helps them to recruit and retain customers by taking appropriate actions in fulfilling the demand of customers in different geographical locations.

As a fourth example, *mining health data and/or disease reports* [45-47]—such as coronavirus disease 2019 (COVID-19) epidemiological data—for health analytics or health informatics allows users (e.g., health scientists, decision makers)

© IEEE 2022. This article is free to access and download, along with rights for full text and data mining, re-use and analysis.

DOI 10.1109/DataCom53700.2021.00009

to get a better understanding of the disease. This inspires them to come up with big data intelligence and computing solutions to flight, control and/or combat the disease. It also allows them to prepare for adequate resources (e.g., sufficient staff and beds in regular wards or intensive care units (ICU) in hospitals) to meet the demands in different spatial locations. By doing so, it improve our health and well-being.

For big data in the aforementioned four examples and many real-life applications, there is a common aspect—namely, spatial component. To elaborate, values of many attributes in these social data (e.g., census microdata), transportation data (e.g., on-time performance for public transit buses), business data (e.g., sales transactions in various businesses), and/or health data (e.g., spatial changes or trends in COVID-19 epidemiological data) may be stable or may fluctuate among different spatial locations. Big data intelligence and computing—especially, spatial data analytics—helps detects and discovers characteristics of the data and their changes among different spatial locations.

For instance, in business world, sales of certain merchandise items or products (e.g., essentials food items such as bread and milk, formal clothes like ties and suits) can be similar among various spatial locations. Sales of other items or products may vary from locations to locations (e.g., customers living in warm and humid coastal regions may buy raincoats or rain boots, whereas customers living in cool and dry prairies may buy puffer jackets and snow boots). Hence, on the one hand, mining these data at a coarse granularity level gives users may only provide a few summary patterns but insufficient details to see the differences. For example, mining national data gives users (e.g., business owners) a summary information—say, national sales—in merchandise items or products. However, such a summary may not provide sufficient details on the customer demand on certain geographical locations. Lack of the information may lead to unnecessary storage for some items (e.g., storing lots of snow boots for warm and humid coastal regions). The matter may be worse for fragile and/or perishable items. On the other hand, mining these data at a fine granularity level requires mining solution to deal with huge volumes of data. The mining results provide users with abundant of patterns, which may sometimes be excessive and time-consuming to comprehend. For example, mining data from every municipality may lead to collections of patterns from more than 5,000 municipalities in Canada.

Similar comments apply to medical world. For instance, as of October 10, 2021, there have been more than 1.6 million COVID-19 cases in Canada and close to 237 million COVID-19 cases worldwide. Out of them, more than 28,000 Canadians lost their life to COVID-19 and close to 5 million global citizens deceased due to COVID-19. Mining these huge volumes of COVID-19 epidemiological data for some useful patterns (e.g., characteristics of COIVD-19 cases) require scalable solution. Moreover, as characteristics of COIVD-19 cases may vary from locations to locations, big data intelligence and computing solution for *spatial data mining or data analytics in supporting health analytics* is needed. When dealing with spatial data, a logical question to ask is: At what spatial granularity should we conduct the mining? On the one hand, mining data for the whole dataset gives users (e.g., health scientists, decision makers, civilian) a summary information. However, such a summary may not provide sufficient details (e.g., differences in the demand for spots in the hospitals and/or ICU at different regions). Lack of the information may lead to undesirable management of healthcare staff and/or health supplies. On the other hand, mining local data gives users abundant of patterns, which may sometimes be excessive and time-consuming to comprehend. As such, some important patterns may be hidden in the haystack of patterns and may be oversighted by the users.

Faced with these challenges, we present in this paper a big data intelligence and computing solution. It analyzes and mines COVID-19 epidemiological data for spatial characteristics of COVID-19 cases. It provides users with patterns at several appropriate levels of granularity in the spatial hierarchy. Hence, our ***key contributions*** of this work include our design and development of such a big data intelligence and computing solution for health analytics of COVID-19 data with the spatial hierarchy. It incorporates external data like population data to mine and analyze COVID-19 epidemiological data at multiple granularity levels of the spatial hierarchy.

We organize the rest of this paper by providing background and discussing related works in the next section. Section III describes the design of our big data intelligence and computing solution for health analytics of COVID-19 data with the spatial hierarchy. Section IV shows evaluation results of our implemented solution on real-life COVID-19 epidemiological data. Last but not least, we draw conclusions in Section V.

II. Background and Related Works

We aim to design a big data intelligence and computing solution for health analytics. For demonstration in this paper, we apply the resulting solution to COVID-19 epidemiological data. Like (a) the severe acute respiratory syndrome (SARS) that broke out during 2002–2004 and (b) the Middle East respiratory syndrome (MERS) that broke out during 2012–2015, COVID-19 is also a viral disease. Specifically, COVID-19—which was formerly known as 2019 novel coronavirus (2019-nCoV) and 2019-nCoV acute respiratory disease—is caused by SARS coronavirus 2 (SARS-CoV-2). Unlike SARS and MERS that affected only a certain number of countries and regions, COVID-19 affects worldwide. Specifically, it was reported in late 2019. The World Health Organization (WHO) declared it as a Public Health Emergency of International Concern on January 30, 2020, and later declared it as a global pandemic on March 11, 2020. Sadly, COVID-19 is still prevailing in October 2021.

