Journal of Medical Internet Research

Impact Factor (2018): 4.945 - ranked #1 medical informatics journal by Impact Factor
Volume 23 (2021), Issue 8    ISSN: 1438-8871    Editor in Chief: Gunther Eysenbach, MD, MPH

Contents

Viewpoint

Every Child on the Map: A Theory of Change Framework for Improving Childhood Immunization Coverage and Equity Using Geospatial Data and Technologies (e29759)
Sarah Chaney, Patricia Mechael, Nay Thu, Mamadou Diallo, Carine Gachen. ................................................................. 2

Original Papers

Web-Based Guidance Through Assisted Reproductive Technology (myFertiCare): Patient-Centered App Development and Qualitative Evaluation (e25389)
Ellen Sparidaens, Rosella Hermens, Didi Braat, Willianne Nelen, Kathrin Fleischer. ................................................................. 11

Artificial Intelligence–Based Prediction of Lung Cancer Risk Using Nonimaging Electronic Medical Records: Deep Learning Approach (e26256)
Marvin Yeh, Yu-Hsiang Wang, Hsuan-Chia Yang, Kuan-Jen Bai, Hsiao-Han Wang, Yu-Chuan Li. ................................................................. 30

Construction of Genealogical Knowledge Graphs From Obituaries: Multitask Neural Network Extraction System (e25670)
Kai He, Lixia Yao, JiaWei Zhang, Yufei Li, Chen Li. ................................................................. 44

Evaluating Epidemiological Risk by Using Open Contact Tracing Data: Correlational Study (e28947)
Stefano Piotto, Luigi Di Biasi, Francesco Marrafino, Simona Concilio. ................................................................. 58

Association Between Public Opinion and Malaysian Government Communication Strategies About the COVID-19 Crisis: Content Analysis of Image Repair Strategies in Social Media (e28074)
Nasaai Masngut, Emma Mohamad. ................................................................. 69

Review

Ecological Momentary Assessment and mHealth Interventions Among Men Who Have Sex With Men: Scoping Review (e27751)
Viktor Clark, Sunny Kim. ................................................................. 21
Every Child on the Map: A Theory of Change Framework for Improving Childhood Immunization Coverage and Equity Using Geospatial Data and Technologies

Sarah Cunard Chaney¹, MSc, MPH; Patricia Mechael¹, MHS, PhD; Nay Myo Thu², MSc, MBBS; Mamadou S Diallo³, PhD; Carine Gachen⁴, MSc

¹HealthEnabled, Cape Town, South Africa
²UNICEF, Juba
³Data and Analytics Unit, Department of Analysis, Planning & Monitoring, UNICEF, New York, NY, United States
⁴Gavi, the Vaccine Alliance, Health Information Systems and Digital Health Information, Geneva, Switzerland

Abstract

The effective use of geospatial data and technologies to collect, manage, analyze, model, and visualize geographic data has great potential to improve data-driven decision-making for immunization programs. This article presents a theory of change for the use of geospatial technologies for immunization programming—a framework to illustrate the ways in which geospatial data and technologies can contribute to improved immunization outcomes and have a positive impact on childhood immunization coverage rates in low- and middle-income countries. The theory of change is the result of a review of the state of the evidence and literature; consultation with implementers, donors, and immunization and geospatial technology experts; and a review of country-level implementation experiences. The framework illustrates how the effective use of geospatial data and technologies can help immunization programs realize improvements in the number of children immunized by producing reliable estimates of target populations, identifying chronically missed settlements and locations with the highest number of zero-dose and under-immunized children, and guiding immunization managers with solutions to optimize resource distribution and location of health services. Through these direct effects on service delivery, geospatial data and technologies can contribute to the strengthening of the overall health system with equity in immunization coverage. Recent implementation of integrated geospatial data and technologies for the immunization program in Myanmar demonstrate the process that countries may experience on the path to achieving lasting systematic improvements. The theory of change presented here may serve as a guide for country program managers, implementers, donors, and other stakeholders to better understand how geospatial tools can support immunization programs and facilitate integrated service planning and equitable delivery through the unifying role of geography and geospatial data.

Introduction

Maps are powerful tools for public health decision-makers to better understand the relationship between the location of populations and health system resources, indicators or predictors of health status, and their patterns over space and time. The visual power of the map is aided by modern advances in technology, computing, and handheld devices that can record the location of any place on the earth and transmit geospatial data for analysis, sharing, and use. The use of geography to analyze patterns of disease, distribution of populations, and

https://www.jmir.org/2021/8/e29759

(J Med Internet Res 2021;23(8):e29759) doi:10.2196/29759

KEYWORDS
geospatial data; immunization; health information systems; service delivery; equity mapping; theory; framework; children; immunization; vaccine; equity; geospatial; data; outcome; coverage; low- and middle-income; LMIC
inventories and locations of health services come together to create a catalyst for improving health systems.

Immunization programs in low- and middle-income countries are beginning to harness digital maps and geospatial data to display and analyze complex information for program improvements [1-6]. The effective use of geospatial data can show program managers which locations have not received adequate immunization services, provide more accurate denominators, and inform what vaccination delivery strategies should be used to optimize coverage and equity. It can also improve monitoring of immunization programs.

Applications of geospatial technologies for immunization are often approached as simple solutions to system challenges without careful consideration of the greater ecosystem or planning for widespread adoption and sustainability [7]. Interventions are often deployed as pilot technology-focused projects without sustained resources or commitment to support the underlying enabling environment, human capacities, and governance systems that will contribute to a long-lasting impact on decision-making and health outcomes [8]. Gavi, the Vaccine Alliance, supports a systematic approach to understanding the range of geospatial data and technology implementation experiences to guide sustainable and effective systems and governance for improving immunization services that can reach every child with life-saving vaccines while strengthening primary health care systems [9]. Geospatial data and technology applications for immunization align with GAVI’s 2021-2025 strategy and the global Immunization Agenda 2030 strategy [9,10]. In order to provide life-saving services to children who default on the vaccination schedule and “zero-dose” children who have never received a vaccine, new geo-enabled approaches to planning and delivering services are needed to expand the reach of effective vaccination for all children.

### Theory of Change

Complex interventions benefit from collaborative efforts to understand the underlying series of events and changes that will lead to the desired result [11]. A theory of change is a process and framework to help describe this causal pathway and to support critical thinking throughout the project design, implementation, and evaluation cycle [11]. A theory of change for the use of geospatial technologies for immunization programming describes the potential for geospatial technologies to contribute to real-world impacts by optimizing routine immunization program design, implementation, and monitoring to reach all children with immunization services (Figure 1). It was developed as part of a collaboration between GAVI and UNICEF (United Nations Children’s Emergency Fund) to review the state of evidence in the published and grey literature and through consultations with implementers, donors, immunization, and geospatial technology experts, as well as country-level implementation teams [12]. The theory of change is meant to guide future investment and planning of geospatial technologies and systems for immunization programs within a broader context of health system strengthening, to coordinate donor and partner collaboration, and optimize investments in foundations and systems for long-term sustainability and effective use of immunization data for decision-making.

**Figure 1.** Theory of change for the use of geospatial technologies for immunization programming (originally published and adapted from [12]), with permission from Gavi, UNICEF, and HealthEnabled.
Evidence From Research and Implementation Experiences

Overview

Geospatial data and technologies contribute to the following three interrelated immunization outcomes in the theory of change that together strengthen immunization campaigns and routine immunization program coverage and equity:

1. Increase the number of children immunized through improved target setting
2. Optimize immunization resource distribution and location of services
3. Improve the quality, timeliness, and perception of immunization services with equity in coverage between communities

These three outcome pillars are supported by foundations and enablers in the health system and a foundation of essential data that serve to guide the collection, management, and sustainable use of geospatial data and technologies for health. The theory of change is based on evidence and implementation experiences described for each of the three expected outcomes below.

Increase the Number of Children Immunized Through Improved Target Setting

Despite years of improvements in global vaccination coverage and strengthening systems for service delivery, many children remain underimmunized or never come in contact with routine immunization programs [13]. Delivering life-saving immunization services to all children requires an enormous amount of coordination, planning, and resources; microplans are the local-level operational workplans used by immunization managers to systematically compile relevant local data, prioritize activities, maintain adequate stock, and find solutions to service delivery barriers [14]. UNICEF and the World Health Organization (WHO)’s Reach Every District (RED) strategy encourages the use of maps for local-level microplanning activities, which are traditionally hand-drawn sketches of the catchment area based on local knowledge [15]. These sketch maps are often not to scale; inaccurate or incomplete; and do not contain crucial information for microplanning such as distances, road conditions, or geographic barriers that may delay or discourage vaccinator teams from reaching remote areas during door-to-door campaign activities [3,16]. Health system data may contain overlapping borders; settlements that fall outside health boundaries; and inconsistencies in naming, spelling, and classification of service delivery units and settlements [2,4]. The planning tools and delivery strategies to reach all children with immunization services need to expand beyond the current methods to incorporate new digital tools that support local immunization managers to identify and reach areas that have been historically left off maps and microplans [17].

Children who have never received a vaccination can be clustered in settlements or neighborhoods, increasing their risk of contracting a vaccine-preventable disease without the benefit of herd immunity in their communities [18,19]. Due to a variety of socioeconomic and geographic barriers, these children are left “off the map” both literally and figuratively. Geospatial data and technologies for immunization programming can help identify these underreached communities by pinpointing the physical location of all settlements relative to the area’s immunization service delivery locations. Spatially accurate maps are created through a combination of satellite images and field-based data collection to georeference and validate landmarks, inhabited settlements, and infrastructure through a participatory process with district and local immunization managers. They are then used to plan and execute realistic action plans that include outreach activities. Microplans developed with geospatial technologies and data are a cost-effective way to identify settlements missed with traditional microplanning activities that rely on hand-drawn paper maps [2,5,6,16]. With more accurate and reliable information about the locations, characteristics, and number of settlements within their catchment area, managers can plan and prioritize their activities to vaccinate more children and monitor progress both from the local and central levels.

Optimize Immunization Resource Distribution and Location of Services

Deciding how many vaccinators are needed for each catchment area, how many vaccines to send, and where to deploy fixed and outreach vaccination services depends on the number of people being served in each area, their distribution in the area, and the current unmet need for immunization services. The target population, or denominator, is often estimated from the most recent national census, adjusted each year by adding a fixed rate of growth [16]. Unfortunately, outdated census data, variation in growth rates, and population migration and mobility contribute to overestimation of the target population, leading to wasted resources, or underestimation with subsequent shortages and unvaccinated children [16,20,21]. Even with good population estimates, the location of settlements in relation to services measured by distance or travel time impact access and coverage. There is a relationship between complete and timely vaccination status and shorter distance or travel time to the nearest vaccination service, demonstrating how important the location of immunization services and geographic accessibility is for maximal immunization coverage [1,22-25]. To calculate unmet need, aggregate vaccination coverage data for the entire country or province can hide pockets of low coverage and settlements with unvaccinated children, leaving these communities vulnerable to vaccine-preventable diseases [26,27]. These data limitations impact immunization program planning and resource distribution, thereby preventing the timely delivery of life-saving vaccines to all children.

