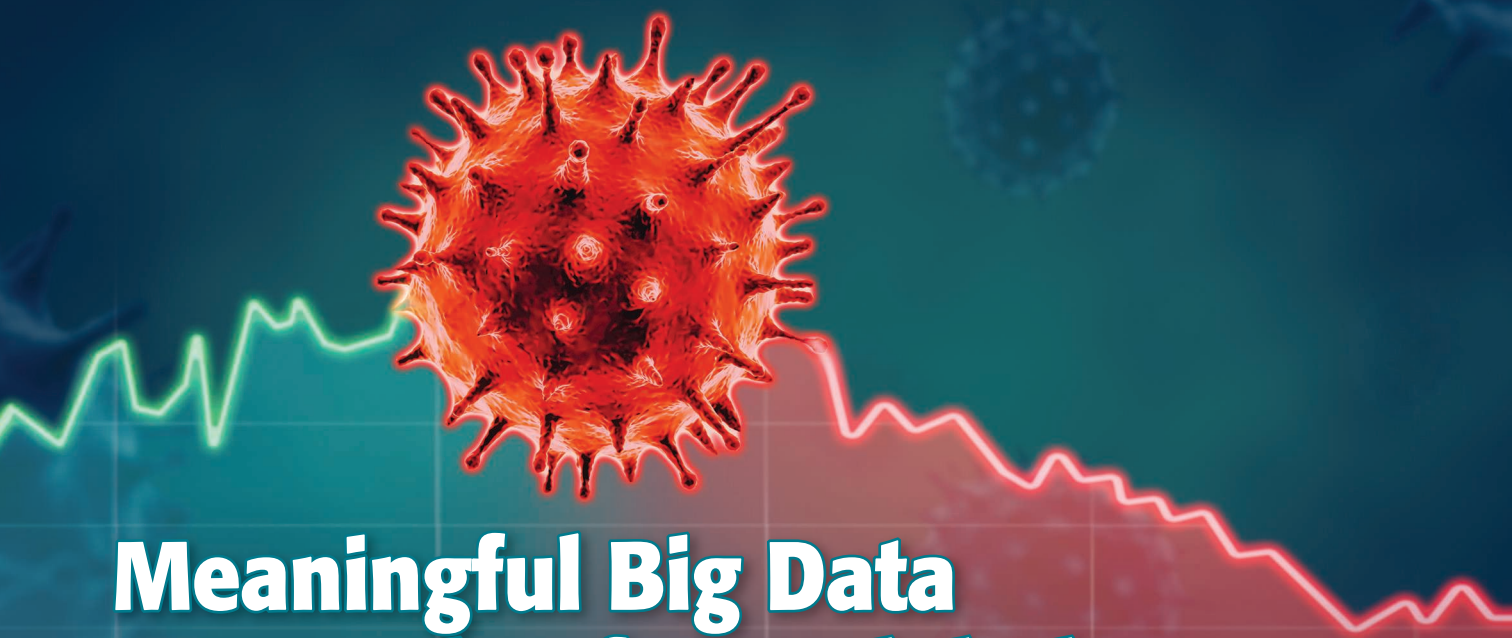

Meaningful Big Data Integration for a Global COVID-19 Strategy

Joao Pita Costa, Marko Grobelnik, Flavio Fuart, and Luka Stopar
Quintelligence & Jozef Stefan Institute, SLOVENIA

Gorka Epelde
Vicomtech & Biodonostia, SPAIN

Scott Fischaber
Analytics Engines, UK

Piotr Poliwoda
IBM, IRELAND

Debbie Rankin, Jonathan Wallace, Michaela Black, Raymond Bond, and Maurice Mulvenna
Ulster University, UK

Dale Weston
Public Health England, UK

Paul Carlin
Open University, UK

Roberto Bilbao
BIOEF, SPAIN

Gorana Nikolic, Xi Shi, and Bart De Moor
KU Leuven, BELGIUM

Minna Pikkarainen and Jarmo Pääkkönen
University of Oulu, FINLAND

Anthony Staines, Regina Connolly, and Paul Davis
Dublin City University, IRELAND

***Abstract*—**With the rapid spread of the COVID-19 pandemic, the novel Meaningful Integration of Data Analytics and Services (MIDAS) platform quickly demonstrates its value, relevance and transferability to this new global crisis. The MIDAS platform enables the connection of a large number of isolated heterogeneous data sources, and combines rich datasets including open and social data, ingesting and preparing these for the application of analytics, monitoring and research tools. These platforms will assist public health authorities in: (i) better understanding the disease and its impact; (ii) monitoring the different aspects of the evolution of the pandemic across a diverse range of groups; (iii) contributing to improved resilience against the impacts of this global crisis; and (iv) enhancing preparedness for future public health emergencies. The model of governance and ethical review, incorporated and defined within MIDAS, also addresses the complex privacy and ethical issues that the developing pandemic has highlighted, allowing oversight and scrutiny of more and richer data sources by users of the system.

Introduction

The COVID-19 outbreak was declared a Public Health Emergency of International Concern by the World Health Organization (WHO) on 30 January 2020 [35]. With its rapid expansion, health stakeholders are keen to find technologies to monitor and combat the spread and impact of the disease. Along with this, the world has seen the multiplication of surveillance efforts to monitor the epidemic by official global health agencies such as the WHO and the European Centre for Disease Control (ECDC). Businesses

Digital Object Identifier 10.1109/MCI.2020.3019898
Date of current version: 14 October 2020

Corresponding Author: Joao Pita Costa (joao.pitacosta@ijs.si).

The MIDAS intelligent system unleashes the potential of a range of analytics to explore the confidential and sensitive data owned by health authorities within a safe and secure environment.

and research institutions have rapidly refocused their monitoring platforms to assess the impact of this common threat. Examples include the WHO COVID-19 Dashboard by ArcGIS [2], the Coronaviruswatch platform by the UNESCO Research Centre for Artificial Intelligence (IRCAI) [34] and RavenPack's Coronavirus News Monitor [29], alongside many other global and local initiatives. Such platforms enable patterns and trends in disease behavior monitored throughout the population. They can also inform public health measures to reduce transmission and allow the impact of these to be assessed.

The novel MIDAS public health platform [21], presented in this paper and shown in Figure 1, goes a step beyond existing platforms, particularly in responding to the coronavirus pandemic, by providing its users in public health authorities with insightful information from a combination of sources including world news, social media and published science, alongside local public health data from the health institution itself and other relevant data sources. The MIDAS platform was co-created with academia, industry, and crucially, health professionals, policy-makers, public health authorities and citizens, to align innovative technology with concrete public health priorities and workflows [4]. It was developed to connect typically heterogeneous, isolated health data, and integrate it with additional social data sources, to enable the application of advanced data analytics techniques and visual analytics tools to support policy decision-making in public health institutes across Europe [7]. Further details, code repositories and demonstrators are openly available at www.midasproject.eu.

The MIDAS intelligent system unleashes the potential of a range of analytics to explore the confidential and sensitive data owned by health authorities within a safe and secure environment. The system achieves this through data visualization widgets, encapsulating the analytics processes, within user-customizable dashboards that summarize each user's selected output priorities, such as monitoring mental health or child obesity across the population within the COVID-19 confinement restrictions. The main contribution of this paper is the description of the MIDAS platform resources, how these were rapidly refocused to address COVID-19—related public health priorities, and how they can help researchers and public health policy-makers achieve a deeper and more complete understanding of the SARS-COV-2 disease.