Due to its global impacts, there have been numerous works related to COVID-19. To name a few, in social sciences, researchers studied crisis management related to the COVID-19 outbreak [48]. In medical and health sciences, researchers examined how to manage clinical and treatment information [49]. Some researchers developed vaccine [50]. In natural sciences and engineering aspect, researchers have explored techniques data mining, data science, machine learning, mathematics, and/or statistics to contribute to the COVID-19 research. For instance, some researchers focused on artificial intelligence (AI)-driven informatics to track, test, diagnose, and/or treat COVID-19. These include the detection of COVID-19 cases by AI-driven analyses of chest computed tomography (CT) images of potential COVID-19 patients [51].

In contrast, we examine huge volumes of alphanumeric COVID-19 epidemiological data (cf. CT images). Regarding related works on COVID-19 epidemiological data, many of them—especially, those notable dashboards—aim to report the numbers of new cases and deaths. Visualizing these numbers in graphical forms may make it easier for laypersons to comprehend the information. However, when visualizing the information with *bubble maps*, the numbers (of new cases or deaths) are indicated by radii of the bubbles. As such, bubbles may overlap with, or contain, some other bubbles. Similarly, when visualizing the information with *choropleth maps*, the numbers (of new cases or deaths) are indicated by shading or coloring. The darker the shade or color, the higher is the number. As such, small locations may not be easily visible.

In addition to the aforementioned number cases and deaths, *common characteristics associated with these COVID-19 cases* are also important. For example, it is to desirable to mine the COVID-19 epidemiological data for revealing useful knowledge like their transmission methods that led to COVID-19, whether or not they show any symptoms (i.e., symptomatic vs. asymptomatic), their hospitalization requirements (e.g., ICU, regular wards, no hospitalization). To elaborate, knowing the transmission methods help the users (e.g., decision makers, health officers) to take appropriate actions to break the transmission links. Knowing information (e.g., symptoms) about the symptomatic cases helps the users (e.g., healthcare providers) to detect COVID-19 cases, whereas knowing information about the asymptomatic cases helps the users to take appropriate actions to prevent these asymptomatic cases from spreading the disease (e.g., by self-quarantine). Similarly, knowing hospitalization requirements helps the users to plan for the potential demand from patients. Consequently, there have been some works on analyzing and visualizing these characteristics of COVID-19 cases [52-55].

Moreover, it is also desirable to mine the COVID-19 epidemiological data for revealing additional useful knowledge like changes (e.g., spatial changes) in characteristics associated with the COVID-19 cases. These changes can show the development of COVID-19 and measure the effects of the actions (e.g., lockdowns, physical or social distancing, stay-at-home orders) and/or at different geographical locations. Consequently, there have been some works on spatial analytics and temporal analytics of COVID-19 cases [54, 55]. However, they mostly focused on a *single* granularity level (e.g., national, provincial, or regional differences; but not their combinations in different levels). In contrast, we examine COVID-19 data at *multiple* spatial granularity levels.

III. Our Big Data Intelligence and Computing Solution

To support health analytics of COVID-19 epidemiological data, our big data intelligence and computing solution first builds a spatial hierarchy. It then incorporates population data to analyze COVID-19 epidemiological data for the discovery of useful patterns at multiple demographic granularity levels in this demographic hierarchy.

A. Spatial Hierarchy

Our solution first builds a spatial hierarchy to support health analytics of COVID-19 epidemiological data. Here, local data may be generated and collected from local healthcare providers or health service facilities (e.g., clinics, hospitals). These data can then be gathered and reported to higher-up in the spatial hierarchy. For instance, the local data can be generalized into municipal health data and/or municipal COVID-19 epidemiological data. These data can then be generalized into data for the corresponding health administrative units within a province or state. The provincial data can be grouped and generalized into data for a specific national regions, and then country and continent. Finally, at the top of the spatial hierarchy would be the worldwide data. See Fig 1. Depending on the application and/or focus, users have flexibility to pick and choose appropriate layers in the spatial hierarchy.

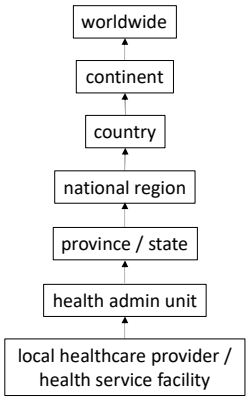

Fig. 1. General spatial hierarchy for health analytics of data

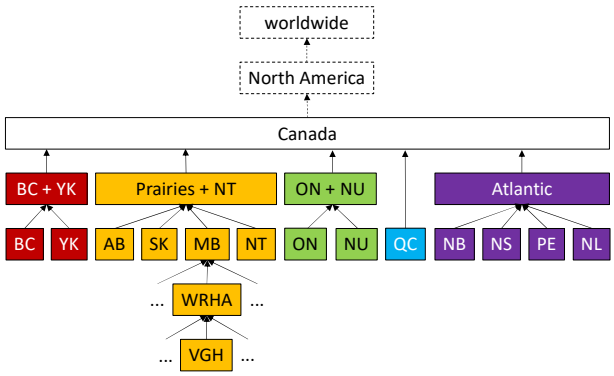

Fig. 2. Specific spatial hierarchy for health analytics of Canadian COVID-19 data