Tools and approaches that utilize geospatial technologies can help immunization managers make more targeted decisions for where and how to focus activities and resources. Precise estimates of population density and distribution for small geographic areas can be generated with a combination of satellite image data, statistical modelling, and sampled survey information to create accurate program targets for planning and monitoring purposes [28,29]. Population distribution estimates can be combined with spatial data on the location of vaccine service posts, road and transportation infrastructure, and geographic barriers to quantify the movement opportunity for people to reach existing services, inform new strategies and...
location of services, and prioritize outreach activities to remote communities [30,31]. Modelled vaccination coverage for small subnational units of measurement can be generated using multiple sources of data to identify pockets with low coverage and, when combined with data on other socioeconomic indicators, can help suggest solutions to overcome the social, gender-related, economic, geographic, or other factors that are preventing access to immunization services [32]. Improved granular data that is visualized to show geographic trends for local populations can help target delivery strategies and resources to increase immunization coverage in the areas that need it the most [33,34].

**Improve the Quality, Timeliness, and Perception of Immunization Services With Equity in Coverage Between Communities**

A number of underlying factors contribute to nonvaccination—from service delivery challenges in the immunization program to community demand, including the caregiver’s perceived quality of immunization services, trust, and respect within the community [35,36]. Pockets of communities that do not receive quality and timely immunization services are susceptible to vaccine-preventable disease outbreaks. Measuring and monitoring these geographic and socioeconomic pockets of inequity is the first step toward promoting equality in coverage [37,38]. Timely and accurate data on program performance such as tracking supply and logistics, frequency of outreach services, and drop-out-rates can be used to improve the quality of services by providing entry points for supportive supervision, improve planning, identify problems, and initiate rapid corrective action for better overall service delivery [8,39]. Vaccine-preventable disease surveillance systems require rapid communication of data that facilitate feedback up and down the surveillance chain for coordinated and appropriate investigation and response [40]. In order to respond quickly to gaps and challenges, local and subnational immunization managers need to have the skills to use data that is collected accurately, transferred quickly, and presented in a way that can trigger action [8].

Mobile technologies and cellular networks provide opportunities to improve data collection, transfer, analysis, and use [41]. The combination of near real-time communication with automatic collection of accurate location data enables field-based teams of vaccinators to report on the number and location of doses delivered and any barriers encountered during immunization campaigns into an integrated dashboard where managers can monitor progress and respond appropriately to challenges and missed communities [42-45]. These daily reports of progress during campaign activities can help inform the next day’s strategy or provide evidence to extend or alter activities to reach all children in the target area [43-45]. For routine immunization services, supervisors can track the progress of mobile vaccination sessions as part of a geo-enabled digital microplan to identify and respond to missed settlements and improve monitoring of the microplan implementation [3,46].

The collection of geographic information linked to reports of suspected vaccine preventable diseases can facilitate rapid and coordinated action to prevent outbreaks, identify high-risk areas that need vaccination services, and facilitate risk-mapping to predict future outbreaks [40,47]. The transparent sharing of data can promote a common understanding of expectations and challenges between vaccinators and supervisors.

**Implementation Experiences: Myanmar Case Study**

The current knowledge base shows that geospatial data and technology applications for immunization have the potential to stimulate programmatic improvements and increase immunization coverage. However, real-life examples of comprehensive and sustainable systems using geospatial data and technologies for immunization are rare. Myanmar provides an example of how the process of integrating geospatial data and technology for immunization microplanning validates the progression of incremental steps outlined in the theory of change.

In 2016, the national immunization program in Myanmar undertook a review as part of a health system commitment to creating a geo-enabled health information system. The assessment uncovered gaps in immunization coverage for children living in geographically and socially hard-to-reach communities, such as migrant worker settlements, remote villages, ethnic minority communities, and conflict-affected areas [48]. The local-level operational immunization workplans lacked reliable population information, and boundaries were out of date. This limited the ability of health workers to plan and undertake the daily logistics of immunization service delivery. In response to these gaps in coverage, the program took steps to support the microplanning process with geospatial data and technologies.

A phased pilot approach began in late 2017 in one township to begin building foundations, local capacity, and standard procedures and to demonstrate the benefits of using geospatial data and technologies for local-level immunization microplanning [48]. Subsequent expansion to a larger region in 2018 built on the foundations and lessons learned from the first pilot, as well as made improvements in the processes and implementation approach. Each expansion phase to a new area lasted 6 to 9 months to ensure that local capacity and systems were strengthened along the way.

The field implementation process created an up-to-date geo-referenced master list of facilities, settlements, and health area boundaries. A master list establishes a standardized, complete, up-to-date, and uniquely coded list of all features essential to the delivery of immunization services. Through this collaborative process, standard definitions were established for the geographic objects relevant to the microplanning process (eg, vaccination sites, facilities, and communities), and procedures were established for standard data collection. Every location where people lived, including temporary migrant settlements, were identified, defined, and included in the master list. Satellite images aided in settlement identification and catchment area delineation. Health workers were important stakeholders in the process to validate and review the maps and make necessary adjustments to their immunization microplans based on available transportation routes, distances, and geographic features in coordination with their supervisors. Online and printed maps showing accurate spatial relationships...
between key immunization assets and communities were produced and made available for national immunization program staff to plan vaccination campaigns and routine service delivery.

The interim results from Myanmar’s phased implementation approach include immediate effects of the collaborative process, map production, and distribution. With settlements and communities well defined, including characteristics and locations of temporary settlements, health workers were able to include these previously overlooked populations in their immunization microplans. The addition of missed settlements improved target population estimates, allowing for improvements in service delivery planning. The transparency and sharing of microplans and maps enabled supervisors to provide better support to health workers and encouraged accountability at all levels. Health officials were able to see the need for expanded health facility distribution with a clear visualization and accurate distances displayed in new microplanning maps.

These experiences validate the expected outputs for the integration of geospatial data and technologies in the theory of change (Table 1). Myanmar’s process of integrating geospatial data and technologies for immunization microplanning demonstrates how the complex challenge of delivering effective vaccinations to every child in countries with underlying health system challenges can benefit from these applications. Based on the implementation experiences in Myanmar’s program, it seems likely that continued expansion and improvements in the geo-enablement of their immunization program will lead to the desired immunization outcomes and overall expanded coverage as the theory of change suggests.

Table 1. Summary of geo-enabled microplanning implementation results from the Myanmar Central Expanded Program on Immunization.

<table>
<thead>
<tr>
<th>Myanmar’s geo-enabled microplanning experiences</th>
<th>Corresponding geospatial data and technology theory of change output</th>
</tr>
</thead>
<tbody>
<tr>
<td>● Settlements that were previously missed are defined, identified, and included in the microplan</td>
<td>Improved identification of zero-dose and underimmunized children through more accurate microplanning and identification of missed settlements to implement appropriate vaccination strategy</td>
</tr>
<tr>
<td>● Visualization of accurate geospatial relationships in catchment areas serve as a tool to plan vaccination sessions</td>
<td>Improved planning and allocation of immunization resources through strengthened use of geospatial data, analysis, and visualization</td>
</tr>
<tr>
<td>● Target population denominator is closer to actual community density and distribution</td>
<td></td>
</tr>
<tr>
<td>● Standardized definitions and categorization of settlements and immunization resources help streamline planning process</td>
<td></td>
</tr>
<tr>
<td>● Visualization serves as an advocacy tool to demonstrate to senior health officials the need for improvements in the equitable distribution and allocation of resources</td>
<td></td>
</tr>
<tr>
<td>● Enhanced geo-enabled microplanning process encourages accountability of health workers and supervisors with transparency and shared expectations and service delivery plans</td>
<td>Improved service delivery through better planning, monitoring, and tracking of immunization activities for rapid problem identification and corrective action</td>
</tr>
</tbody>
</table>

Additional lessons from Myanmar’s experience reinforce the importance of the enabling environment and foundations in the theory of change, built on UNICEF’s guidelines and detailed approach to support the enabling environment for geospatial data and technologies in immunization programs [49]. The high-level commitment to transition to a geo-enabled national health information system in the Department of Public Health in Myanmar, with support and technical guidance from the WHO, GAVI, UNICEF, and the Health GeoLab Collaborative (a center of excellence for the Asia-Pacific region) laid a strong foundation of advocacy, governance, policies, and capacity for the management and use of geospatial data and technologies in the broader health sector [50]. The geo-enhanced microplanning process was further supported by the creation of common master lists for the geographic objects essential to the immunization program, an element recommended in the theory of change as an essential foundation to the sustained and effective use of geospatial data and technologies for immunization programs. A commitment to improving the supportive environment with a dedicated plan and resources to address needs and gaps in these enablers and foundations will promote the sustainable and effective use of geospatial data and technologies and the application of future geospatial innovations for immunization programs.

Discussion: Applying the Theory of Change

As more immunization programs begin to incorporate geospatial data and technologies to help achieve and measure improvements in equitable immunization service delivery, the theory of change for the use of geospatial technologies for immunization programming can guide discussions, decision-making, and consensus building for investment, development, and coordination. The theory of change represents a thought process aimed at understanding the underlying sequence of events that can contribute to sustained and effective improvements and should be considered a roadmap that is subject to change, improvements, and fine-tuning as more country-level experiences bring insights into best practices and real-world challenges. The three pillars and supportive foundations and enablers can help initiate conversations and identify needs and gaps in country immunization programs to make sound decisions for short-term and long-term planning and contribute to improving the broader health system through shared geospatial data, technologies, and resources.

The theory of change may also serve as a framework for operational research and evaluations by suggesting quantifiable...
research objectives that will contribute to the evidence base and help clarify the relationships and determinants of effective application and use of geospatial data and technologies. The use of geospatial data and technologies within immunization programs can improve not only the systematic collection and use of quality and transparent data for programming but also for measuring improvements and incremental achievements throughout the project cycle.

The framework presented here is grounded in lessons from a handful of implementation experiences and existing evidence from the literature. As more countries gain practical experiences in integrating geospatial data and technologies into national immunization programs, best practices will suggest improvements to this theory of change and will help guide other programs on the incremental steps, foundations, planning, and budgeting recommendations that contribute to the sustainable integration of spatial data for immunization programming. A number of global and regional centers are developing and testing practical guidance and also providing technical support, resources, and training to help national programs apply geospatial data and technologies for immunization and other health systems [49,51,52].

**Conclusions**

Effective data use will be necessary to make additional gains in global immunization coverage. Technology can help improve the collection, visualization, and use of data to detect and address inequalities in coverage [8,38]. However, the quality and value of immunization data ultimately depends on the people who are collecting, analyzing, and using the data, not just the technology they are using [8]. Geospatial data and technologies are a means to an end. They can strengthen data-driven decision-making if they are aligned with immunization outcomes in ways that address program needs and reinforce people’s confidence and trust in the resulting data products and analyses. Optimizing the deployment of immunization services to make them accessible for newly identified communities will pave the way for anchoring primary health care services in underserved areas. A focus on investing in and building sustainable and equitable health and immunization systems with strong leadership and capacity to use the geospatial tools and technologies that are appropriate for each country program will be critical for delivering life-saving vaccines to all children.