The MIDAS integrated platform consists of software developed for data ingestion, processing, analytics and visualization. These services are deployed locally within health authority sites. The platform offers a bespoke co-created collection of software and services enabling secure access to and exploration of sensitive health data as well as open and social data brought in from external applications, via the MIDAS dashboard. A common user authentication service allows single sign-on across the services. Data downloads are prohibited to mitigate privacy concerns and risk. If a MIDAS user, such as a policy-maker, requires access to the raw and/or prepared data, they must make a data access request to the data gatekeepers to

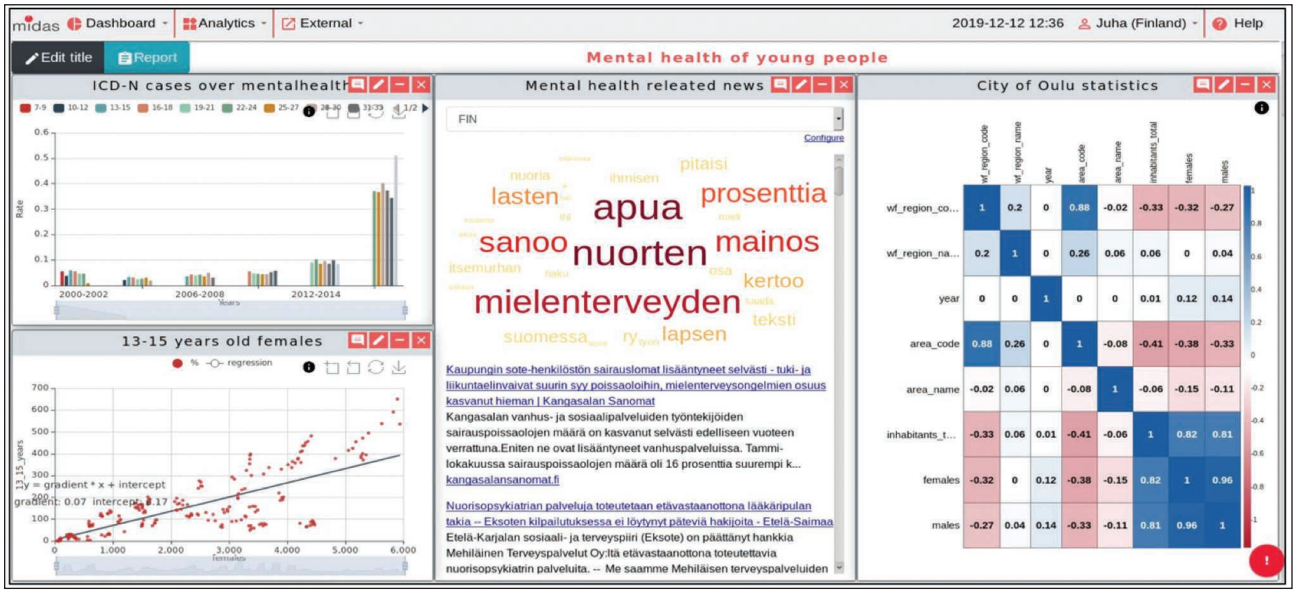

FIGURE 1 The MIDAS platform dashboard refocused to explore COVID-19 over proprietary data, worldwide news, social media campaigns and biomedical research.

secure a data access agreement (DAA) in compliance with relevant ethics and privacy legislation.

The core data platform is based on Apache HDFS, Apache Spark, and Apache Hive. In the system architecture diagram in Figure 2, the light grey boxes indicate user-facing web applications, whilst the dark grey boxes represent services and applications required for the platform that are not accessible to the end user. The MIDAS Platform runs across the health authority network (in blue) as well as externally (in orange). Data can be pulled into the platform from private or public data sources. The MIDAS platform has been successfully tested, validated [6] and evaluated by four pilot sites in public health institutes throughout Europe (Finland, Northern Ireland, Ireland and the Basque Region) realizing success across different priority policy areas. Each of these pilot study priorities is closely related to COVID-19, either in terms of the associated risk (e.g. diabetes, obesity and the ageing population), the impact of confinement restrictions (e.g. measures of mental health and childhood obesity), to a wide range of accessible social and spatial variables, such as measures of deprivation, social isolation, and access to and use of healthcare services [3]. From the outset MIDAS embedded a user-centered, co-creative, agile approach to its design and development, engaging with a wide-range of stakeholders, having the needs of each health policy pilot site driving the development in relation to both data analytics and visualizations for their individually selected health policy focus [5].

Making Data Meaningful Within a Local Context

The ability to enhance the usefulness of data located within health policy sites by integrating it with other global data sources provides significant potential value to the user. MIDAS provides tools such as a cross-filter analytics dashboard, an easy to use and understand interactive map visualization that updates its content automatically when the user selects different countries on the displayed map (as shown in Figure 3). This can be a useful tool for the initial exploration of many types of data within a specific topic, e.g. COVID-19. The form of visualizations is not uniform and can be customized based on the data type and research question (i.e. they can include line charts, bar charts, maps, tables, and other plots). The categorical variables used in the cross-filter are predefined. Users can select subgroups for analysis and the graphs will automatically update with data from the selected subgroups. An example of the cross-filter tool in the MIDAS platform is shown in Figure 3, created for the Irish pilot. The categorical variables are Gender, Region, and Age Group, and users can select subgroups either by clicking the buttons on the panel, selecting the region on the map, or the subgroups in the line chart or bar chart.

The application of the cross-filter tool can present results more intuitively on the map and can make cross-group comparisons easier through interaction with the user. It is not necessary to have the same components as the example of the Irish pilot shown in Figure 3, however all of the components can be replaced according to the categorical variables and data types. The flexibility of the tools developed for the MIDAS platform allows the user to apply the most suitable forms of visualization to their data.

To help users follow the results, emergency ICD-10 WHO codes U07.1 and U07.02 from [36] have been assigned to diagnose an identified presence of the virus. For example, the

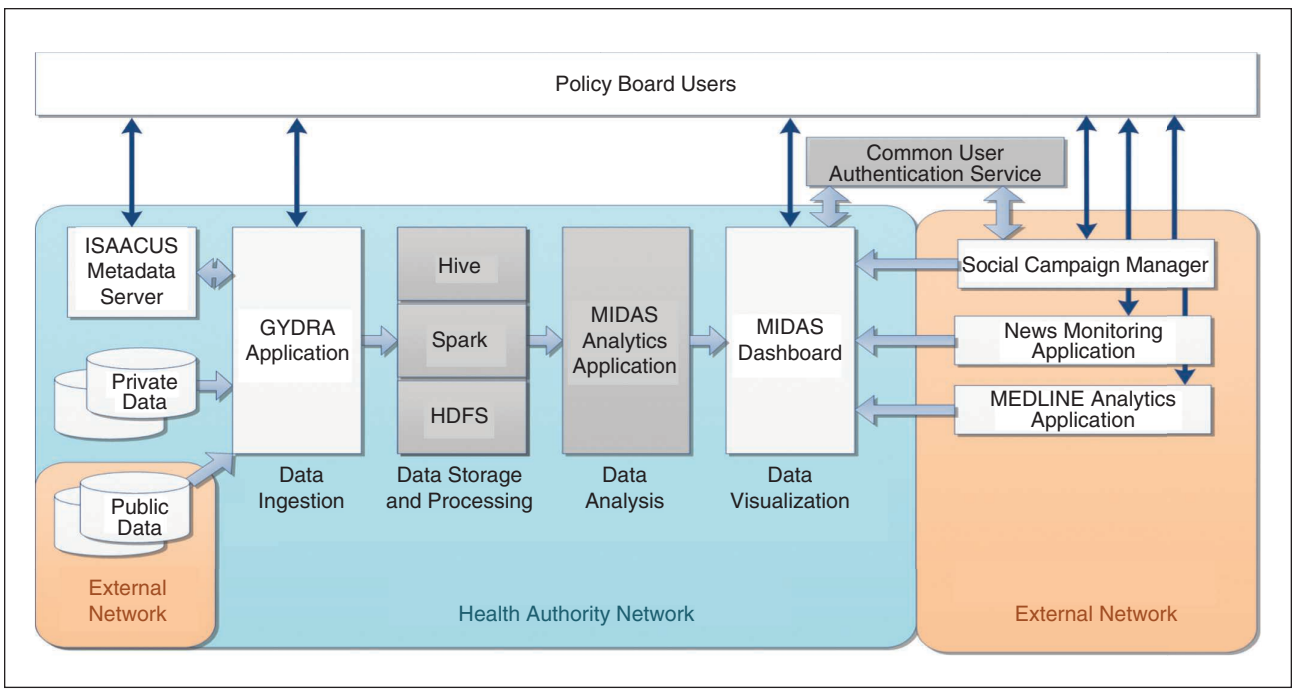

FIGURE 2 MIDAS platform architecture overview, designed to host Big Data to improve decision-making in public health.