As a preview of our evaluation, we apply our big data intelligence and computing solution to real-life Canadian COVID-19 epidemiological data. Fig. 2 shows the corresponding spatial hierarchy for health analytics of Canadian COVID-19 data. Here, local data were generated and collected from local health service facilities (e.g., St. Boniface Hospital, Victoria General Hospital (VGH)). These data were then gathered to become data for a provincial health region administrated by a provincial health services authority. Examples include Toronto Central Local Health Integration Networks in the province of Ontario (ON), Winnipeg Regional Health Authority (WRHA) in the province of Manitoba (MB),

and Vancouver Coastal Health in the province of British Columbia (BC). These data can be generalized into provincial data, which can then be generalized into data for five national regions: (1) Atlantic, (2) Quebec (QC), (3) ON + Nunavut (NU), (4) Prairies + Northwest Territories (NT), and (5) BC + Yukon (YK). These regional data are then generalized into Canadian national data at the top of the spatial hierarchy. The hierarchy could be further extended to include North American and global worldwide data.

B. Incorporation with Population Data

Observed that population is not evenly distributed. Spatial regions with higher numbers of population may have higher chances of having more COVID-19 cases. Hence, after building the spatial hierarchy, our solution incorporates population data. By doing so, it shows both (1) the absolute frequency and (2) the relative frequency with respect to its population.

C. Health Analytics at Multiple Spatial Granuarity Levels

After building the spatial hierarchy and incorporating the population data, our solution analyzes and mines data at different spatial granularity levels in this hierarchy. To elaborate, mining patterns at the top (i.e., coarsest) level of the hierarchy provides users with overview of the COVID-19 situation and relevant epidemiological information at the top level in the taxonomy (e.g., entire country or worldwide). However, the COVID-19 situations may not necessarily evenly distributed among all spatial locations at lower (i.e., finer) levels of the hierarchy. For example, the summary characteristics of Canadian COVID-19 cases may not reflect the local characteristics of Manitoban cases. Moreover, the summary characteristics are usually dominated by the observed characteristics from those with highest numbers of cases (e.g., Quebec). As such, it is desirable to be able to mine patterns at lower levels of the hierarchy.

However, doing so may incur high computational costs and result in numerous mined patterns and due to numerous numbers of spatial units (e.g., more than 70 local health service facilities managed by WRHA) at these lower levels. Although summary characteristics of COVID-19 cases at higher level may not reflect the local characteristics in all these finer-grained spatial units, some may follow the same or similar patterns as their higher levels in the hierarchy.

Hence, we design and develop our solution so that it returns patterns to users if these patterns are different from (i.e., not covered by) those mined from their parents or ancestors in the hierarchy. Here, we making the following observations.

Observation 1. Pattern mining at each spatial unit at a fine granularity level can be performed independently.

Observation 2. The frequency of a pattern at a coarser granularity in a spatial hierarchy can be obtained by aggregating (e.g., summing) the frequencies of the pattern of the corresponding time units at a finer granularity. For instance, frequency of Atlantic Provinces can be obtained by summing the frequencies of the four corresponding provinces.

Observation 3. Patterns that are frequent at a coarser granularity in a spatial hierarchy must be frequent locally in at least one of the corresponding spatial units at a finer granularity.

Based on these observations, our solution mines patterns (e.g., frequent patterns) as follows. Due to Observation 1, our solution achieves scalability by conducting spatial analytics from finer grained data *in parallel*. Then, based on Observation 2, once patterns are discovered, their associated information (e.g., frequencies of patterns) is aggregated to a higher spatial granularity level. By doing so, it saves extra efforts in re-computing frequencies of patterns at higher granularity levels from scratch. Moreover, based on Observation 3, patterns that are frequent at a coarser granularity in a hierarchy must be frequent locally in at least one of the corresponding spatial units at a finer granularity. Hence, once we gathered patterns that are frequent at the coarser granularity, our solution checks the aggregated frequencies to determine if the patterns are frequent at the coarser granularity:

- If not, our solution returns the patterns discovered from the finer granularity to the users as *outlying or exceptional patterns* that are not covered by patterns discovered from the coarser granularity.
- If so, our solution *recursively* applies similar process to aggregate frequencies to the next coarser granularity level (until it reaches the top level) and determines whether or not the patterns at the level are frequent.

With this recursive approach, our solution first returns patterns at the top (i.e., coarsest) level of the hierarchy, which provide users with overview (e.g., of the COVID-19 situation and relevant epidemiological information in our evaluation application) at the top level in the taxonomy (e.g., the entire country). Then, it also discovers any patterns (e.g., outlying or exceptional patterns) from finer granularity levels that are not covered by patterns discovered from coarser granularity levels.