**Acknowledgments**

The authors thank the Monitoring an Evaluation team at the GAVI Secretariat and all the key informants for providing guidance on the original landscape and theory of change. Special thanks to Rocco Panceira at UNICEF for providing insights and suggestions on the manuscript. The research was supported by Gavi, the Vaccine Alliance (contract no. M&E88730520).

**Authors' Contributions**

SCC led landscape analysis, evidence review, cases study development, and primary author of paper. PM led framing and co-authoring of the landscape; led development of theory of change; and contributed to the structure, review, and revision of the manuscript. NMT provided content for the case study on Myanmar experiences and performed an overall review of the manuscript. MSD provided technical guidance and reviewed and revised the manuscript. CG provided guidance, framing, and input for the landscape and theory of change, as well as technical guidance, review, and revision of the manuscript.

**Conflicts of Interest**

None declared.

**References**


11. Vogel I. Review of the use of Theory of Change in international development. UK Department of International Development. 2012. URL: https://assets.publishing.service.gov.uk/media/57a08a5ded915d3cfd00071a/DFID_Toc_Review_VogelV7.pdf [accessed 2021-01-10]


52. WHO GIS Centre for Health. World Health Organization. URL: https://www.who.int/data/gis [accessed 2021-06-04]

Abbreviations

GAVI: Gavi, the Vaccine Alliance
RED: Reach Every District
WHO: World Health Organization

Edited by G Eysenbach; submitted 19.04.21; peer-reviewed by E Chukwu, M Das; comments to author 18.05.21; revised version received 14.06.21; accepted 05.07.21; published 03.08.21.

Please cite as:
Chaney SC, Michael P, Thu NM, Diallo MS, Gachen C
Every Child on the Map: A Theory of Change Framework for Improving Childhood Immunization Coverage and Equity Using Geospatial Data and Technologies
J Med Internet Res 2021;23(8):e29759
URL: https://www.jmir.org/2021/8/e29759
doi:10.2196/29759
PMID:34342584

©Sarah Cunard Chaney, Patricia Mechael, Nay Myo Thu, Mamadou S Diallo, Carine Gachen. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 03.08.2021. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.
Original Paper

Web-Based Guidance Through Assisted Reproductive Technology (myFertiCare): Patient-Centered App Development and Qualitative Evaluation

Ellen Marie Sparidaens¹, MD; Rosella P M Hermens², PhD; Didi D M Braat¹, MD, PhD; Willianne L D M Nelen¹, MD, PhD; Kathrin Fleischer³, MD, PhD

¹Department of Obstetrics and Gynaecology, Radboud University Medical Center, Nijmegen, Netherlands
²Scientific Institute for Quality of Healthcare (IQ Healthcare), Radboud University Medical Center, Nijmegen, Netherlands
³The Fertility Partnership Center of Reproductive Medicine, Düsseldorf, Germany

Corresponding Author:
Ellen Marie Sparidaens, MD
Department of Obstetrics and Gynaecology
Radboud University Medical Center
PO Box 9101
Nijmegen, 6500HB
Netherlands
Phone: 31 (0)243617351
Email: ellenmarie.sparidaens@radboudumc.nl

Abstract

Background: Providing patient-centered fertility care is known to improve quality of life and can reduce anxiety and depression. In a previous study, we established the need for a web-based app providing personalized information and interactive functionalities among couples undergoing intracytoplasmic sperm injection with surgically retrieved sperm.

Objective: This study aimed to design, develop, and qualitatively evaluate a multifaceted web-based app for infertile couples undergoing intracytoplasmic sperm injection with surgically retrieved sperm during their treatment trajectory.

Methods: The web-based app was developed in three phases: (1) we established a patient-centered functional design, (2) developed the app in collaboration with medical and technical professionals, and (3) qualitatively evaluated the app among couples using a think-aloud method.

Results: The basis of the app is the couple’s visualized treatment trajectory. The app provides personalized and interactive functionalities; for example, customized information and communication options. During qualitative evaluation, myFertiCare was highly appreciated and received a median score of 8 out of 10. The main improvements made upon conclusion of the think-aloud sessions were related to faster login and easier app navigation.

Conclusions: A patient-centered web-based app aimed at guiding couples through their fertility treatment course was systematically designed, developed, and positively evaluated by patients and medical and technical professionals.

(J Med Internet Res 2021;23(8):e25389) doi:10.2196/25389

KEYWORDS
eHealth; infertility; interactive; mobile apps; patient education; patient-centered care; personalized; topic

Introduction

According to the Institute of Medicine, health care should be safe, effective, timely, efficient, equitable, and patient-centered [1]. Patient-centeredness is described as “providing care respectful of and responsive to individual patient preferences, needs and values, and ensuring that patient values guide all clinical decisions” [1]. Providing patient-centered fertility care can improve quality of life and can reduce anxiety and depression [2]. It has been reported that patient-centeredness in fertility care needs improvement [3,4]. Crucial aspects of patient-centeredness in fertility care are provision of adequate information, continuity of care, and active involvement of patients in their treatment course [5-8]. The internet is known to be an important source of information and support for infertile couples [9-11]. A possible instrument for improving...
patient-centeredness is the use of eHealth tools; that is, web-based apps in health care [12]. Previous eHealth initiatives in fertility care aimed mainly at information provision, support, and mental health promotion [13]. These initiatives only contained few interactive web-based components. Aarts et al [13] concluded that these initiatives could be improved by including more interactive and dynamic elements as their key components. Infertile couples are known to specifically prefer personalized information and appreciate being able to communicate with both their treatment team and fellow patients [3,10]. eHealth tools are a promising strategy to empower patients in managing their own treatment trajectory.

In a previous study [14], we established the need for a web-based app providing personalized information and interactive functionalities among couples undergoing intracytoplasmic sperm injection (ICSI) with surgically retrieved sperm. We hypothesized that a web-based app is specifically suitable for this group of patients because of the psychological and physical burden of the multidisciplinary treatment on both partners. Therefore, the aim of this study was to design, develop, and qualitatively evaluate a multifaceted web-based app for use by couples undergoing ICSI with surgically retrieved sperm during their treatment trajectory.

Methods

Systematic Approach

myFertiCare was developed in three phases: (1) we established a patient-centered, functional design for the app; (2) developed myFertiCare in collaboration with medical and technical professionals; and (3) had myFertiCare qualitatively evaluated for usability, with a think-aloud method.

Phase 1: Establishment of a Patient-Centered Functional Design

The functional design of myFertiCare is based on (1) literature review; (2) interviews with an expert panel comprising a gynecologist, a urologist, an embryologist, an expert in patient-centered innovation, and a board member of Freya, the Dutch association for infertility problems; and (3) interviews with a patient panel. This was part of our previous study [14] on the informational needs of couples undergoing ICSI with surgical sperm retrieval. The patient panel consisted of 11 couples, a number that was determined through data saturation. We conducted semi-structured interviews with each couple individually. The data were analyzed using a constant comparative method. The functional design that followed from this process was verified by the clinic’s fertility treatment team and supplemented with their suggestions. Both the expert and the patient panels were enthusiastic about the idea of a web-based app to guide couples through their treatment trajectory. The overall opinion was that the more functionalities an app provides, the better the app, so that people are motivated to use it. The participants specifically valued personalized and interactive functionalities. Various functionalities were suggested, such as being able to view appointments, test results, and information about lifestyle advice; information about the clinic’s fertility treatment team; and communication with physicians and peers. The interviewees also highlighted the safeguarding of confidential information, which they stated should be at the core of app development [14].

Phase 2: Development of myFertiCare

Based on the preferences of the expert and the patient panels, myFertiCare was developed in close collaboration with medical professionals from the department of Reproductive Medicine, Radboud University Medical Center (RUMC), and technical experts from a Dutch company specializing in eHealth. Together, they formed the project team. Development was an iterative process. The desired functionalities of myFertiCare were categorized by the medical professionals as must-have functionalities that had to be available before the app could be made available on the internet, or as nice-to-have functionalities that could be developed later. Subsequently, the must-have functionalities were developed by the technical experts and tested in a test environment by both the technical experts and the medical professionals. Technical adjustments were made as necessary, and the testing cycle was started over again. Once all the must-have functionalities were developed and tested by the technical and medical experts, myFertiCare was made available through the hospital website, the App Store, and Google Play Store to couples undergoing ICSI with surgical sperm retrieval. myFertiCare was also incorporated in the existing hospital information systems. After the app was live, all the nice-to-have functionalities were developed through the same development cycle. After each functionality was iteratively developed and tested, it was immediately made available to all myFertiCare users. The duration of the whole development trajectory was approximately 1.5 years.

Phase 3: Qualitative Evaluation of myFertiCare for Usability Through the Think-Aloud Method

After all the must-have and nice-to-have functionalities were implemented, we began the qualitative evaluation of myFertiCare. In total, 21 couples, who consecutively visited the fertility clinic, were invited by their physicians to participate in the think-aloud sessions. Six couples agreed to participate, which accounted for 9 participants (4 men and 5 women) (Tables 1 and 2). Reasons for nonparticipation were being too busy, already having too much going on, or simply not wanting to participate. Think-aloud is a research method in which participants verbalize their thought processes while interacting with a tool [15]. It provides a valid source of data about participants’ thought processes and can be used effectively in qualitative studies [15]. Our aim was to identify usability flaws and to provide suggestions for design modification.

The participants were individually provided with 16 tasks to perform using myFertiCare while thinking out loud (Figure 1). Of these tasks, 11 were the same for every participant and 5 focused on the specific phase of treatment that an individual was in. By completing these tasks, the participants explored all the functionalities of myFertiCare. The researcher observed the participants and asked questions for clarification where needed. The researcher took notes, also focusing on nonverbal communication of participants. In addition, the sessions were audio-taped. After completing each task, the participants answered 3 task-linked questions (“I find this task easy,” “I find
this information useful,” and “I find this information is in a logical spot”) with a 5-point Likert scale to rate responses. They could also add free comments.

Immediately after each think-aloud session, the participants completed a self-developed questionnaire on their experiences using myFertiCare (Figure 2). The questions were about participants’ attitudes toward usability, privacy, understandability of information, and the usefulness of various functionalities of myFertiCare. The questionnaire consisted of 20 questions: 17 with responses rated on a 5-point Likert scale, 2 to be answered with “yes” or “no,” and 1 open question. Again, the participants could also write free comments.

Table 1. Demographic characteristics of the study participants (N=9).

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Men (n=4)</th>
<th>Women (n=5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (years), median (range)</td>
<td>33 (27-47)</td>
<td>30 (28-36)</td>
</tr>
<tr>
<td>Daily internet usage in a private setting (minutes), median (range)</td>
<td>60 (45-60)</td>
<td>90 (60-180)</td>
</tr>
<tr>
<td>Treatment-related use of the internet, n</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>Use of myFertiCare prior to the think-aloud session, n</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Educational status, n</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Medium</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>High</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Ethnic background, n</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Caucasian</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Non-Caucasian</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Those who already have children, n</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>No</td>
<td>4</td>
<td>5</td>
</tr>
</tbody>
</table>

*aLow educational status includes no education and lower general secondary education. Medium educational status includes higher general secondary education and intermediate vocational education. High educational status includes higher vocational education and a university degree.