Irish cross-filter tool, shown in Figure 3, allows the monitoring of diabetes cohorts and their Around the Clock prescription use by individual counties. With the MIDAS cross-filter tool it is possible to repurpose it to closely follow the COVID-19 codes and obtain a more accurate epidemiological overview over the regional county map.

Using the exploratory data analytics dashboards in the MIDAS platform, we have evidenced the known link [18] between diabetes and pneumonia outcomes. This insight from the available data is also highly useful in the response to COVID-19, considering that diabetes is a risk factor for the progression and poorer prognosis for COVID-19 patients [13], [16]. Using the MIDAS platform to understand the cohort of the population with diabetes in Ireland enables more targeted responses to the COVID-19 pandemic, for example, targeting local areas, which have high prevalence or higher historical hospitalizations with comorbidities, with specific messaging. This permits a more effective and rapid service delivery and may allow the identification of potential COVID-19 hot spots in advance.

The MIDAS pilot in the Basque Region, focusing on the topic of childhood obesity, is useful and important given the known associations of obesity in children with the emergence of comorbidities (e.g. diabetes and hypertension). Moreover, the Basque public health authorities are interested in monitoring the effect of the COVID-19 pandemic on childhood obesity using the MIDAS tools. Behavioral changes during lockdown in children and adolescents with obesity participating in a longitudinal observational study in Italy have been published recently [24]. No changes in vegetable intake were reported during lockdown, whilst in contrast, potato chip, red meat, and sugary drink intake increased significantly. Time spent in sports activities decreased, sleep-time increased and screen time increased. Taking into account the severity of the COVID-19 cases identified in people diagnosed with obesity, the MIDAS platform could be utilized to enhance our understanding of the adverse collateral effects and lasting impact of the COVID-19 pandemic lockdown on the adiposity level of adolescents. This includes the evolution of childhood obesity during the pandemic, and the evolution and adoption of policies based on the monitoring and visualization of a rapidly changing context.

The topic of mental health was studied in the context of the MIDAS pilot in Finland. This pilot ran a social media campaign in 2019 through the Twitter chatbot capabilities within the MIDAS platform (described in a later section of this paper). The existing data, ingested into the MIDAS platform previously and located within the policy sites, can be useful in this new COVID-19 framework, helping public health authorities to gain a better understanding of the health scenario over this difficult to assess topic, complemented by the news published around it. The additional ingestion of online participatory surveillance systems can contribute to a better understanding of the impact of the pandemic on mental health.

The COVID-19 situation has increased global concerns in relation to the mental wellbeing of children and young people. In Finland, the MIDAS pilot focused on the prevention of mental health problems in young people. As an example, within this context, the MIDAS platform could be used as a tool to support decision-making related to the organization of mental healthcare services and resources for young people. The COVID-19 pandemic has created a situation whereby long isolation periods at home and difficult economic situations are commonplace. This has raised concern amongst decision-makers in relation to child abuse and domestic

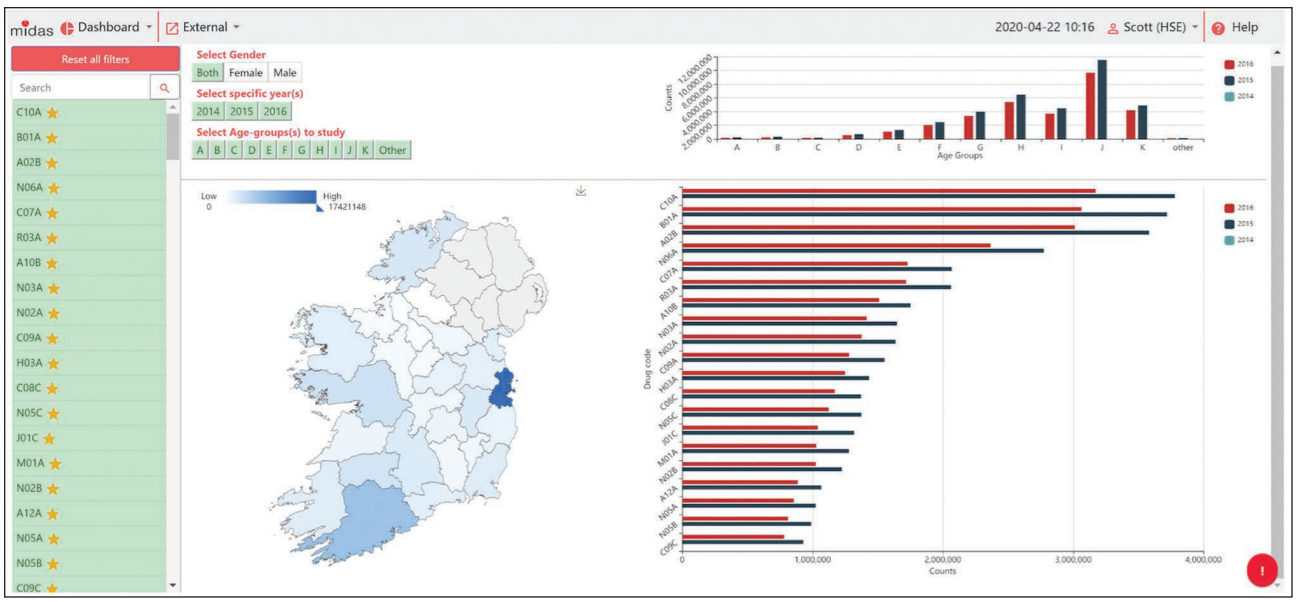

FIGURE 3 The cross-filtering at the MIDAS platform and its usage by the public health authorities in Ireland.

violence. This is an example of a complex and multi-level issue where a range of heterogeneous data is required at a policy-making level to enable better analysis and visualization of the situation. The hindering factor for the real use of the MIDAS platform in mental health cases in Finland is the current legislation often restricting the combined usage of the anonymized social and healthcare data collected on individuals in policy level decision making.

… the MIDAS platform could be utilized to enhance our understanding of the adverse collateral effects and lasting impact of the COVID-19 pandemic lockdown.

As an example of how the MIDAS platform could be used in the COVID-19 context, the lockdown effects on childhood obesity and mental health could be analyzed by updating the available longitudinal clinical with data covering the lockdown period, reusing existing data preparation techniques, or modifying them to allow the inclusion of new countries and health systems and reusing existing data analytics and visualization tools (on the period of interest). In the case of childhood obesity, these could focus on body mass index (BMI) and epidemiological analysis of new diabetes cases, covering person (gender, age-group), time (year) and location dimension (trust level and primary care unit level). Furthermore, the tool could be extended by plugging and ingesting new data sources such as physical activity captured by wearables, nutrition-related periodic questionnaires, or by alternatively adding grocery shopping aggregated data that, once integrated to comparable aggregation levels and described in metadata, could be analyzed side-by-side with current visualizations or further analyzed using statistical or machine learning technique.