As an extension, since we have the frequency information, our solution can also return to users any frequent patterns with large differences in frequencies (especially, relative frequencies) when compared with frequent patterns discovered at coarser levels. By doing so, it detects patterns with significant changes in frequencies.

IV. Evaluation

A. Setup

For evaluation, we applied our solution to analyzing real-life Canadian COVID-19 epidemiological data[1, 2] [56], including data on VOC[3] and vaccination[4]. We also incorporated Canadian population data [57] (especially quarterly population estimate for Q3 of 2021). Note that, although we applied to Canadian data, our solution is applicable to data from other countries

[1] https://www.ctvnews.ca/health/coronavirus/tracking-every-case-of-covid-19-in-canada-1.4852102

[2] https://health-infobase.canada.ca/covid-19/epidemiological-summary-covid-19-cases.html

[3] https://www.ctvnews.ca/health/coronavirus/tracking-variants-of-the-novel-coronavirus-in-canada-1.5296141

[4] https://www.ctvnews.ca/health/coronavirus/coronavirus-vaccination-tracker-how-many-people-in-canada-have-received-shots-1.5247509

and/or regions. Here, the COVID-19 epidemiological data capture various information about COVID-19 cases, which include:

- Spatial details (e.g., episode provincial health services authority, province, national region)
- transmission methods (e.g., domestic acquisition, international travel)
- indicating sign of the disease (e.g., asymptomatic, symptomatic)
- COVID variants[5], which include (a) variants of concern (VOC), (b) variants of interest (VOI), and (c) variants under monitoring (VUM). Examples of VOC include alpha (e.g., Phylogenetic Assignment of Named Global Outbreak (PANGO) lineage B.1.1.7 and its descendent Q lineages[6]), beta (B.1.351 and its descendent lineages), gamma (P.1 and its descendent lineages), and delta (e.g., B.1.617.2 and its descendent AY lineages). Omicron (e.g., B.1.1.529 and BA lineages) was recently added to the list of VOC on November 26, 2021. Lambda (lineages C.37 and C.37.1) and mu (B.1.621 and B.1.621.1) are examples of VOI. Note that the former VOI kappa (B.1.617.1) is an example of current VUM. However, other former VOI—such as epsilon (B.1.427 & B.1.429), zeta (P.2), eta (B.1.525), theta (P.3), and iota (B.1.526)—are not.
- Vaccination status (e.g., ineligible for vaccination; unvaccinated; received first, second, and/or third doses)
- hospital status (e.g., ICU, regular ward, not hospitalized)
- recovery status (e.g., recovered, deceased)

There are unstated values (which would be skipped for our evaluation) for some records in these data. This may partially due to privacy concerns and fast dissemination on reporting the cases.

B. Results and Observations

With the aforementioned evaluation setup, we first built a spatial hierarchy as depicted in Fig. 1 to capture information related to Canadian COVID-19 cases from January 25, 2020 (first COVID-19 case in Canada) to October 10, 2021. With data (and aggregated data) in this hierarchy, we applied our big data intelligence and computing solution to discover patterns from multiple granularity levels. As of October 10, 2021, we observed from the **top (i.e., national) level** of the spatial hierarchy:

- There have been 1,662,584 cumulative COVID-19 cases in Canada. By incorporating population data, we observed that these 1,662,584 cases account for 4.3% of the entire Canadian population. Moreover, there have been 28,203 cumulative deaths, which give a national case fatality rate is 1.7%.
- 99% of the Canadian COVID-19 cases were acquired domestically through the community. Among them, (a) 92% did not require hospitalization, and their recovery rate is 0.99; (b) 6% were admitted to regular wards in the hospital, and their recovery rate is 0.85; and (c) 2% were admitted to the ICU, and their recovery rate is 0.70.
- 417,047 (i.e., 25%) of the Canadian COVID-19 cases have been identified as VOC. Among them, 64% were alpha and 30% were delta, with the remaining 5% and 1% as gamma and beta variants respectively.
- 76.6% of all Canadians have been vaccinated with one dose, and 71.7% of all Canadians have been fully vaccinated with at least two doses. When considering only those eligible for vaccination (i.e., aged 12^{+}), 87.6% and 82.0% were vaccinated with one and at least two doses respectively. Moreover, less than 1% of Canadians received their third dose.

Next, we moved down to a finer granularity level in the hierarchy. At the five (national) **regional level**, we observed the following that are not covered by (or different from) the aforementioned observations:

- In terms of percentage of COVID-19 cases with respect to the regional population (see Fig. 3), Atlantic Provinces have a much lower infection rate (of 0.6%) than the national rate (of 4.3%), whereas Prairies + NT have a much higher infection rate (of 6.3%). As shown in Table I, despite their outlying infection rates, case fatality rates for both regions were similar to the national rate.

  Quebec, on the other hand, has an infection rate (of 4.8%) similar to the national rate, its case fatality rate (of 2.7%) is much higher than the national rate (of 1.7%).