*bOne man from Indonesia and 1 woman from Suriname.

Table 2. Demographic characteristics of the couples who participated in the study (N=6).

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Couples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Socioeconomic status, n</td>
<td></td>
</tr>
<tr>
<td>Low</td>
<td>3</td>
</tr>
<tr>
<td>Medium</td>
<td>3</td>
</tr>
<tr>
<td>High</td>
<td>0</td>
</tr>
<tr>
<td>Stage of treatment, n</td>
<td></td>
</tr>
<tr>
<td>Before surgical sperm retrieval</td>
<td>1</td>
</tr>
<tr>
<td>After surgical sperm retrieval and before ICSI b</td>
<td>2</td>
</tr>
<tr>
<td>During first ICSI cycle</td>
<td>2</td>
</tr>
<tr>
<td>After at least one full ICSI cycle</td>
<td>1</td>
</tr>
<tr>
<td>Duration of infertility (months), median (range)</td>
<td>28 (16-47)</td>
</tr>
</tbody>
</table>

*aAccording to the Dutch Social and Cultural Planning Office: Low socioeconomic status included a status score of ≤-1; medium socioeconomic status included a status score between –1 and +1; high socioeconomic status included a status score of >1.

bICSI: intracytoplasmic sperm injection.
Figure 1. Results obtained from think-aloud sessions.

<table>
<thead>
<tr>
<th>Task</th>
<th>Mmen (M) n=4</th>
<th>Women (W) n=5</th>
<th>Easy</th>
<th>Useful</th>
<th>In logical spot</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Go to Rodboudmc website and open myFertiCare</td>
<td>M: 4/5</td>
<td>W: 4/5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. Log in using DigID</td>
<td>M: 4/5</td>
<td>W: 4/5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. Accept invitation for the forum</td>
<td>M: 4/5</td>
<td>W: 4/5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. Search for data, time, location and doctor for your next appointment</td>
<td>M: 4/5</td>
<td>W: 4/5</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Participants before surgical sperm retrieval (n = 5)

<table>
<thead>
<tr>
<th>Task</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>5a. Search for what to bring on the day of surgical sperm retrieval</td>
<td>M</td>
<td>W</td>
<td>M</td>
<td>W</td>
<td>M</td>
<td>W</td>
</tr>
<tr>
<td>5b. Search for information about appointment with the urologist</td>
<td>M</td>
<td>W</td>
<td>M</td>
<td>W</td>
<td>M</td>
<td>W</td>
</tr>
<tr>
<td>5c. Search for location of group meeting on the day of the intake</td>
<td>M</td>
<td>W</td>
<td>M</td>
<td>W</td>
<td>M</td>
<td>W</td>
</tr>
<tr>
<td>5d. Search for the duration of surgical sperm retrieval</td>
<td>M</td>
<td>W</td>
<td>M</td>
<td>W</td>
<td>M</td>
<td>W</td>
</tr>
<tr>
<td>5e. Search for follow-up appointment after surgical sperm retrieval</td>
<td>M</td>
<td>W</td>
<td>M</td>
<td>W</td>
<td>M</td>
<td>W</td>
</tr>
</tbody>
</table>

Participants after surgical sperm retrieval (n = 8)

<table>
<thead>
<tr>
<th>Task</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>6a. Search for what you need to bring on the day of ovum pick-up</td>
<td>M</td>
<td>W</td>
<td>M</td>
<td>W</td>
<td>M</td>
<td>W</td>
</tr>
<tr>
<td>6b. Search for how to sign up for new ICSI when period started</td>
<td>M</td>
<td>W</td>
<td>M</td>
<td>W</td>
<td>M</td>
<td>W</td>
</tr>
<tr>
<td>6c. Search for telephone number to sign up for new ICSI cycle</td>
<td>M</td>
<td>W</td>
<td>M</td>
<td>W</td>
<td>M</td>
<td>W</td>
</tr>
<tr>
<td>6d. Search for information about anaesthesia during ovum pick-up</td>
<td>M</td>
<td>W</td>
<td>M</td>
<td>W</td>
<td>M</td>
<td>W</td>
</tr>
<tr>
<td>6e. Search for when to do pregnancy test after embryo transfer</td>
<td>M</td>
<td>W</td>
<td>M</td>
<td>W</td>
<td>M</td>
<td>W</td>
</tr>
</tbody>
</table>

All participants (n = 9)

<table>
<thead>
<tr>
<th>Task</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>7a. Search for the name of your primary provider</td>
<td>M</td>
<td>W</td>
<td>M</td>
<td>W</td>
<td>M</td>
<td>W</td>
</tr>
<tr>
<td>8a. Make a note</td>
<td>M</td>
<td>W</td>
<td>M</td>
<td>W</td>
<td>M</td>
<td>W</td>
</tr>
<tr>
<td>9a. Post message on the forum</td>
<td>M</td>
<td>W</td>
<td>M</td>
<td>W</td>
<td>M</td>
<td>W</td>
</tr>
<tr>
<td>10. Search for the duration of the walk from the parking lot to the outpatient clinic</td>
<td>M</td>
<td>W</td>
<td>M</td>
<td>W</td>
<td>M</td>
<td>W</td>
</tr>
<tr>
<td>11. Search for the checklist with lifestyle advice and click the items you meet</td>
<td>M</td>
<td>W</td>
<td>M</td>
<td>W</td>
<td>M</td>
<td>W</td>
</tr>
<tr>
<td>12. Go to your personal health record and view your test results</td>
<td>M</td>
<td>W</td>
<td>M</td>
<td>W</td>
<td>M</td>
<td>W</td>
</tr>
<tr>
<td>13. You can log out</td>
<td>M</td>
<td>W</td>
<td>M</td>
<td>W</td>
<td>M</td>
<td>W</td>
</tr>
</tbody>
</table>

* Mean scores on five-point Likert scales: 1 = completely disagree; 3 = completely agree

* This functionality appeared to be unavailable during the first think-aloud sessions due to a technical error. The error was solved after four sessions, so for this task men, n = 1; women, n = 4

ICSI: Intracytoplasmic sperm injection
The think-aloud sessions were conducted until saturation was reached. The duration of a session was approximately 15 to 20 minutes. The audio-taped sessions were transcribed verbatim. Data were analyzed anonymously. An open coding method was applied. We coded quotes that identified usability flaws or provided suggestions for modification of the app’s design. A second researcher verified the coding process. Differences were discussed until consensus was reached. Ethical approval was proposed but was not required according to the local research ethics committee (CMO Arnhem Nijmegen, file# 2016-2485). All participants provided written informed consent.

Results

Phases 1 and 2: Design and Development of myFertiCare

Based on the results of phase 1 [14], myFertiCare was developed as a web-based app available on the RUMC website, Google Play Store, and Apple App Store for couples undergoing ICSI with surgical sperm retrieval at the RUMC. Patients log in using their digital identity (DigiD), which is provided by the government of the Netherlands to assure safe access to all
governmental institutions. This guarantees the safety of couples’ medical data. The apps of both partners are synchronized, so that individuals can also see their spouse’s information.

Figure 3. Screenshot of the myFertiCare app.

myFertiCare contains personalized and interactive functionalities that are divided over 5 tabs:

1. Treatment trajectory: this is the basis of the app. The treatment trajectory is visualized as a subway map in which every stop stands for one of the appointments a couple must have in order to move forward. Couples can see their past appointments and future scheduled appointments with the corresponding data, time, physician, and location, but they can also see future appointments that are not scheduled yet. Thus, couples are better prepared for the upcoming trajectory and know what to expect. Each stop on the subway map provides information about the specific phase of the treatment trajectory and provides advice on how to prepare for the appointment, and, if applicable, informs them of anything they should bring with them for the appointment. Furthermore, users receive support messages before or after certain appointments to comfort them or provide some advice. These support messages are sent via the app or via text message.

2. Notes: users can write notes that are synchronized with their spouse’s notes. For example, couples can compose a topic list with questions they want to ask during their upcoming physician’s appointment.

3. Care providers: an overview of the whole treatment team is provided through photographs, with an individual’s primary care provider on top. Users can ask medical questions to the treatment team, and they are answered within 24 hours.

4. Forum: patients can communicate with peers on the forum. The forum is supervised by a clinician.

5. Lifestyle advice: this is provided as separate checklists for men and women. The aim is to improve treatment outcomes; that is, to improve the chance of retrieving semen through percutaneous epididymal sperm extraction or testicular sperm extraction and concomitantly the probability of conception. Users can click the boxes of the checklist, which are also synchronized with those of their partner.

In addition to the 5 tabs, myFertiCare provides a main menu with general information (eg, contact details and app settings) and a link to the user’s personal health record. In the personal health record, users can see their own test results and read the correspondence between their primary care provider and their family physician.

Finally, for couples who are not yet being treated at RUMC and thus do not have login details, a preview version of myFertiCare is available. In this version, they can view the general treatment trajectory and consult the checklist with lifestyle advice. Thus, they can prepare themselves for their intake appointment.

Phase 3: Qualitative Evaluation of myFertiCare for Usability With the Think-Aloud Method

The think-aloud sessions yielded both positive and negative feedback. Given the aim of the study, we focused on opportunities to improve the app. As described earlier, every participant completed 16 tasks (11 generic and 5 personalized). This resulted in 21 different tasks. After each task, the participants answered 3 task-linked questions (“I find this task easy,” “I find this information useful,” “I find this information is in a logical spot”) on a 5-point Likert scale (1="totally disagree," 5="totally agree"). Figure 1 shows a summary of the results. In general, participants considered the tasks (ie, the functionalities that myFertiCare provides) easy and useful. They also considered that the information was provided logically.

Although the scores for all tasks were high, the participants named some discomforts and suggestions for improving the app.
design. They commented that logging in with their DigiD was too cumbersome, since it consists of a username, password, and verification via text message. It was also noted that moving along the visualized treatment trajectory was difficult. The participants attempted to slide through the treatment trajectory, which was not possible. Instead, they had to click on stops to move to this specific stop. Furthermore, they noticed that the app did not open with the most recent appointment, which was the mode they preferred. When using the forum, participants regretted that they could not delete an erroneous message they had posted. Finally, participants expressed the need for a home button to lead them to the home screen of the app.

At the end of each think-aloud session, the participants completed a questionnaire about their experience using myFertiCare. Figure 2 shows a summary of the results. The participants allocated high scores to all surveyed items that related to usability, understandability of information, the usefulness of various functionalities, and privacy. The men were consistently slightly more critical than the women. The space for writing free comments revealed no additional information. All participants felt that myFertiCare provides an added value to them. All but 1 participant intended to use myFertiCare in the future. In conclusion, myFertiCare was rated 8 out of 10 (Figure 2).

Guided by the think-aloud sessions, we made various improvements in app design. We made it possible for myFertiCare users to create a 4-digit entry code after the first login with DigiD, so that fast but equally safe access was enabled for future use. Furthermore, opening myFertiCare with the most recent appointment was made possible, while proceeding through the treatment trajectory. We added an option to remove a message from the forum after it has been posted as well. A home button was incorporated, which leads users to the app’s home screen.