Ingesting Useful Open Data Sources

The MIDAS platform includes heterogeneous datasets prepared and deployed in different pilot site locations, addressing their own specific challenges. These include city and government generated controlled datasets (i.e. health and social care data exports mainly at individual person level) and government open data (aggregated data) on air and water quality, national statistics (e.g. deprivation, education level or unemployment level per municipality), or city planning. These data sources are selected by each user to address various priority health policy questions across each pilot region.

The integration of different controlled public health data sources containing individual level data for the MIDAS pilot cases was completed by the data owners prior to loading these into the MIDAS platform, while the linking identifiers and datasets provided to the consortium were agreed to be provided on an anonymous basis. A different instance of the MIDAS platform was securely hosted at each policy site, to ensure the data owner retained control over the data being loaded. These heterogeneous datasets have been used to provide combined solutions in different sites, combining data at aggregated and individual level, by providing mappings at agreed location aggregation levels. Applying this process to the context of COVID-19, it is easy to map and analyze the relationship of clinical variables and open data indicators (i.e. COVID-19-related indicators such as number of cases, intensive care units or beds, and recovered people that have been published through different organizational open data agencies) at an aggregated location level (e.g. primary care unit or trust). The COVID-19 pandemic motivated the availability of diverse open datasets and indicators [1], presenting new opportunities (new sources for monitoring, modelling and forecasting the pandemic) and challenges (the need to correctly pre-process and integrate the data sources made available).

In the MIDAS platform the GYDRA Big data preparation tool (renamed from its initial in-memory processing version TAQIH) [30] has been developed for the preparation, ingestion and loading of the selected datasets. GYDRA has two main aspects: (i) an easy to use and interactive web-based interface (mimicking traditional data quality assessment and improvement flow) allowing non-technical users to use it; and (ii) data synchronization functionality to allow data owners and policy-makers to iteratively prepare and automatically deploy the prepared data to the analytics platform (relying on Apache Hive technology, and a defined metadata approach for the platform).

Within GYDRA's data preparation user interface items are placed from left to right following the usual iterative pipeline in exploratory data analysis. The "*General Stats and Features"* GYDRA sections provide global and detailed views of the data content, distribution and quality. The *"Missing Values"* section deals with the completeness of data. The *"Correlations"* section presents the correlations amongst variables, to help identify possible redundancies amongst variables or incoherent data. Finally, the *"Outliers"* section identifies outliers for each variable. Based on the insights identified during this analysis, a transformation pipeline can be configured to drop features and observations, handle missing values and outliers, or define operations to create new features from existing features or to change specific values. Within the MIDAS project each dataset from each pilot site has required a different preparation recipe. However, common tasks for each dataset include checking the number of columns per row (to avoid issues with separators being present in the content), merging data exported in chunks, format changing (e.g. for date fields loading), recoding of categorical values (after defining integration mappings), dropping features with few occurrences, dropping some meaningless outlying occurrences and creating new tables for specific analysis.

The metadata generated by the data synchronization functionality introduced above describes the data organization after

multi-resource data is ingested via the GYDRA tool and deployed to the analytics platform. This enables the automatic generation of generic data analysis and visualizations, as well as the development of specialized applications and machine learning models exploiting the secondary use of platform loaded data.

Figure 4 depicts the data ingestion, preparation and synchronization process of the GYDRA tool. Using the established data ingestion, preparation and synchronization methodology, it is straightforward to ingest, map and load COVID-19 related controlled and open datasets to the MIDAS platform. Additionally, the GYDRA tool relies on the interactive definition of dataset transformation pipelines which, once defined and refined, can be used to dynamically process and load partial and complete dataset updates. In this sense, the feature enables the ingestion of more dynamic data sources (dynamic in contrast to data export dumps provided at certain time periods by Health Care providers). In order to produce (re)usable COVID-19 data ingestions and mapping pipelines, analytics and visualizations within the MIDAS platform, it is necessary to work with the clinical and scientific community to achieve private and secure means to access data sources and agreed data models (as requested by [12]).

Moreover, the combination of the MIDAS developed GYDRA data preparation tool, alongside synthetic dataset generation strategies, can enable hospitals and healthcare providers, to: 1) refine and prepare their datasets (with the required metadata description), and; 2) share synthetically generated privacy-preserving datasets with the scientific community, that follow statistical patterns similar to the real data, and have proven to be reliable for training machine learning models [28]. These mechanisms would enable users to load a controlled dataset into the MIDAS platform and to develop in-house analytics, whilst simultaneously allowing the scientific community to develop AI models based on synthetic datasets that can later be fed back to the policy-makers through the MIDAS platform. This methodology provides a way to upscale and expedite the development of machine learning solutions through privacy preserving data sharing.

Extracting Insight from Published Biomedical Research

As the pandemic developed, MIDAS contributed to the many efforts to help biomedical researchers gain a better understanding of the disease (see examples in [14], [38] and [20]). To this end we have utilized the knowledge base MEDLINE [22] that serves the well-established biomedical search engine PubMed [23]. This open dataset stores structured information on more than 30 million records dating back to 1966. The comprehensive controlled vocabulary associated with MEDLINE—the MeSH Headings—delivers a functional system of indexing published biomedical science from journal articles and books. The MEDLINE articles are hand-annotated by humans with the established MeSH headings as health-related topics. These allow the user to explore a certain biomedical related topic (e.g. *"Biomarkers"* with the MeSH ID D015415), relying on curated information made available by the North American National Library of Medicine (NLM). The controlled vocabulary MeSH extends from 16 major health categories (covering topics such as anatomical terms, diseases, and drugs), each of which will be further