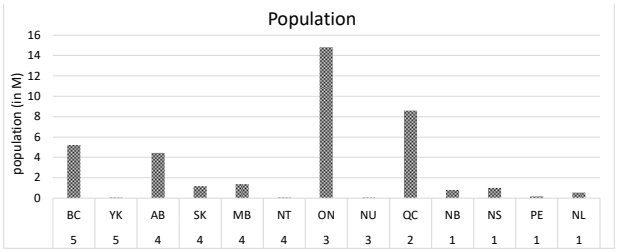

Fig. 3. Population of 13 provinces & territories within the 5 national regions

TABLE I. INFECTION AND CASE FATALITY RATES OF 5 NATIONAL REGIONS

	Infection rate (i.e., cases ÷ population)	**Case fatality rate (i.e., deaths ÷ population)**
1 Atlantic	0.6%	1.3%
2 Quebec	4.8%	2.7%
3 Ontario + Nunavut	4.0%	1.7%
4 Prairies + NT	6.3%	1.1%
5 BC + Yukon	3.7%	1.0%
Canada	4.3%	1.7%

[5] https://www.who.int/en/activities/tracking-SARS-CoV-2-variants/

[6] https://cov-lineages.org/lineage_list.html

- In terms of transmission methods, 97% and 93% of COVID-19 cases in Quebec and Atlantic Provinces respectively were acquired domestically (i.e., 3% and 7% of their cases were imported through international travel). Between them, hospital status for cases in Atlantic Provinces was similar to the national figure. However, in Quebec, 87% of domestically acquired cases did not require hospitalization (with a recovery rate dropped to 92%), 10% were admitted to regular wards (with a recovery rate dropped to 72%). See Tables II and III.

  For Ontario + Nunavut, despite its percentage of domestic acquisition matches the national percentage (of 99%), it has the lowest non-hospitalization percentage (of 80%) and highest hospitalization percentage (with 16% in regular wards + 4% to the ICU) among domestic acquired cases in the five national regions.

TABLE II. BREAKDOWN OF HOSPITAL STATUS OF DOMESTICALLY ACQUIRED CASES IN THE 5 NATIONAL REGIONS

	No hospitalized	Regular wards	ICU	
1 Atlantic	95%	4%	1%	100%
2 Quebec	87%	10%	2%	100%
3 ON + NU	80%	16%	4%	100%
4 Prairies + NT	96%	4%	1%	100%
5 BC + YK	95%	4%	1%	100%
Canada	92%	6%	2%	100%

TABLE III. RECOVERY RATE OF DOMESTICALLY ACQUIRED CASES WITH 3 DIFFERENT HOSPITAL STATUS IN THE 5 NATIONAL REGIONS

	No hospitalized	Regular wards	ICU
1 Atlantic	99%	91%	80%
2 Quebec	92%	72%	70%
3 ON + NU	99%	86%	66%
4 Prairies + NT	100%	86%	71%
5 BC + YK	99%	92%	78%
Canada	99%	85%	70%

- Quebec is the only region with VOC distribution percentage that was similar to the national percentage. To elaborate, Table IV reveals that 90% of variants in Atlantic Provinces were alpha and 5% were delta; 85% of variants in Ontario + Nunavut were alpha and 11% were delta. The ratio of alpha to delta cases dropped to almost equal (with 51% alpha + 46% delta) in Prairies + NT. In contrast, BC + YK have more (49% of variants) delta and fewer (28%) alpha variants.

TABLE IV. BREAKDOWN OF COVID-19 VARIANTS OF CONCERNS IN THE 5 NATIONAL REGIONS

	Alpha	Beta	Gamma	Delta	
1 Atlantic	90%	4%	1%	5%	100%
2 Quebec	63%	1%	1%	36%	100%
3 ON + NU	85%	1%	3%	11%	100%
4 Prairies + NT	51%	0%	3%	46%	100%
5 BC + YK	28%	0%	23%	49%	100%
Canada	64%	1%	30%	5%	100%

Then, we moved further down to a finer granularity level in the hierarchy. At the **provincial level**, we observed the following that are not covered by (or different from) the aforementioned observations:

- Among the four Atlantic Provinces, PEI has a lower infection rate (of 0.2% as shown in Fig. 4 and its zoom-in view in Fig. 5), no death due to COVID-19 (as shown in Fig. 6), and no case with beta or gamma variants. No delta variants were observed in NB. See Fig. 7.