Discussion

Principal Findings

We designed, developed, and qualitatively evaluated an eHealth app for fertility care in accordance with a methodological framework, based on couples’ information needs and input provided by health care providers. The basis of the app is the visualized treatment trajectory. The app provides both personalized and interactive functionalities, including customized information and communication options. On thorough qualitative evaluation, myFertiCare received high usability ratings. The participants felt that myFertiCare provides an added value during their treatment. The app was rated with a median score of 8 out of 10. The most important improvements after the think-aloud sessions were related to faster login and easier navigation through the app.

A large part of research in fertility care is aimed at the female partner. We chose to include both partners when developing the app, since it is recognized that men should have a well-defined role as an equal partner during fertility treatment, particularly in cases of male infertility [16]. A previous study by Sylvester et al [11] reported that men registered a marked time lag between diagnosis and treatment initiation. They felt “they were in a maze without a map” and expressed the need for detailed information about the treatment plan, including a timetable, so that they could control and manage their lives [11]. With a visualized treatment trajectory as the basis of myFertiCare, we aimed to meet this need and guide couples through their entire treatment trajectory. In our opinion, patient satisfaction with information provided by the clinic is an important indicator of the quality of fertility care, although in fertility care, the focus is often on live birth rates. Alper et al [17] further endorsed this idea.

There is literature available on eHealth initiatives in fertility care [13], primarily on online support groups. In general, there is a lack of initiatives that provide interactive and dynamic elements, and there is a lack of methodological standards for these complex interventions [13]. There has been 1 web-based initiative that provides both information and peer support, which showed high patient appreciation [18]. Furthermore, a web-based community has been reported in which couples could communicate and share information with professionals and peers [19].

Compared to previous initiatives, a strength of myFertiCare is that it provides a large variety of personalized and interactive functionalities centered around the visualized treatment trajectory of the couple. Another methodological strength of our study is its 3-phase systematic approach: first, a functional design for the app was developed through a qualitative assessment of the informational needs of patients; second, myFertiCare was actually developed; and third, myFertiCare was qualitatively evaluated for usability through the think-aloud method. All 3 phases were carried out in collaboration with patients and medical and technical professionals, which is important for successful eHealth development and implementation [20]. Our qualitative evaluation of myFertiCare for usability is crucial since usability evaluations are critical to the success of adopting an interactive eHealth app [20,21]. The think-aloud method is preferred for uncovering usability problems, and it provides understanding of how users interact with myFertiCare [22]. Furthermore, the think-aloud method is especially suitable since we used both a concurrent method (ie, while performing the task) and a retrospective method (ie, immediately after performing the task) to report participants’ thinking, a method that has been suggested for producing optimal data quality [23].

Limitations

Our study also has some limitations. It could be argued that the study population for the think-aloud sessions was relatively small. However, studies have shown that for a think-aloud test, 5 participants are enough for success in identifying usability problems in a user interface [24]. Since we included 9 participants and achieved data saturation, we are confident that we have identified all the possible usability problems. Furthermore, it is known to be a challenge for a researcher to remain consistent when it is necessary to intervene in a think-aloud session; for example, when a participant is unable to complete a task, clarification on a participant’s comment is required, or a participant sidesteps the functionality of interest.
[25]. In these situations, it is important to explain to a participant that it is the aim of the study to identify problems and to invite them to approach the problem otherwise. It has been reported that researchers often unintentionally intervene in theoretically inconsistent ways [25]. We made a conscious effort to achieve a reliable data set by being aware of these limitations and through triangulation of research methods (namely the think-aloud, task-linked questions, and researcher’s observations) and the recording, transcribing, and coding of the interviews.

Practical Implications

This study provides a framework for patient-centered design, development, and evaluation of an eHealth app. Our systematic approach, in which patients and professionals participated in every phase of the process, is particularly suitable in the current era where patient-centeredness is highly valued. Furthermore, we obtained insight into the various functionalities that patients appreciate in a web-based app. The framework we developed for myFertiCare supports professionals in fertility care for guiding patients through their treatment trajectory and delivering patient-centered care. In the near future, myFertiCare will also be evaluated quantitatively. Expansion of eHealth tools to cover the whole fertility care journey and expansion to other medical disciplines is considered of high value. Development of eHealth tools from a patient’s viewpoint is an opportunity to empower patients in managing their own treatment trajectories in the current era of patient-centered care.

Conclusions

We designed, developed, and qualitatively evaluated a multifaceted web-based app, myFertiCare, through a systematic approach in which patients and medical and technical professionals participated in every phase. This app aims to guide couples undergoing ICSI with surgically retrieved sperm through their treatment trajectory. myFertiCare provides personalized and interactive functionalities, facilitating the provision of patient-centered care and empowering patients to manage their own treatment trajectory. The app had a high usability rating and was highly appreciated by both male and female partners.

Acknowledgments

The authors wish to thank all study participants and the medical and technical experts, who contributed to the development and evaluation of myFertiCare. This study did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Authors' Contributions

All authors contributed to the design of the study. EMS performed data collection and analysis, which was supervised by WLDMN and KF. EMS drafted the manuscript. All authors critically revised the manuscript and approved the final version for submission for publication.

Conflicts of Interest

None declared.

References


Abbreviations

DigiD: digital identity
ICSI: intracytoplasmic sperm injection
RUMC: Radboud University Medical Center
Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.
Ecological Momentary Assessment and mHealth Interventions Among Men Who Have Sex With Men: Scoping Review

Viktor Clark¹, BA, MS; Sunny Jung Kim¹, BA, MA, MS, PhD
Department of Health Behavior and Policy, School of Medicine, Virginia Commonwealth University, Richmond, VA, United States

Corresponding Author:
Viktor Clark, BA, MS
Department of Health Behavior and Policy
School of Medicine
Virginia Commonwealth University
830 East Main Street
Richmond, VA, 23219
United States
Phone: 1 8046283443
Email: nowackv@mymail.vcu.edu

Abstract

Background: Ecological momentary assessment (EMA) is a research design that allows for the measurement of nearly instantaneous experiences within the participant's natural environment. Using EMA can help improve recall bias, ecological validity, and patient engagement while enhancing personalization and the ubiquity of interventions. People that can benefit from the use of EMA are men who have sex with men (MSM). Previous EMA studies have been successful in capturing patterns of depression, anxiety, substance use, and risky sexual behavior. These findings are directly relevant to MSM, who have high rates of each of these psychological and behavioral outcomes. Although there is a driving force behind the growing literature surrounding EMAs among MSM, no synthesizing reviews yet exist.

Objective: The aims of this study were to (1) synthesize the literature across fields on how EMA methods have been used among MSM, (2) better understand the feasibility and acceptability of EMA interventions among MSM, and (3) inform designs for future research studies on best evidence-based practices for EMA interventions.

Methods: Based on 4 library databases, we conducted a scoping review of EMAs used within interventions among MSM. The eligibility criteria included peer-reviewed studies conducted in the United States and the use of EMA methodology in an intervention for MSM. Modeling after the Centers for Disease Control and Prevention's Compendium of Evidence-Based Interventions as the framework, we applied a typology that used 8 distinct review criteria, for example, sample size, design of the intervention, random assignment, design of the follow-up investigation, rate of retention, and rate of engagement.

Results: Our results (k=15, N=952) indicated a range of sample sizes; the smallest sample size was 12, while the largest sample size was 120. Of the 15 studies, 7 (47%) focused on outcomes related to substance use or outcomes related to psychological experiences. Of the 15 studies, 5 (33%) implemented an EMA intervention across 30 days. Of the 15 studies, 2 studies (13%) used random assignment, and 2 studies (13%) had quasi-experimental designs. Of the 15 studies, 10 studies (67%) reported acceptable retention rates greater than 70%. The outcomes that had event-contingent prompts (ie, prompts after engaging in substance use) were not as effective in engaging participants, with overall engagement rates as low as 37%.

Conclusions: Our systematic scoping review indicates strong evidence that the EMA methodology is both feasible and acceptable at high rates among MSM, especially, when examining psychological and behavioral outcomes such as negative or positive affect, risky sexual behavior, or substance use. Further research on optimal designs of EMA interventions for MSM is warranted.

(J Med Internet Res 2021;23(8):e27751) doi:10.2196/27751

KEYWORDS

mHealth; men who have sex with men; mobile health; interventions; mental health; sexual health; ecological momentary assessment; behavior
Introduction

Ecological Momentary Assessment

Developed originally from the field of personality and social psychology, ecological momentary assessment (EMA) is a research design with methodological components that allow researchers to measure experiences as close to that moment as possible and within the participant’s natural environment [1]. Methodological strategies in EMAs have included prompting participants at various time intervals using self-report surveys or triggers specific to locations, events, or both [2]. Researchers have indicated that recall-based self-reports can be inaccurate or unreliable measurements of participants’ actual lived experiences [2]. The goal of EMA is to minimize retrospective recall issues and enhance researchers’ ability to measure lived experiences of people in the moment [1].

The ability to measure and potentially intervene in lived experiences in the moment is especially important to impact dysfunctional thoughts, capture psychological distress, or even intervene in harmful behavior [2]. In addition, one of the key benefits and goals of EMA is to provide high levels of ecological validity, or validity that comes from collecting data and implementing an experiment in a participant’s natural setting in real-world contexts. High ecological validity can enhance the ability for research findings to be applied to real-world scenarios, increasing the likelihood of generalizability [3]. Research has also found EMA methods to outperform traditional paper-pencil measurements in the ability to determine needs of clinical interventions more precisely. A primary reason that EMA measurements outperform traditional paper-pencil measurements is that repeated measurements minimize the effect of participants’ current state on results [4]. Finally, technology-based interventions incorporating EMA methods have shown promise in terms of feasibility and acceptability of enhancing intervention outcomes [5].

Men Who Have Sex With Men

Men who have sex with men (MSM) have been found to show high rates of both psychological distress and engagement in various risky behaviors [6]. Specifically, studies have found MSM to endure higher levels of depression, anxiety, substance use, and risk of contracting HIV [7,8]. EMAs have been used among MSM in daily diaries since 2007 [7] and have evolved tremendously into the realm of internet use [9], smartphone technology [10], and interventions [11]. Use of EMAs among MSM is a growing area of research. EMAs have been shown to be highly effective in reaching people who have a history of substance use or other risky behavior, due to the minimization of stigma and enhancement of self-control over privacy, confidentiality, and anonymity [10,12,13].

Scoping Review

The primary purpose of a scoping review is to synthesize current literature surrounding a topic area. Thus, the synthesis produced from a scoping review acts as a summary of available literature, a means to identify key concepts, and a precursor to a systematic review [14]. To the best of the authors’ knowledge, neither a scoping review nor a systematic review has yet been produced on the topic of EMA use among MSM, due to the limited literature surrounding the topic. As a result, the authors intended to conduct a scoping review by applying the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses-Extension for Scoping Reviews) [15], as a contribution toward a future systematic review based on an increase in literature. Authors also acknowledge the limitations of a scoping review, mainly, an inability to make quality assurances, a lack of strong validity, and an inability to hypothesize based on the review.