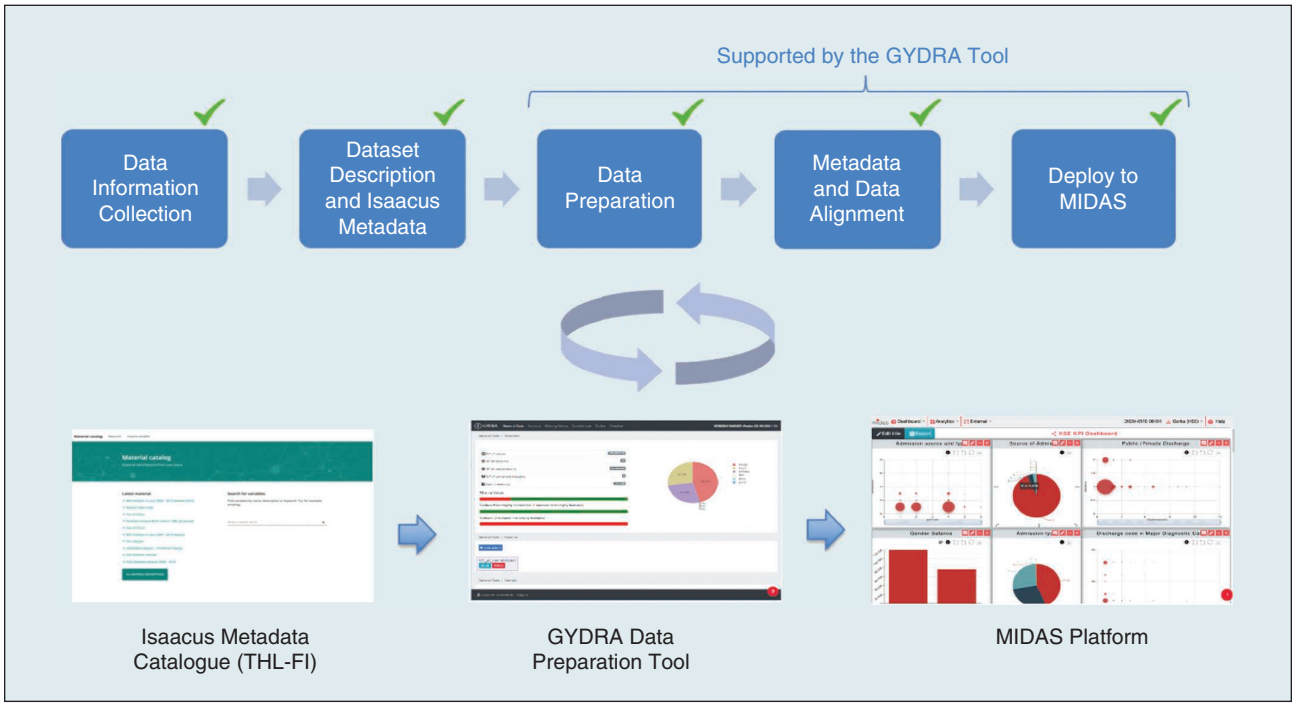

FIGURE 4 GYDRA tool-based data ingestion, preparation and synchronization with the MIDAS visualization and analytics platform.

distinguished from the most general to the most specific in up to 13 hierarchical depth levels. Although the recent introduction of the supplementary concept *"COVID-19"* on January 13, 2020, the MeSH heading "Coronavirus" was introduced in 1994 referring to the group of related viruses and prior known strains. These are located in the MeSH tree under the Coronaviridae family, introduced in 1999. MEDLINE includes 5976 research articles on coronavirus, relating to 4097 other health topics and 1706 substances. The articles that are hand-annotated with the MeSH class *"Coronavirus"* can help researchers better understand the new strain from the available scientific literature. With this in mind, the MIDAS platform offers an exploratory tool (see Figure 5) that allows the user to explore the published research through: (i) a query, based on keywords and operators, or an advanced query based on the syntax of the Lucene language; and (ii) a target pointer, which the user interacts by dragging it over the tag cloud to explore the results on the subtopics it relates to and to reprioritize the search results obtained. An example of such a precise syntax query (including two types of search categories—MeSH headings as health topics, or Chemicals)—*MeshHeadingList.desc:" Coronavirus" NOT ChemicalList.NameOfSubstance: "Viral Proteins"*—provides the user with, e.g. a subset of MEDLINE articles that are hand-annotated with the MeSH heading *"Coronavirus"* but are not labelled with the substance *"Viral Proteins."* The user can further explore the subset restricting it to labelled publications with the *"Biomarkers"* MeSH heading to explore new treatments in this specific context.

It is used to annotate scientific articles, news articles and reports relating to the COVID-19, allowing for the utilization of the MeSH headings as search topics …

The MIDAS platform also includes an exploratory dashboard that provides access to all MEDLINE records and enables users to explore these directly, and save samples based on queries. These are stored as JSON files in an elasticsearch based database, utilizing robust and well-established technology [11]. The external MIDAS MEDLINE explorer does not require the user to have expert technical knowledge and allows the average biomedical researcher to explore MEDLINE with further insight but little technical skills required. It also enables the user to rapidly build several data visualization modules (based on the Kibana open-source data visualization technology) that are easily configurable and based on templates, over the queried data (saved as a subset of records). It includes a variety of charts, tag clouds, heat maps and lists, as well as dashboards that integrate the created dynamic data visualization modules (see Figure 6). These enable MIDAS users to build and share dashboards that analyze published biomedical research on COVID-19 in relation to relevant topics for this study (similarly to diabetes, mental health etc.), and recover the biomedical articles relating to it. Moreover, this dashboard is served with a powerful API that allows the user to access and query the data from other systems. An extended instance of it enables easy ingestion of new articles, reports or news that can be annotated using a classifier, loaded, visualized and explored in the

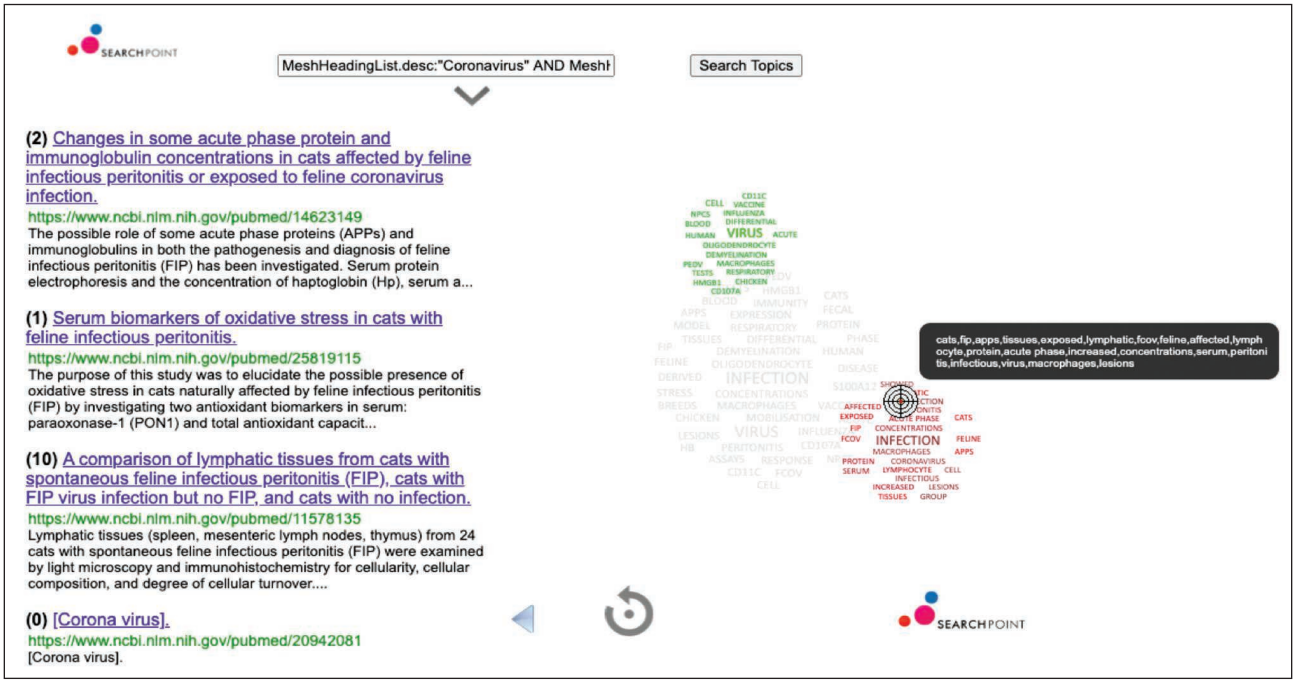

FIGURE 5 Exploring the biomarkers related to the Coronavirus in the published research using the MIDAS MEDLINE Explorer.

MEDLINE dashboard and explorer tools, and further mined through the available API.