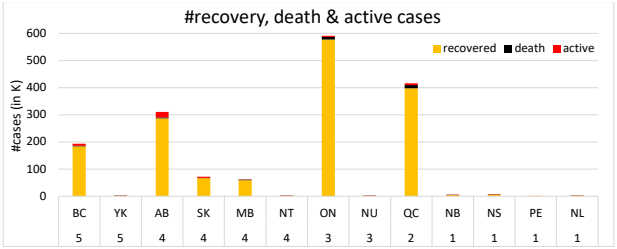

Fig. 4. Numbers of recovered, deceased and active cases in 13 provinces & territories within the 5 national regions

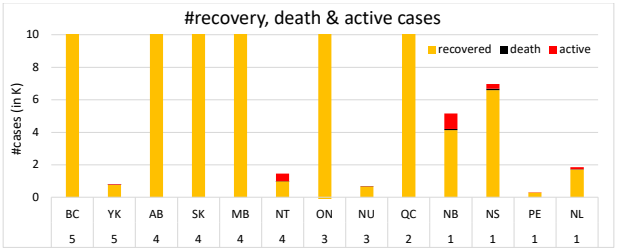

Fig. 5. A zoom-in view on numbers of recovered, deceased and active cases

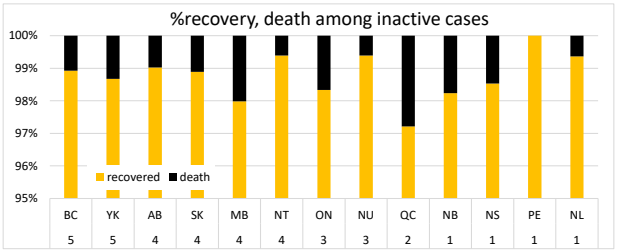

Fig. 6. Infection and death rates for 13 provinces & territories within the 5 national regions

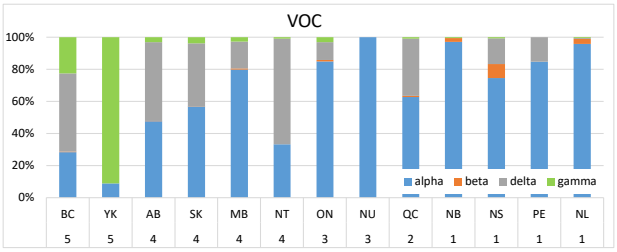

Fig. 7. VOC for 13 provinces & territories within the 5 national regions

- Between Ontario and Nunavut, the latter has a lower infection rate (of 1.7%) and a lower case fatality rate (of 0.6%). No beta, gamma or delta variants were observed there. This may explain why its full vaccination rate was

lower (with 75.2% those who aged 12^{+} received two doses).

- Among the three Prairie Provinces and NT, both Manitoba and NT have lower infection rates (of 4.5% and 3.2% respectively) than the regional rate of 6.3%, but Alberta has a higher infection rate (of 7.0%). Despite its low infection rate, Manitoba has a higher case fatality rate (of 2.0% when compared with the regional rate of 1.1%). In terms of VOC, although no beta variants were observed in NT, there were more delta variants than alpha variants there. In contrast, Manitoba has significantly more alpha variants (80%) than delta variants (17%). Some people in Alberta and Saskatchewan received their third dose. See Fig. 8.

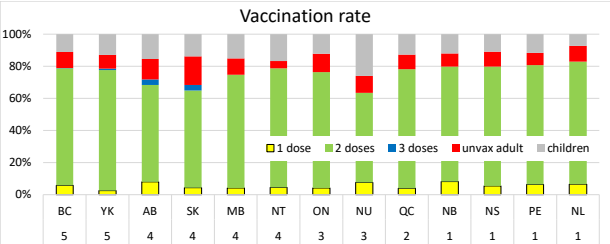

Fig. 8. Vaccination rate for 13 provinces & territories within the 5 national regions

- Between BC and Yukon, the latter has a lower infection rate (of 1.9%). No beta or delta variants were observed there. This is similar to the observations made from the comparison between Ontario and Nunavut.

C. *Comparisons with Related Works on Functionalities*

After we demonstrated the functionalities and practicality of our big data intelligence and computing solution in supporting health analytics of real-life COVID-19 data, we then compared the functionalities of our solution with those of the related works:

- Many of the related works reported mostly the numbers of COVID-19 cases and deaths, but did not provide privacy-preserving details and epidemiological characteristics of COVID-19 cases. In contrast, our solution provides these details and characteristics (e.g., transmission methods, hospital status).

- Some related works provided overall data distribution of COVID-19 cases, but they confined to showing the distribution of single dimensions/attributes. In contrast, our solution provides multi-dimensional information like relationships among attributes (e.g., relationships among transmission methods, hospital status, and recovery status).

To recap, when compared with related works, our solution provides additional functionalities beyond many related works. Specifically, it provides information (e.g., characteristics of COVID-19 cases) beyond just the numbers of cases and deaths, and relationships among these characteristics.

V. Conclusions

In this paper, we presented a big data intellgient solution for health analytics on big COVID-19 epidemological data. Our solution builds spatial hierarchy to capture data at multiple granularity levels in this hierarchy. In addition, knowing that population may not evenly distributed among different geographical areas, our solution also integrating population data with epidemological data. It then aggregates attribute values from a finer granularity level to a coarser granularity level of the hierarchy, and discovers in a top-down fashion. To minimize redundancy, it returns to users (a) patterns discovered from the top level and (b) only exceptional patterns that are not covered by parent or ancestor levels. By doing so, the numbers of returned patterns are comperhendible to users becauses the users can obtain a summary on patterns at the top level and essential patterns at lower (i.e., finer granularity) levels. Evaluation on real-life COVID-19 cases from Canada demonstrated the practicality of our soluton for health analytics of COVID-19 epidemological data—especially, in providing rich knowledge about characteristics of COVID-19 cases. The discovered knowledge helps users (e.g., researchers, epidemiologists, policy makers, civilian) to get a better understanding of the disease, and thus take an active role in fighting, controlling, and/or combating the disease. As *ongoing and future work*, we explore ways to incorporate temporal data into our current big data intelligence and computing solution to analyze and visualize the resulting knowledge. We also transfer knowledge learned from the current work to health analytics of other diseases and/or big data analytics in other real-life applications.