Scoping Review for EMAs Among MSM

Despite the benefits and clinical implications of using EMA methods for at-risk populations, there have been no reviews compiling the literature of how EMAs have been used among MSM. One growing method for synthesizing theoretical and empirical evidence in the literature is the scoping review [16]. Scoping reviews are considered as means to describe key findings across literature, identify gaps in research, and inform the design of future research studies [16]. Two major benefits of conducting a scoping review are the ability to examine the breadth of the topic of EMA methodologies that are applied to MSM, specifically, within interventions, and the ability to identify knowledge gaps and future directions for the expansion of this area of research. The aims of this study were to: (1) synthesize the literature across fields on how EMA methods have been used among MSM, (2) better understand the feasibility and acceptability of EMA interventions among MSM, and (3) to inform designs for future research studies on best evidence-based practices for EMA interventions.

Methods

Eligibility Criteria

We identified several eligibility criteria that needed to be met for a publication to be included in this scoping review. The eligibility criteria included MSM samples; EMA interventions or assessments or determinations of EMA’s efficacy, acceptability, and feasibility; a publication date within the past 5 years; peer-reviewed studies conducted in the United States; and quantitative data analyses.

Information Sources

We chose 4 prominent databases to retain studies from: Ovid Medline, which focuses on biomedical scholarly literature; Psychological and Behavioral Science Collection, which focuses on mental processes and emotional and behavioral experiences; PsycInfo, which focuses on behavioral and social science research; and Cumulative Index to Nursing and Allied Health Literature, which provides access to health research, specifically, in nursing and other allied health. EMAs have predominately been applied to behavioral or psychological health [17,18] and physical health [19,20]. Therefore, these databases were determined as most relevant and applicable to the specific topic of this scoping review (Figure 1).
Keywords and Search Process

To ensure that we conducted an inclusive review of the literature across databases, we used multiple search terms for our target populations. Step 1 involved searching the terms “homosexuality” and “homosexual.” In step 2, we used the search term “sexual minority men.” Finally, in step 3, we used the search terms “men who have sex with men” and “MSM.” In step 4, we combined all these terms to achieve the most comprehensive review of literature pertaining to our target population of MSM. In step 5, we added the keywords “ecological momentary assessment” or “EMA,” “daily diaries,” and “experience sampling” to identify all EMA-related literature from these databases. In step 6, we combined our MSM search terms with the EMA terms to narrow down to only the most directly relevant articles. Upon reaching step 7, we applied the term “intervention” to further narrow the articles included in this review. Step 8 was comprised of combining all search terms.
from our MSM terms, EMA terms, and intervention terms, which resulted in a total of 129 articles.

**Inclusion Criteria, Exclusion Criteria, and the Iterative Process**

Once we had the initial studies from our database searches, we combined all the study titles and previewed the articles. Articles were further narrowed based on 2 additional inclusion criteria and 2 additional exclusion criteria; studies were excluded if they were not published in the past 5 years and if they were duplicative across databases. Studies were included if they were conducted in the United States and included quantitative data and analyses (Figure 1).

As suggested by the clarity of guidelines in scoping reviews [21], our search and review were conducted in an iterative manner over time. We conducted our first search in January 2020, our second search in May 2020, our third search in September 2020, and the last search in March 2021, as presented, to examine quarterly changes in the literature.

**Analysis**

Modeling after the Centers for Disease Control and Prevention (CDC) Compendium of Evidence-Based Interventions (EBIs) [22], we primarily used 8 distinct criteria to review the final set of eligible publications [9,11,13,23-34]. The EBI criteria and best practices in the compendium, developed and defined by the CDC’s Prevention Research Synthesis project, posit a series of systematic review components for interventions. EBI criteria have been shown to generate significant effects and strong evidence of efficacy in HIV-related outcomes [22]. Based on the compendium of EBIs, our review criteria included (1) citation, sample size, and duration of study; (2) location; (3) random assignment (yes or no); (4) key aspects of the intervention; (5) follow-up (yes or no); (6) occurrence of follow-up after intervention; (7) rate of retention; and (8) rate of engagement (Multimedia Appendix 1). We also conducted a secondary analysis comprised of additional review criteria: recruitment strategy, description of intervention and comparison of arms, specific measurement tools, outcomes of interest, and outcome results (Multimedia Appendix 2).

**Results**

**Risk of Quality in Individual Studies**

Using the CDC compendium of best evidence-based risk reduction for individual-level interventions, we evaluated the strength of each study. None of the studies reviewed met full criteria for best standards of risk reduction in individual-level interventions. However, we continued evaluating based off CDC compendium criteria to determine study designs, intervention elements, and highest standards currently achieved.

**Study Selection**

A total of 129 articles were identified in the preliminary search: 24 from the Psychology and Behavioral Science Collection, 5 from OVID Medline, 58 from PsycInfo, and 42 from Cumulative Index to Nursing and Allied Health Literature. Among these 129 articles, we excluded 114 studies for not meeting additional screening criteria. These 4 criteria that the studies had to meet were: (1) the study was published in the past 5 years, eliminating 44 studies; (2) not a duplicative study, eliminating 38 studies; (3) a US-based study, eliminating 12 studies; and (4) the study used a quantitative analysis, eliminating 9 studies. The qualifying criteria led to a final set of 15 studies in this review [9,11,13,23-34].

**Sample Sizes**

The average sample size across the 15 studies reviewed was 63.5 (SD 31.85). The CDC recommends that each sample is greater than 50 participants per study arm. Among the 15 studies reviewed, 10 (67%) had sample sizes >50, and 4 (27%) had sample sizes ≥100 [9,31,33,34]. The largest sample size was 120, in a study that had a single intervention arm [31]. Of 15 studies, 4 (27%) had multiple intervention arms [11,26,27,30].

**Demographics**

The only 2 demographic measurements reported by all 15 articles were race/ethnicity and age [9,11,13,23-34]. Within the total sample of 952 participants that was developed from a composite of all article samples, the majority of participants (476/952, 50%) were white, and this sample had a mean age of 38.75 years (SD 8.5). Of 15 studies, across 9 studies (60%), income was also reported, with the majority of the composite sample making <$40,000 annually [9,11,13,25,26,31-34]. The final demographic measurement that was majorly reported, across 12 of the 15 studies (80%), was education; the majority (420/952, 44.1%) of this composite sample had at least some college education [9,11,13,24-26,28,29,31-34]. A detailed synopsis of demographics can be found in Multimedia Appendix 3.

**Key Aspects of the Interventions**

From this review, we determined that multiple studies had similar key aspects across the interventions implemented. First, 7 of the 15 studies (47%) focused on multiple types of substance use, including nicotine use, alcohol use, or other substance use (eg, cocaine, methamphetamine, and cannabis). Second, affect and stigma were discussed as primary outcomes for 5 of 15 studies (33%) [11,28-30,34] and secondary outcomes for 3 additional studies of the 15 studies (20%) [27,32,33]. Among 15 studies reviewed, 6 (40%) of them focused on examining feasibility or acceptability of EMA methodologies within the intervention [11,13,23,25,31]. Finally, 60% (9/15) of the studies focused on sexual behavior among men who have sex with men [9,11,23,25,28,30,32-34].

**Random Assignment**

Random assignment occurred in 2 of the 15 studies (13%) included in our review [11,26]. The CDC recommends random assignment as a gold standard, to rule out biases in a systematic way across multiple intervention arms. Since the majority of the studies (11/15, 73%) included in this review had only 1 arm, random assignment was not implemented [9,13,23-25,28,29,32-34]. In the remaining 2 studies that had nonrandomized designs with multiple arms, one study used a quasi-experimental design implementing clinical cutoffs for hypersexuality to determine group membership [30], while the other study assigned groups based on whether participants were recruited in-person or online [9]. Random assignment was considered as
potentially unethical in many of these studies due to their focus on substance use and sexual behavior; thus, a quasi-experimental design was better suited.

**Description of Intervention and Comparison Arms**

Across the articles, there were a wide variety of intervention strategies used for implementation, for example, impacting the duration, mechanism for the intervention, and tasks for intervention participants. The average duration of the intervention was 31.5 days; the shortest intervention was a single-day cross-sectional survey [13], and the longest was a 90-day intervention [31]. Interventions were conducted primarily via mobile devices (14/15, 94%) [9,11,23-34]. Among the interventions conducted through mobile devices, 11 of 14 studies (79%) used texting prompts [13,28-30]. Tasks to be completed were predominately surveys (10/15, 67%) [11,13,24-31] administered through text message (6/15, 40%) [11,24-27,31]. The second-most used modality for data collection was daily diaries (5/15, 33%) [9,23,32-34]. The CDC recommends a clear intervention description, which we found across all the included studies [22].

**Follow-up and Occurrence of Follow-up**

Follow-up assessments were administered in 2 studies of the 15 (13%) [11,31], one of which included follow-ups at 3 different time points: 60 days, 90 days, and 120 days [31], while the other followed up with participants after 4 weeks [11]. According to the CDC compendium, there should be a follow-up, and it should, specifically, occur more than 30 days after completion of the intervention [22].

**Rate of Retention**

Rate of retention was measured in 33% (5/15) of the included studies [11,25-27,31]. Among those that measured retention, the average retention rate was 77.58%, the lowest retention rate was 29.2% [31], and the highest retention rate was 93% [26]. For a high-quality intervention, the CDC recommends a 70% study retention rate [22].

**Rate of Engagement**

Of the 15 studies, 13 (87%) reported rate of engagement [9,13,23-29,31-34]. Engagement rate was defined as an overall rate of completion for text or online surveys, text prompts, or daily diaries, depending on the study modality and was reported consistently across all studies. The average overall engagement rate was 76.93%, the lowest overall engagement rate was 37.3% [9], and the highest was 98.7% [13]. Engagement is a key component of retention, and the 70% retention rate is recommended as the benchmark for an acceptable engagement rate [22].

**Location**

The location of the catchment areas and study sites varied. The majority of studies (9/15, 60%) were conducted on the East Coast [9,13,23,25,28,29,32-34]. Within those conducted on the East Coast, the majority of these studies (5/9, 56%) were concentrated in the Northeast [9,23,32-34]. The second-most researched area was the West Coast (4/15, 27%) [11,26,27,31], with a focus on San Francisco (3/15, 20%) [26,27,31]. Of the 15 studies, 1 (7%) was conducted in the Northwest [24] and 1 (7%), in the upper Midwest [29].

**Recruitment Strategies**

Of the 15 studies, 12 (80%) used more than 2 recruitment strategies. Of those that used at least 2 recruitment strategies, 14/15 (93%) studies paired their strategies with social media (e.g., Instagram or Facebook) [9,11,23-34]. Of the 15 studies, 3 (20%) used cohorts from larger or alternative study sites [11,13,30], one of which used multiple recruitment strategies, including social media [30]. The most popular recruitment strategy that was paired with social media was the use of community-based organizations with in-person recruitment (6/15, 40%) [11,25-27,30,31].