One of the highly innovative technologies derived from the research carried out to build the MIDAS platform is the MeSH classifier [23]. It is an automated text classifier that has learned over human hand annotation of MeSH classes from more than 25 million biomedical articles (part of the corpus was used for evaluation) to perform the assignment of MeSH classes to any given snippet of text. It uses advanced text mining algorithms for this classification [19], and can classify any input text (including news, reports and health records) with this well-accepted health taxonomy. It is used to annotate scientific articles, news articles and reports relating to the COVID-19, allowing for the utilization of the MeSH headings as search topics over the corpus of these documents. It enhances the searchable information and allows for data visualization modules where the user can see the health topics (based on the most frequent MeSH classes) associated to the news over a query. This classifier was evaluated over: (i) scientific articles, part of the annotated MEDLINE dataset; and (ii) over news articles, hand annotated by some of the health experts in MIDAS on topics such as infectious diseases, diabetes, childhood obesity and mental health. The evaluation of the classifier resulted in an F1 measure of 0.43 in the MeSH tree depth level three for the classification of scientific articles, whilst F1 measures range between 0.55 to 0.85 for news articles in specific health domains (including diabetes, mental health and infectious diseases). The details will be published in [27]. The MeSH classifier offers a web portal and an API to enable a diversity of usages and integrations in other solutions. The web portal (accessible through the MIDAS COVID-19 toolset at the web portal *www.midasproject.eu/covid-19/*) provides the positioning of the MeSH categories that were assigned to the input text snippet, their similarity percentage and the MeSH tree branches to which the class belongs. This classifier allows us to generate useful metadata (based on the MeSH categories assigned to news articles, new research articles or medical reports) enabling its usage in the MEDLINE explorer and dashboard described previously. This explorer is served with an API that allows access to the structured Coronavirus dataset, and that can be enriched with other reports and annotated with the MeSH classifier. This allows researchers to leverage the existing knowledge generated in the current research.

Worldwide News Monitoring

The MIDAS news monitoring dashboard is fed by Newsfeed technology, collecting and analyzing more than 100 thousand news articles daily in real-time through the Event Registry technology, offering the MIDAS user insightful data visualizations to explore health-related news [17]. Since January 2020, the MIDAS news engine collected more than 13 million news articles on coronavirus-related topics across more than 60 languages. These included over 120 thousand articles on COVID-19 and diabetes, more than eight thousand articles on COVID-19 and retirement homes and elderly care, over 18 thousand articles on COVID-19 and obesity, more than 116 thousand articles on COVID-19 and mental health, and approximately 191 thousand articles on COVID-19 related to nursing. The news explorer within MIDAS allows the user to explore the overall sentiment of the news and the categories associated with it. From the total amount of collected coronavirus articles, approximately 36% have a positive sentiment and 0.69% relate to patient education whilst 0.71% relate to testing facilities. Recently, IRCAI released a worldwide news monitoring dashboard dedicated to COVID-19 based on the same news engine [30]. This general purpose health news monitoring dashboard exhibits the news on the epidemic outbreak in real-time and allows the reader to explore the information provided by country. However, the user cannot customize the news feed except using preset filters. MIDAS improves the usability of that by providing a news stream (see the visualization module at the center of Figure 1) where the user can personalize the search query and even include blogs. It allows further exploration of COVID-19 related news specific to topics on the user's own health policy priorities, such as home care or childhood obesity. It includes a tag cloud to have a first grasp over the main topics under discussion. Alongside this useful tool, the MIDAS news dashboard [25] allows the user to further explore the news based on data visualization modules including related concepts, entities and categories, or even the sentiment of the news article selection. This analysis is particularly important to avoid bias in the health news search [26], and to explore the several dimensions of misinformation caused by the *infodemia* [37], in conjunction with the pandemic. The search engine uses Wikipedia terms, to ensure the multilingual potential of the dashboard. These include the following COVID-19 related terms: *"Coronavirus"* (on the Coronavirus virus family, available in 73 languages), *"Coronavirus disease 2019"* (corresponding to the specific COVID-19 sort, available in 136 languages), and *"2019–20 Coronavirus pandemic"* (that writes on the pandemic itself, available in 132 languages). Moreover, we can backtrack the news articles about Coronavirus in Italy, discussing the triage of passengers commuting from Wuhan, China, in the timeline exploration visualization module, to January 20th this year, at the beginning of the European epidemic. We can further access these articles' entities to identify main actors and related topics. We can also explore the sentiment over these news articles and, in some cases, their impact on social media through the number of times a particular news article was shared. All these features are offered over comprehensive and well documented APIs.

The usefulness of the MeSH classifier, described in the latter section, is extended in the MIDAS platform through its integration with the news dashboard. With this integration, the user can use the MeSH heading terms together with keywords in a query, when exploring a certain news topic, in a similar fashion to the usage of the well-accepted PubMed workflow, providing data visualization modules that include those classes. A meaningful example is the visualization module *"Article Categories"* where a MIDAS user can see the distribution of news

articles subsequent to the query throughout the related MeSH classes. Figure 6 shows that 6.64% of the news on Coronavirus talks about topics related to the Mesh heading *Organisms/Viruses/RNA Viruses*. These new capabilities enhance the monitoring of health news over structured information, allowing a MIDAS user to have an understanding of media coverage in closer conjunction to the biomedical research itself.

Campaigning Through Social Media

A Twitter chatbot campaign led by MIDAS helped assess the global efforts of people during the pandemic. The aim of the social media campaign was to check-in with the global Twitter community during the "One World: Together at Home" initiative led by Global Citizen, asking questions concerning their hopes, technology usage and feelings towards a more connected world. The Global Citizen initiative was an event that aimed to support the WHO's COVID-19 Solidarity Response Fund, which supports and equips healthcare workers around the world. Hundreds of thousands of pieces of protective medical equipment, and 1.5 million diagnostic kits were provided to countries around the world through this fund. As part of the efforts this global initiative brought together change makers from over 150 countries and helped to raise funds for the cause. However, the globally televised event itself was not a fundraising telethon. It focused on entertainment, messages of solidarity and showing support for healthcare workers [31], [33]. The social media campaign was connected to a Twitter account managed by the IBM Corporate Social Responsibility team (@IBMOrg). Using the Social Campaign Manager's ability to spawn chatbots at will, the team was able to create a user-friendly conversation-led questionnaire asking the public a series of questions. The campaign included a number of multiple choice questions ranging from simple yes, no, maybe answers, to questions requesting the selection of one or more items from a list of options, or the answer to open questions where the respondent was free to write their response in free-form