Acknowledgment

This work is partially supported by NSERC (Canada) and University of Manitoba.

References

[1] A. Kobusinska, et al., "Emerging trends, issues and challenges in Internet of Things, big data and cloud computing," FGCS 87, 2018, pp. 416-419.

[2] A.M. Olawoyin, et al., "Preserving privacy of temporal big data," IEEE BigData 2020, pp. 4042-4051.

[3] M. Yamamoto, et al., "Motion capture system towards big data collection of craftsmanship," IEEE DataCom 2019, pp. 17-24.

[4] A. Alsaig, et al., "Foundational issues on big data science and engineering," IEEE DASC-PICom-DataCom-CyberSciTech 2018, pp. 995-1000.

[5] C.K. Leung, et al., "A data science model for big data analytics of frequent patterns," IEEE DASC-PICom-DataCom-CyberSciTech 2016, pp. 866-873.

[6] C.K. Leung, et al., "A visual data science solution for visualization and visual analytics of big sequential data," IV 2021, pp. 224-229.

[7] P. Braun, et al., "Game data mining: clustering and visualization of online game data in cyber-physical worlds," Procedia Computer Science 112, pp. 2259-2268.

[8] J. Kim, et al., "KNN-SC: novel spectral clustering algorithm using k-nearest neighbors," IEEE Access 9, 2021, pp. 152616-152627.

[9] C.K. Leung, et al., "Effective classification of ground transportation modes for urban data mining in smart cities," DaWaK 2018, pp. 83-97.

[10] M.T. Alam, et al., "Mining frequent patterns from hypergraph databases," PAKDD 2021, Part II, pp. 3-15.

[11] M.E.S. Chowdhury, et al., "A new approach for mining correlated frequent subgraphs," ACM TMIS 13(1), 2021, pp. 9:1-9:28.

[12] L.V. S. Lakshmanan, et al., "The segment support map: scalable mining of frequent itemsets," ACM SIGKDD Explorations 2(2), 2000, pp. 21-27

[13] C.K. Leung, et al., “Fast algorithms for frequent itemset mining from uncertain data,” IEEE ICDM 2014, pp. 893-898.

[14] J. Liu, et al., “Efficient mining of extraordinary patterns by pruning and predicting,” ESWA 125, 2019, pp. 55-68.

[15] R.K. MacKinnon, et al., “DISC: efficient uncertain frequent pattern mining with tightened upper bounds,” IEEE ICDM Workshops 2014, pp. 1038-1045.

[16] S.Z. Ishita, et al., “New approaches for mining regular high utility sequential patterns,” Applied Intelligence, 2021. DOI:10.1007/s10489-021-02536-7

[17] K.K. Roy, et al., “Mining weighted sequential patterns in incremental uncertain databases,” Information Sciences 582, 2022, pp. 865-896.

[18] J.J. Cameron, et al., “Stream mining of frequent sets with limited memory,” ACM SAC 2013, pp. 173-175.

[19] C.K. Leung, F. Jiang, “Frequent pattern mining from time-fading streams of uncertain data,” DaWaK 2011, pp. 252-264.

[20] D.L.X. Fung, et al., “Self-supervised deep learning model for COVID-19 lung CT image segmentation highlighting putative causal relationship among age, underlying disease and COVID-19,” BMC Journal of Translational Medicine 19, 2021, pp. 318:1-318:18.

[21] C.K. Leung, et al., “Health analytics on COVID-19 data with few-shot learning,” DaWaK 2021, pp. 67-80.

[22] M. Wang, et al., “Leveraging reinforcement learning to allocate bandwidth in the agent-based resource management system,” IEEE DataCom 2019, pp. 70-76.

[23] F. Jiang, C.K. Leung, “A data analytic algorithm for managing, querying, and processing uncertain big data in cloud environments,” Algorithms 8(4), 2015, pp. 1175-1194.

[24] C.K. Leung, Y. Hayduk, “Mining frequent patterns from uncertain data with MapReduce for big data analytics,” DASFAA 2013, Part I, pp. 440-455.

[25] C.K. Leung, F. Jiang, “Big data analytics of social networks for the discovery of "following" patterns,” DaWaK 2015, pp. 123-135.

[26] C. Wang, C. Lee, “Data analysis of portfolio optimization using artificial neural network in China's stock market,” IEEE DataCom 2019, pp. 33-38

[27] H. Khaled, et al., “Parallel study of 3-D oil reservoir data visualization tool using hybrid distributed/shared-memory models,” IEEE DASC-PICom-DataCom-CyberSciTech 2018, pp. 1016-1021.