**Specific Measurement Tools**

The majority of the studies (10/15, 67%) were conducted with EMA surveys [11,13,24-31]. Of these 10 studies, 2 (20%) used the same scale to measure affect, the Positive and Negative Affect Scale [9,30]. Of these 10 studies, 2 others (20%) administered the Difficulties with Emotion Regulation Scale to measure emotion dysregulation [28,29]. Many studies used questions such as how many partners a participant engaged in sex with over the past 24 hours (9/15, 60%) [9,11,23,25,30-34], how many standard alcoholic drinks a participant consumed in the past 24 hours (6/15, 40%) [9,24,26,32-34], and what types of drugs were used over the past 24 hours (11/15, 73%) [9,23,24,26,27,29-34].

**Outcomes of Interest**

There were 3 prominent outcomes of interest across the included studies: risky sexual behavior, substance use, and acceptability. Of the 15 studies, 7 (47%) measured substance use status, including use of nicotine, alcohol, and other nonprescription drugs [9,11,24,27,29,33,34]. Of the 15 studies, 6 (40%) measured sexual behavior, especially, risky, unprotected sexual behavior defined by condomless sex or sex with partners who were of unknown HIV status [11,23,30,32-34]. Of the 15 studies, 4 (27%) assessed the acceptability of EMAs implemented in an intervention by measuring response rates, completion rates, and study retention rates [9,25,26,31].

**Outcome Results**

Intervention studies using EMA methods have demonstrated success in longitudinally measuring substance use, compared with studies that relied on non-EMA measurements such as timeline follow-back surveys [9,23,32]. Additionally, EMA methods generated greater acceptability than other methods: daily diaries had high rates of response completion. The highest response rate was 97.3% [9,34], and the lowest was only 84% [23].

**Discussion**

**Principal Findings**

In our scoping review, we aimed to provide an overview of the growing literature on a relatively novel measurement: ecological momentary assessment (EMA). We found that among men who have sex with men (MSM), the majority of EMAs have been...
used to intervene on risk-taking behaviors such as alcohol and drug use or unprotected sex with multiple partners. Although risk-taking behaviors have often been stigmatized, the use of EMA through smartphone technology has been seen as a highly effective way to safely assess risk-taking behaviors [10,12,13]. Overall, EMA was seen as an acceptable and feasible method, with daily diaries as the most acceptable tool [9,23,32-34] to collect the experiences of MSM. A unique facet of MSM research was the successful use of recruitment strategies beyond technology-based recruitment, which included assistance in initiation and engagement from community-based organizations [11,25-27,30,31]. This scoping review was used as a synthesizing method with a wide array of review dimensions and criteria such as quantitative interventions among MSM. This allowed our results to provide a comprehensive set of typological frameworks that may be useful in designing and implementing an EMA-integrated intervention for behavioral change. Basing typographic dimensions off preexisting frameworks offered by the CDC [22], we incorporated the most salient components for intervention research. This allowed for a better assessment of the strength of existing EMA interventions among MSM. Also, we conducted 4 time points of literature searches (ie, January 2020, June 2020, September 2020, and March 2021), for inclusion of more studies, which increased the comprehensiveness of this scoping review.

Our study determined that there were limited psychometrically sound EMA measurements that were fully validated. Given the growing research attention on EMAs within the context of behavioral intervention, future studies may aim to develop and validate EMA measurements. EMA has been widely used as a just-in-time assessment and monitoring tool, but it also can be a great measurement resource to predict behavioral outcomes. We suggest future research should focus on developing predictive models and analytic methods, using intensive longitudinal data from EMAs to understand behavioral changes or outcomes over time. In terms of an analytic perspective, since EMAs lead to extensive longitudinal data, the risk of missingness and the handling of missing data will become more prevalent. Therefore, studies on appropriate analytic approaches to manage missing data from EMAs will be essential.

Although EMAs may reduce recall bias, due to the repetitiveness of measurement, they can also increase participant bias and burden [35,36]. Future research should take into consideration EMA designs that are engaging but protective of data anonymity and confidentiality, to prevent participant biases such as the social desirability effect and the halo or devil effect. To avoid priming of such participant biases, details and information in EMA-based interventions should be presented in a judgment-free manner. In order to reduce psychological reactions, future research should consider developing EMAs as self-motivated mechanisms, with use options such as event-contingent prompts, daily diaries, text prompts, or other mechanisms. Additionally, future studies should examine putative mechanism factors such as resilience and social support, to develop a comprehensive, integrative intervention program for MSM [37].

Limitations
There were limitations imposed by the scope and design of the study. First, the inclusion of the major library databases focused on studies most relevant to the population and methodological strategy of interest, but we excluded other minor library databases. Therefore, a future direction may include an inclusion of minor databases in the review for EMA interventions among MSM. Second, many of the studies included in this review were feasibility and acceptability tests as well as pilot studies with inconsistent assessments of outcomes, thus minimizing the effectiveness of a meta-analysis or systematic review. Although we presented results from a scoping review to provide an overview and the state of EMA in behavioral medicine, future research may conduct a systematic review or meta-analysis, as the prevalence of empirical evidence from randomized controlled trials using EMAs in this area is likely to increase [38].

Conclusions
Leveraging evidence-based intervention designs with validated ecological momentary assessments can advance our understanding of factors and processes in behavioral changes and health outcomes. These approaches can be further empowered through technology-based behavioral medicine and social medicine. In this scoping review paper, we provided a typology of EMA-based intervention research that was designed to promote health behavior and psychological well-being. Advancements in psychometric tests to validate EMAs will be critical. As the empirical evidence and theories in this field are emerging, we hope our review offers some guidance and synthesis of the literature to develop and evaluate technology-based EMA health interventions.

Authors' Contributions
VC conceived the study. VC and SJK were responsible for the study design and concept. SJK was responsible for accurate implementation of scoping review methodology. VC was responsible for the literature search and literature review. SJK supervised the study design and the review process. VC was responsible for constructing figures and multimedia appendices. VC and SJK collaborated on the discussion and study impact. VC and SJK drafted the first draft of the manuscript. All authors extensively reviewed and approved the manuscript before submission for peer review.

Conflicts of Interest
None declared.

Multimedia Appendix 1
Primary scoping review analysis.

Multimedia Appendix 2
Secondary scoping review analysis.

Multimedia Appendix 3
Demographics.

References


Abbreviations

CDC: Centers for Disease Control and Prevention
EBI: evidence-based intervention
EMA: ecological momentary assessment
MSM: men who have sex with men
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses - Extension for Scoping Reviews

© Viktor Clark, Sunny Jung Kim. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 03.08.2021. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.
Original Paper

Artificial Intelligence–Based Prediction of Lung Cancer Risk Using Nonimaging Electronic Medical Records: Deep Learning Approach

Marvin Chia-Han Yeh¹,², MD, PhD; Yu-Hsiang Wang³, MD; Hsuan-Chia Yang⁴,⁵, PhD; Kuan-Jen Bai⁶,⁷,⁸, MD; Hsiao-Han Wang¹,²,⁴,⁹*, MD; Yu-Chuan (Jack) Li¹,²,⁴,⁵,⁹*, MD, PhD

¹Department of Dermatology, Wan Fang Hospital, Taipei Medical University, Taipei, Taiwan
²Research Center of Big Data and Meta-analysis, Wan Fang Hospital, Taipei Medical University, Taipei, Taiwan
³School of Medicine, Taipei Medical University, Taipei, Taiwan
⁴Graduate Institute of Biomedical Informatics, College of Medical Science and Technology, Taipei Medical University, Taipei, Taiwan
⁵International Center for Health Information Technology, Taipei Medical University, Taipei, Taiwan
⁶Division of Pulmonary Medicine, Department of Internal Medicine, Wan Fang Hospital, Taipei Medical University, Taipei, Taiwan
⁷School of Respiratory Therapy, College of Medicine, Taipei Medical University, Taipei, Taiwan
⁸Pulmonary Research Center, Wan Fang Hospital, Taipei Medical University, Taipei, Taiwan
⁹Department of Dermatology, School of Medicine, Taipei Medical University, Taipei, Taiwan
* these authors contributed equally

Corresponding Author:
Yu-Chuan (Jack) Li, MD, PhD
Department of Dermatology
Wan Fang Hospital
Taipei Medical University
No 111, Section 3, Xinglong Road
Wenshan District
Taipei, 116
Taiwan
Phone: 886 29307930 ext 2980
Email: jaak88@gmail.com

Abstract

Background: Artificial intelligence approaches can integrate complex features and can be used to predict a patient’s risk of developing lung cancer, thereby decreasing the need for unnecessary and expensive diagnostic interventions.

Objective: The aim of this study was to use electronic medical records to prescreen patients who are at risk of developing lung cancer.

Methods: We randomly selected 2 million participants from the Taiwan National Health Insurance Research Database who received care between 1999 and 2013. We built a predictive lung cancer screening model with neural networks that were trained and validated using pre-2012 data, and we tested the model prospectively on post-2012 data. An age- and gender-matched subgroup that was 10 times larger than the original lung cancer group was used to assess the predictive power of the electronic medical record. Discrimination (area under the receiver operating characteristic curve [AUC]) and calibration analyses were performed.

Results: The analysis included 11,617 patients with lung cancer and 1,423,154 control patients. The model achieved AUCs of 0.90 for the overall population and 0.87 in patients ≥55 years of age. The AUC in the matched subgroup was 0.82. The positive predictive value was highest (14.3%) among people aged ≥55 years with a pre-existing history of lung disease.

Conclusions: Our model achieved excellent performance in predicting lung cancer within 1 year and has potential to be deployed for digital patient screening. Convolution neural networks facilitate the effective use of EMRs to identify individuals at high risk for developing lung cancer.

(J Med Internet Res 2021;23(8):e26256) doi:10.2196/26256

KEYWORDS
artificial intelligence; lung cancer screening; electronic medical record
Introduction

Lung cancer is a leading cause of cancer death worldwide, and to reduce its mortality, early detection is crucial. The National Lung Cancer Screening Trial (NLST) revealed that screening with low-dose computed tomography (LDCT) can reduce the mortality associated with lung cancer by 20% [1]. Likewise, the Dutch-Belgian Randomized Lung Cancer Screening Trial (NELSON study) recently revealed that screening with LDCT resulted in a 24% decrease in the 10-year cumulative mortality for men and a 33% decrease for women [2]. Multiple organizations have recommended LDCT screening for lung cancer to be used on target populations [3,4]. Given the potential harm due to radiation exposure, false-positive results, and costs associated with LDCT, most organizations only recommend annual screening that targets high-risk individuals; this group is largely identified by epidemiological factors, including age and smoking/cessation history [5]. Furthermore, due to the potential harm associated with false-positive results, the cost-effectiveness of implementing annual LDCT screening remains controversial [6]. Multiple research groups have attempted to overcome this problem by developing risk prediction models to identify patients who might benefit from LDCT screening and to generate criteria that are superior to those introduced by the NLST and related studies [7-14]. These models frequently include self-reported information, such as family history, BMI, socioeconomic status, and smoking/cessation history, and they use conventional regression models for the final risk analysis.

In the era of digital medicine, the use of artificial intelligence has resulted in good performance for predicting image-related tasks, specifically the use of convolutional neural networks (CNNs). In lung cancer research, CNNs have been applied to LDCT and chest radiographic images to facilitate detection and classification of pulmonary nodules; these models demonstrate performance that is comparable to that achieved by human experts [15-19]. The prediction performance is largely based on high-level feature extraction and nonlinear prediction via the use of CNNs. Given proper data conversion, using CNN methodologies to generate predictions using other nonimaging medical data may be possible. Our group recently described a risk prediction model for nonmelanoma skin cancer that was generated using data extracted from electronic medical records (EMRs) [20].