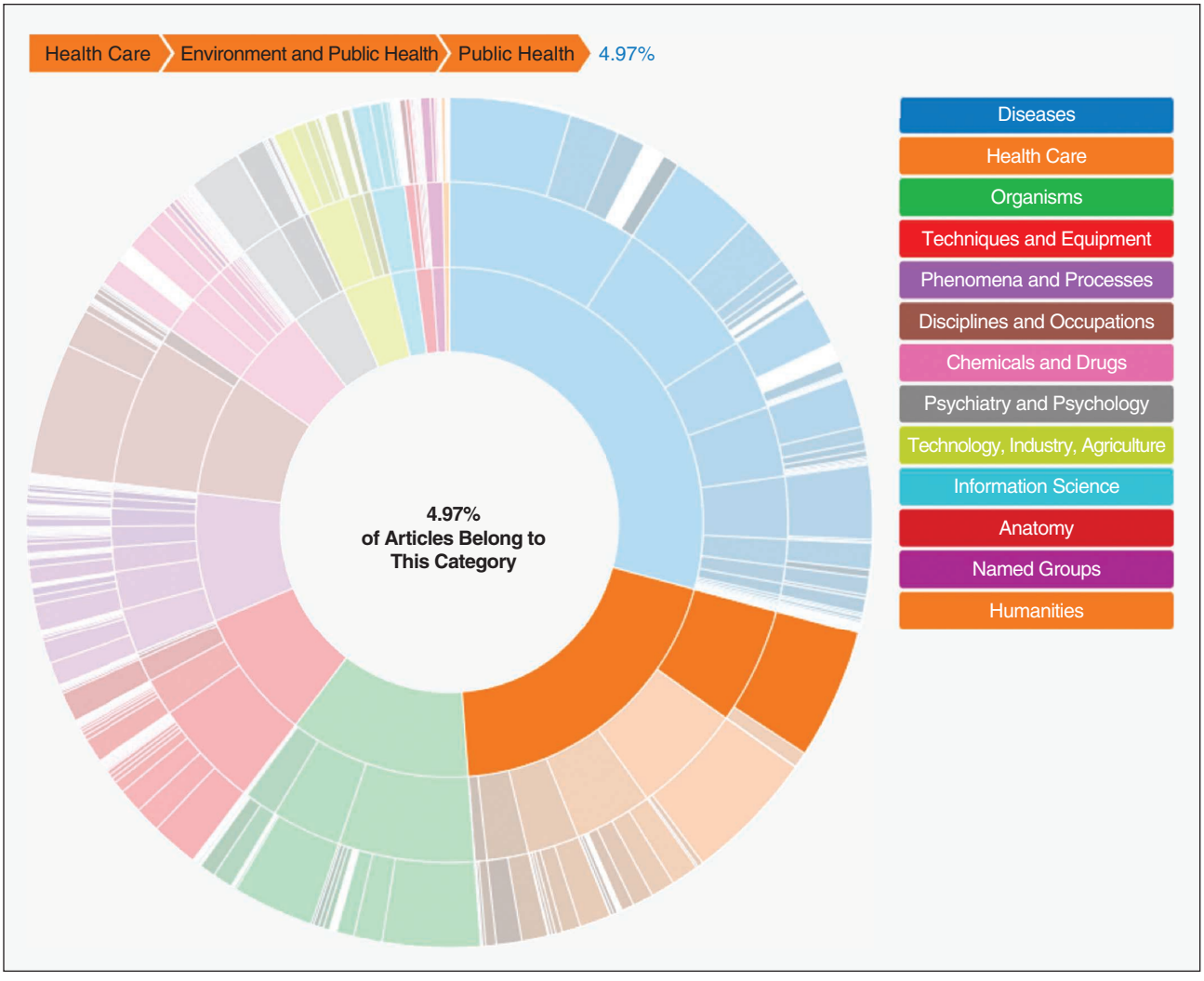

FIGURE 6 Analyzing the Coronavirus news articles through the percentage of health topics using the innovative integration between the MeSH classifier and the MIDAS news dashboard.

… a technological solution, such as the MIDAS platform, allows analysis of heterogeneous datasets, in an environment that allows relationships and policy to be explored.

text. The free form responses were analyzed using IBM's Watson Natural Language Understanding tool, which in turn provided the campaign owners with a general sense of the sentiment of the conversations they had with the chatbot, and emotions expressed in the responses. The aggregated view of these responses shows that the favorite stay at home activity of 38% of the public was the physical exercise, with video chatting at 25%, and hobbies at 20%. When asked how technology has shaped their life since the start of the COVID-19 pandemic a large majority (65%) of people said they rely on [technology] more than ever and a third of the respondents said they use it as much as before. Only 2% of the respondents said they use technology less than before.

Topics trending in the campaign responses were family and friends, health and fitness, resulting societal impacts of the disease, children and education. The scientific impact due to the global efforts worldwide was also mentioned. The system can identify the aggregated sentiment of free-form text responses given by all survey participants and provide the prediction of the mean probability of the emotions (within the categories: sadness, joy, fear, disgust or anger) in the free form responses using IBM's Watson NLU. Of all of the categories mentioned, only technology and computing were mentioned in a neutral context, whereas the majority of responses were given in a positive context which can be seen in green in the graph in Figure 6.

Ethics in the Time of COVID-19

"May you live in interesting times" is often quoted as an ancient Chinese curse, but dig deeper and this origin is erroneous. It is actually attributable to Joseph Chamberlain, a 19th century British politician. The parallels with the current COVID-19 crisis requires little imagination. The transitional world in which we currently live has unheard of restrictions of movements and freedoms normally available in democratic societies [15]. These restrictions are driven by modelling and the epidemiological evidence and, certainly to this point in May 2020, the public appears to have trusted the rationale and approach in large part [32]. Lockdown strategies are a matter of choosing short term loss over long term gain; these are the policy questions that are being dealt with and, as such, require the best evidence available. What is clear is that a technological solution, such as the MIDAS platform, allows analysis of heterogeneous datasets, in an environment that allows relationships and policy to be explored. A key output of the platform development was the realization that this environment should be apolitical, in the sense that policy should be based on science and the relationships of the data used in the system, robustly quality checked and analyzed, [5]. MIDAS therefore proposes a two-pronged approach to ethical and governance assurance: the public as partners and a system of robust ethical and scientific oversight from all parties involved in the MIDAS platform. Public engagement is core and needs to be meaningful. This requires a program of engagement, education and support for the public. Obviously this also requires nuance, resource and openness by science and government, as well as innovative techniques for engagement, such as the Chatbot discussed previously, and a platform such as 'engage' [9] used throughout the MIDAS platform development. In the time of SARS-CoV-2 (COVID-19) this may seem a luxury, but we need to plan now for future outbreaks, pandemics, or other public health emergencies. This public engagement and perhaps the use of opt in/opt out models of data use for public health is a discussion that needs to take place urgently. A measure of control is essential in managing public trust, expectation and compliance in the use of any system. MIDAS mitigates the risk by creating a system to manage this requirement: an Honest Broker Service model (HBS). This system creates an operational structure for review, scientific justification and oversight drawn from all interested parties, the public, government, academia and business. This is the ABCD model: Academia, Business, Client, and Direction. These parties set the bar in respect of the scientific/policy question at hand, allowing scrutiny of the hypothesis by parties not directly invested in specific work, within a framework that allows review for quality assurance and feasibility. This model creates a system that can be trusted, a regulatory framework much as the ones that exist for devices and pharmaceuticals that can allow science to drive decision making. Public messaging and education is integral to acceptance of the model, and of course the operational integrity of the system is dependent on user engagement through data use, a symbiotic relationship between the public and data investigators. A system that includes the public as contributor and gatekeeper, vouchsafed by independent review goes some way to safeguarding this trust.

Conclusion

A substantial part of technology adoption in public health and healthcare is the utility of the tools and the meaningfulness of their outcomes. As a result of being co-created with stakeholders [8], undergoing regular impact evaluations [10], and having usability formally evaluated by policy-makers [6], the MIDAS platform has proved its usefulness and has led to the development of components driven by stakeholder requests. Significant interest was established in the MIDAS platform in what it can offer to new regions, cities and organizations.

Moreover, the platform comprises a representative set of open, anonymized and synthetic data upon which the full range of available analytics and corresponding visualizations reside. This is valuable in an epidemic scenario, enriching the proprietary data of the public health authority, with existing

results in areas close to the disease (e.g. diabetes and old age), to the outcomes of related restriction measures (e.g. mental health and childhood obesity), and that can be refocused with low effort (e.g. child care to elderly care). The potential of (i) specific public health campaigns using social media; and (ii) worldwide news monitoring with a measure of impact in Facebook, can further help in understanding the spread of the disease and misinformation around it. In turn this will contribute to improvements in the public health campaigns that are an essential component for the success of disease control. Finally, the integration and utilization of open datasets, and the use of MEDLINE, in particular, greatly contribute to the understanding of the disease itself, when studying it side-by-side with the local data. A further ambition is to analyze and study how individuals' biological and psycho-emotional status with the actual data performs using adapted mental health and childhood obesity research questions for the COVID-19 pandemic. Results could influence the current pandemic response, alongside the development of health policy recommendations and preventive actions needed for prevention/control of future outbreaks or pandemics. The ethics and governance frameworks used in MIDAS, whilst operationally limited to the project and the HBS In Northern Ireland (with model development in the Basque region, and a similar model adopted within Finland), clearly articulate the demand and need for a system of oversight and review. Therefore, what is needed now is the adoption of the described model at scale, adequately resourced and built upon a funded and meaningful engagement piece. There will also need to be a shift to a position in which HBS becomes the Trusted system for review.