[28] C.K. Leung, C.L. Carmichael, “FpVAT: A visual analytic tool for supporting frequent pattern mining,” ACM SIGKDD Explorations 11(2), 2009, pp. 39-48.

[29] C.K. Leung, et al., “PyramidViz: Visual analytics and big data visualization of frequent patterns,” IEEE DASC-PICom-DataCom-CyberSciTech 2016, pp. 913-916.

[30] M. Prince, F. Lin, “hunting algorithm visualization and performance evaluation through BDI agent simulation,” IEEE DASC-PICom-DataCom-CyberSciTech 2018, pp. 262-269.

[31] F. Jiang, et al., “Big social network mining for "following" patterns,” C3S2E 2015, pp. 28-37.

[32] C.K. Leung, et al., “Interactive discovery of influential friends from social networks,” Social Network Analysis and Mining 4(1), 2014, pp. 154:1-154:13.

[33] C.K. Leung, et al., “Parallel social network mining for interesting 'following' patterns,” CCPE 28(15), 2016, pp. 3994-4012.

[34] S. Li, et al., “Predicting performance of social media postings using data mining methods,” IEEE DataCom 2019, pp. 54-63.

[35] P. Xu, et al., “TF-RNN: a method of rumor detection on social media,” IEEE DataCom 2019, pp. 47-53.

[36] C.K. Leung, “Mathematical model for propagation of influence in a social network,” Encyclopedia of Social Network Analysis and Mining, 2e, 2018, pp. 1261-1269.

[37] Y. Chen et al., “A data science solution for supporting social and economic analysis,” IEEE COMPSAC 2021, pp. 1689-1694.

[38] C.M. Choy, et al., “Natural sciences meet social sciences: census data analytics for detecting home language shifts,” IMCOM 2021, pp.520-527.

[39] A.A. Audu, et al., “An intelligent predictive analytics system for transportation analytics on open data towards the development of a smart city,” CISIS 2019, pp. 224-236.

[40] C.C.J. Hryhoruk, et al., “Smart city transportation data analytics with conceptual models and knowledge graphs,” IEEE SmartWorld 2021, pp. 455-462.

[41] M.D. Jackson, et al., “A Bayesian framework for supporting predictive analytics over big transportation data,” IEEE COMPSAC 2021, pp. 332-337.

[42] C.K. Leung, et al. “Urban analytics of big transportation data for supporting smart cities,” DaWaK 2019, pp. 24-33.

[43] C. Sueyoshi, et al., “An analysis of the number of passengers collected with a practical management support system for regional public transportation service,” IEEE DataCom 2019, pp. 258-261.

[44] S. Ahn, et al., “A fuzzy logic based machine learning tool for supporting big data business analytics in complex artificial intelligence environments,” FUZZ-IEEE 2019, pp. 1259-1264.

[45] C.K. Leung, et al., “A digital health system for disease analytics,” IEEE ICDH 2021, pp. 70-79.

[46] C.K. Leung, et al., “Big data science on COVID-19 data,” IEEE BigDataSE 2020, pp. 14-21.

[47] J. Souza, et al., “An innovative big data predictive analytics framework over hybrid big data sources with an application for disease analytics,” AINA 2020, pp. 669-680.

[48] W. Kuo, J. He, “Guest editorial: crisis management - from nuclear accidents to outbreaks of COVID-19 and infectious diseases,” IEEE Trans. Reliab. 69(3), 2020, pp. 846-850.

[49] A.A. Ardakani, et al., “Application of deep learning technique to manage COVID-19 in routine clinical practice using CT images: results of 10 convolutional neural networks,” Comp. Bio. Med. 121, 2020, pp. 103795:1-103795:9.

[50] A.K. Arshadi, et al., “Artificial intelligence for COVID-19 drug discovery and vaccine development. Frontiers Artif. Intell. 3, 2020, 65:1-65:13.

[51] Q. Liu, et al., “A two-dimensional sparse matrix profile DenseNet for COVID-19 diagnosis using chest CT images,” IEEE Access 8, 2020, pp. 213718-213728.

[52] Y. Chen, et al., “Temporal data analytics on COVID-19 data with ubiquitous computing,” IEEE ISPA-BDCloud-SocialCom-SustainCom 2020, pp. 958-965.

[53] C.K. Leung, et al., “Revealing COVID-19 data by data mining and visualization,” INCoS 2021, pp. 70-83.

[54] D. Deng, et al., “Spatial-temporal data science of COVID-19 data,” IEEE BigDataSE 2021.

[55] S. Shang, et al., “Spatial data science of COVID-19 data,” IEEE HPCC-SmartCity-DSS 2020, pp. 1370-1375.

[56] “Preliminary dataset on confirmed cases of COVID-19,” Public Health Agency of Canada, 2020-2021. DOI:10.25318/132600032020001-eng

[57] “Population estimates, quarterly,” Table 17-10-0009-01, Statistics Canada. DOI:10.25318/1710000901-eng