The MIDAS Open Source Foundation (OSF) will directly facilitate the long-term sustainability and growth of the MIDAS Platform and will provide an opportunity for health authorities, as well as regional, national and pan-European governments to embrace the platform to address strategic health policy development such as for COVID-19 or any future pandemic/global health crisis.

Acknowledgment

This project was funded by the European Union research fund 'Big Data Supporting Public Health Policies,' under GA No. 727721.

References

[1] T. Alamo et al., Covid-19: Open-data resources for monitoring, modeling, and forecasting the epidemic. *Electronics*, vol. 9, no. 5, p. 827, 2020. doi: 10.3390/electronics9050827.

[2] "WHO COVID-19 Dashboard," ArcGIS. [Online]. Available: https://covid19.who.int/

[3] R. Armitage and L. B. Nellums, "COVID-19 and the consequences of isolating the elderly," *Lancet Public Health*, vol. 5, no. 5, p. e256, 2020. doi: 10.1016/S2468-2667(20)30061-X.

[4] M. Black et al., "Meaningful Integration of data, analytics and services of computer-based medical systems: The MIDAS touch," in *Proc. IEEE 32nd Int. Symp. Computer-Based Medical Systems (CBMS)*, 2019, pp. 104–105.

[5] P. Carlin, "D2.2 Good practice guide - Ethics and Governance Workpage at MIDAS H2020," Unpublished.

[6] B. Cleland et al., "Meaningful integration of data analytics and services in MIDAS project: Engaging users in the co-design of a health analytics platform," in *Proc. 32nd British Computer Society Human Computer Interaction Conf. (BCS HCI)*, 2018, pp. 1–4.

[7] B. Cleland et al., "Insights into antidepressant prescribing using open health data," *Big Data Res.*, vol. 12, pp. 41–48, 2018. doi: 10.1016/j.bdr.2018.02.002.

[8] B. Cleland et al., "Usability evaluation of a co-created big data analytics platform for health policy-making," in *Proc. Int. Conf. Human-Computer Interaction '19*, 2019, pp. 194–207.

[9] B. Cleland et al., "The 'engage' system: Using real-time digital technologies to support citizen-centred design in government," in *User Centric E-Government. Integrated Series in Information Systems*, S. Saeed, T. Ramayah, Z. Mahmood, Eds. Cham: Springer-Verlag.

[10] J. Connolly et al., "Impact evaluation of an emerging European health project: The MIDAS model," *Bus. Syst. Res., Int. J. Soc. Adv. Innov. Res. Econ.*, vol. 11, no. 1, pp. 142–150, 2020. doi: 10.2478/bsrj-2020-0010.

[11] "Elasticsearch portal," Elastic. [Online]. Available: https://www.elastic.co/

[12] C. J. Galvin et al., "Accelerating the global response against the exponentially growing COVID-19 outbreak through decent data sharing." *Diagnostic Microbiol. Infect. Dis.* p. 115070, 2020. doi: 10.1016/j.diagmicrobio.2020.115070.

[13] W. Guo et al., "Diabetes is a risk factor for the progression and prognosis of COVID-19" *Diabetes Metab. Res. Rev.*, p. e3319, 2020. doi: 10.1002/dmrr.3319.

[14] "COVID-19 Open Research Dataset Challenge (CORD-19)," Kaggle. [Online]. Available: https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge

[15] M. Karwowski et al., "When in danger, turn right: Covid-19 threat promotes social conservatism and right-wing presidential candidates," 2020, PsyArXiv.

[16] J. B. Kornum et al., "Type 2 diabetes and pneumonia outcomes: a population-based cohort study," *Diabetes Care*, vol. 30, no. 9, pp. 2251–2257, 2007. doi: 10.2337/dc06-2417.

[17] G. Leban et al., "Event registry: Learning about world events from news," in *Proc. Int. Conf. World Wide Web*, 2014, pp. 107–111. doi: 10.1145/2567948.2577024.

[18] S. Madsbad, "COVID-19 infection in people with diabetes," Touch Endocrinology. [Online]. Available: www.touchendocrinology.com/insight/covid-19-infection-in-people-with-diabetes/

[19] C. Manning et al., *Introduction to Information Retrieval*. Cambridge, U.K.: Cambridge Univ. Press, 2008, pp. 269–273.

[20] "COVID-19 SARS-CoV-2 preprints from medRxiv and bioRxiv," medRxiv. [Online]. Available: https://connect.medrxiv.org/relate/content/181

[21] "MIDAS project website," MIDAS. [Online]. Available: http://www.midasproject.eu/.

[22] "MEDLINE - Description of the database," National Library of Medicine (NLM). [Online]. Available: https://www.nlm.nih.gov/bsd/medline.html

[23] "PubMed search engine," NLM. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/

[24] A. Pietrobelli et al., "Effects of COVID-19 lockdown on lifestyle behaviors in children with obesity living in Verona, Italy: A longitudinal study," *Obesity*, vol. 8, no. 8, pp. 1382–1385, 2020. doi: 10.1002/oby.22861.

[25] J. Pita Costa et al., "Text mining open datasets to support public health," in *Proc. WITS Conf.*, 2017.

[26] J. Pita Costa et al., "Health News Bias and its impact in Public Health," in *Proc. Slovenian KDD Conf. (SIKDD'19)*, 2019.

[27] J. Pita Costa et al., "A new classifier designed to annotate health-related news with MeSH headings," *Artif. Intell. Med.*, submitted for publication.

[28] D. Rankin et al. "Reliability of supervised machine learning using synthetic data in healthcare: Model to preserve privacy for data sharing," *JMIR Med Inform.*, vol. 8, no. 7, p. e18910, 2020. doi: 10.2196/18910.

[29] "Coronavirus news monitor," RavenPack. [Online]. Available: https://coronavirus.ravenpack.com/

[30] R. Alvarez et al., "TAQIH, a tool for tabular data quality assessment and improvement in the context of health data," *Comput. Methods Programs Biomed.*, vol. 181, p. 104824.

[31] L. Snapes. "Lady Gaga, Billie Eilish and Paul Mccartney to play coronavirus benefit," *The Guardian*. [Online]. Available: https://www.theguardian.com/music/2020/apr/06/lady-gaga-billie-eilish-and-paul-mccartney-to-play-coronavirus-benefit

[32] J. Stone, "Coronavirus testing: Government accused of 'misleading the public' amid criticism over figures," *The Independent*, May 2, 2020. Accessed: May 26, 2020. [Online]. Available: https://www.independent.co.uk/news/uk/politics/coronavirus-testing-figures-uk-target-criticism-hancock-a9495621.html

[33] "One world: Together at home could be live aid for the coronavirus generation," The Verge, 2020. [Online]. Available: https://www.theverge.com/2020/4/7/21211716/one-world-together-at-home-benefit-concert-lady-gaga-covid-19-global-citizen

[34] "Coronavirus watch portal," UNESCO AI Research Inst. [Online]. Available: http://coronaviruswatch.ircai.org/

[35] "WHO Director-General's opening remarks at the media briefing on COVID-19 - 11 March 2020," World Health Organization. [Online]. Available: https://www.who.int/dg/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19—11-march-2020

[36] "Emergency use ICD codes for COVID-19 disease outbreak," World Health Organization. [Online]. Available: https://www.who.int/classifications/icd/covid19/en/

[37] J. Zarocostas. "How to fight an infodemic," *Lancet*, vol. 395, no. 10225, p. 676, 2020. doi: 10.1016/S0140-6736(20)30461-X.

[38] "Coronavirus Disease Research Community - COVID-19," Zenodo. [Online]. Available: https://zenodo.org/communities/covid-19/