In predicting lung cancer risk, the EMR should be suited to the task of identifying high-risk individuals [21]. In this study, our goal is to develop a risk model for the prediction of lung cancer using data from EMRs. As such, we applied established CNN algorithms to the large data set available in EMRs to identify important patterns associated with the development of lung cancer. In contrast with methods used for traditional regression analysis, we attempted to include evolving sequential information found in EMRs to generate our prediction model. Our goal was to generate a model that facilitated the prospective identification of individuals at higher risk for developing lung cancer; these individuals might then undergo further follow-up examinations, including LDCT. The use of a predictive model to identify individuals at high risk could serve to limit unnecessary radiation exposure and reduce costs associated with LDCT and related interventions.

Methods

Study Population

Deidentified EMRs of 2 million patients who received care between January 01, 1999, and December 31, 2013, were initially sampled from the Taiwan National Health Insurance Research Database (NHIRD). These EMRs included the demographic information, diagnoses, and procedure codes from the International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM) and prescriptions from both outpatient clinical declaration files and in-hospital declaration files. This study included participants between the ages of 20 and 90 years who had at least 4 years of medical records on file. Participants with missing data were excluded. These criteria yielded 1,628,250 EMRs with over 300 million record entries for evaluation and analysis. This study was approved by the Taipei Medical University Institutional Review Board; informed patient consent was waived, as all data were anonymous and deidentified before analysis [22].

Data Preprocessing

Previous validation studies that focused on lung cancer using the NHIRD have shown a positive predictive value (PPV) of 95% [23]. In this study, we provide further validation of the diagnosis of lung cancer using intervention codes (eg, thoracic surgery, subsequent radiotherapy, or chemotherapy) and national catastrophic illness cards (which require definite pathologic proof of a cancer diagnosis). The inclusion and exclusion criteria used in this study are indicated in Figure 1.
The index date for patients with lung cancer was defined as the date of first diagnosis. For the control patients, the index dates were randomly selected from their medical history. ICD-9-CM diagnosis codes and World Health Organization-Anatomical Therapeutic Chemical (WHO-ATC) prescription codes were collected from each case for preprocessing; the date 1 year prior to the index date was used to define the prediction window. The observation window included the 3 years prior to the date included in the prediction window. Thus, we used 3 years of patient medical information to predict the risk of new-onset lung cancer at or within 1 year later (Figure 2). The ICD-9-CM and WHO-ATC codes were preprocessed as described in our previous study [20]. Briefly, the EMRs were classified into 1099 ICD-9-CM code groups and 830 WHO-ATC drug groups. Together, 1929 features were recorded weekly for 157 weeks. For each patient, the diagnoses and medications prescribed at each visit were recorded and converted to an image-like array that preserved temporal information associated with both diagnosis and medication history.

The inputs included age, gender, and an image representing the patient’s 3-year history of diagnosis and medication. The image was input into Xception, a 126-layer neural network, in which feature extraction was performed. The final layer of the Xception network was connected to an average pooling layer and then connected to a fully connected layer with the patient’s age and gender.
We performed 3 subgroup analyses to investigate the performance of the model in different populations. According to the age criteria used in previous trials focused on lung cancer screening [1], patients above and below 55 years of age were included among the subgroups. We also examined patients both with and without previous lung disease [24], including subgroups of patients diagnosed with asbestosis, bronchiectasis, chronic bronchitis, chronic obstructive pulmonary disease (COPD), emphysema, fibrosis, pneumonia, sarcoidosis, silicosis, and tuberculosis. Finally, to focus on the discriminative power of the diagnosis and medication without the confounding effects of age, a subgroup of age- and gender-matched controls was identified.

Model Construction and Evaluation

All patient data were split into training, validation, and testing sets based on their respective index dates. Data with index dates prior to December 31, 2012, were used for training and internal validation, and data with index dates after that date were used for prospective testing. The patients’ age, gender, and image-like arrays described above were used as inputs to generate the model (Figure 2).

Lung cancer risk prediction was treated as a binary classification task using supervised learning. The model was trained to determine whether a given patient was likely to develop lung cancer within 1 year. The Xception architecture [25], which includes a 126-layer CNN-based neural network with a moderate number of parameters, was used for feature extraction. The detailed model structure is shown in Figure 2; the model construction and hyperparameters are listed in Section S1 in Multimedia Appendix 1. During training, class weights based on the population size were set to address data imbalance. To ensure the robustness of the model, a 5-fold cross validation was performed on the model. The performance of the model was assessed by its sensitivity, specificity, and area under the receiver operating characteristic curve (AUC). Model calibration was assessed using a reliability curve and the median absolute error.

To understand the model prediction, occlusion sensitivity analysis was performed by iteratively masking information from a single diagnosis or medication followed by evaluating any changes in the model prediction [26]. In addition, a dimensional reduction technique, t-distributed stochastic neighbor embedding (t-SNE), was performed on the fully connected hidden layer output of the final testing data. We randomly selected 1000 lung cancer patients and 9000 control patients for visualization. The model construction, data preprocessing, model training, and statistical processing were performed using the Python programming language, version 3.6.

Results

Baseline Demographics

A total of 11,617 lung cancer patients and 1,423,154 control patients were identified in our data set. The mean age of the lung cancer group was 66.62 years (SD 14.01); the overall data set included 856,558 (59.7%) men and 578,213 (40.3%) women. The baseline demographics of this patient cohort and the assigned subgroups are summarized in Table 1 and Tables S1-S10 in Multimedia Appendix 1.
Table 1. Demographics of the patients with lung cancer and control patients (N=1,434,771).

<table>
<thead>
<tr>
<th>Group</th>
<th>Patients, n</th>
<th>Age (years), mean (SD)</th>
<th>Male gender, n (%)</th>
<th>Mean diagnosis record count (SD), n</th>
<th>Mean medication record count (SD), n</th>
</tr>
</thead>
<tbody>
<tr>
<td>Whole population</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lung cancer</td>
<td>11,617</td>
<td>66.62 (14.01)</td>
<td>6931 (59.7)</td>
<td>121.62 (113.19)</td>
<td>202.68 (208.97)</td>
</tr>
<tr>
<td>Control</td>
<td>1,423,154</td>
<td>44.95 (16.32)</td>
<td>683,375 (48.0)</td>
<td>66.09 (76.60)</td>
<td>105.99 (135.54)</td>
</tr>
<tr>
<td>Age and gender match (1:10)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lung cancer</td>
<td>11,617</td>
<td>66.62 (14.01)</td>
<td>6931 (59.7)</td>
<td>121.62 (113.19)</td>
<td>202.68 (208.97)</td>
</tr>
<tr>
<td>Control</td>
<td>385,052</td>
<td>66.57 (9.04)</td>
<td>56,730 (48.6)</td>
<td>114.23 (106.76)</td>
<td>184.50 (189.50)</td>
</tr>
<tr>
<td>Age ≥55 years</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lung cancer</td>
<td>9261</td>
<td>71.99 (9.46)</td>
<td>5673 (61.3)</td>
<td>135.12 (116.31)</td>
<td>227.81 (218.12)</td>
</tr>
<tr>
<td>Control</td>
<td>3565</td>
<td>70.79 (12.73)</td>
<td>2244 (63.0)</td>
<td>175.12 (134.36)</td>
<td>297.56 (245.55)</td>
</tr>
<tr>
<td>Age &lt;55 years</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lung cancer</td>
<td>2356</td>
<td>45.50 (7.55)</td>
<td>1258 (53.4)</td>
<td>68.58 (80.42)</td>
<td>103.90 (126.71)</td>
</tr>
<tr>
<td>Control</td>
<td>1,038,102</td>
<td>36.93 (9.85)</td>
<td>496,256 (47.8)</td>
<td>48.23 (51.36)</td>
<td>76.87 (93.45)</td>
</tr>
<tr>
<td>History of lung diseasea</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lung cancer</td>
<td>3565</td>
<td>70.79 (12.73)</td>
<td>2244 (63.0)</td>
<td>175.12 (134.36)</td>
<td>297.56 (245.55)</td>
</tr>
<tr>
<td>Control</td>
<td>182,098</td>
<td>53.01 (18.09)</td>
<td>85,070 (46.7)</td>
<td>125.17 (114.53)</td>
<td>204.85 (204.66)</td>
</tr>
<tr>
<td>No history of lung disease</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lung cancer</td>
<td>8052</td>
<td>64.77 (14.16)</td>
<td>4687 (58.2)</td>
<td>97.94 (93.08)</td>
<td>160.67 (174.80)</td>
</tr>
<tr>
<td>Control</td>
<td>1,270,651</td>
<td>43.77 (15.70)</td>
<td>598,305 (48.2)</td>
<td>57.42 (64.94)</td>
<td>91.48 (115.23)</td>
</tr>
</tbody>
</table>

*aLung diseases included asbestosis, bronchiectasis, chronic bronchitis, chronic obstructive pulmonary disease, emphysema, fibrosis, pneumonia, sarcoidosis, silicosis, and tuberculosis. More information is provided in Table S11 in Multimedia Appendix 1.

Model Performance

For all patients, the model revealed an AUC of 0.821 when the input image-like array included sequential diagnostic information only. By contrast, the AUC was 0.894 when the input features included sequential medication information only; when the sequential diagnostic and medication information was simplified to binary variables, the model performance decreased (AUC=0.827). When both sequential diagnostic and medication information were integrated, the model reached an AUC of 0.902 on prospective testing, with a sensitivity of 0.804 and specificity of 0.837 (Table S12 in Multimedia Appendix 1). The calibration of the model showed a median expected error of 0.125; the reliability curve is shown in Figure S1 in Multimedia Appendix 1.

The model performance at different age cutoffs was then investigated. Screening using an age cutoff of 55 years revealed a superior AUC of 0.871 compared to those obtained when cutoffs of 50 or 60 years were used (0.866 and 0.863, respectively) (Table S13, Multimedia Appendix 1).

Subgroup Analysis

Analyses of the subgroups included one that was both age- and gender-matched, those at ages above and below 55 years, and those with or without lung disease were performed. For this analysis, we identified an age- and gender-matched control subgroup that was 10 times larger than the original lung cancer subgroup. This model revealed an AUC of 0.818 (SD 0.005) with a sensitivity of 0.647 (SD 0.017) and a specificity of 0.873 (0.023 SD), as shown in Table 2 and in Table S14 in Multimedia Appendix 1. For patients above 55 years of age, the model revealed an AUC of 0.869 (SD 0.005) with a sensitivity of 0.784 (SD 0.011) and a specificity of 0.785 (0.016). The PPV in this subgroup was 0.081% (SD 0.011%), and the negative predictive value was 0.993% (SD 0.000%). The performance of the model was inferior in patients below the age of 55 years; however, it still achieved an AUC of 0.815 (SD 0.007). The discriminatory powers of these models were both excellent among patients with and without a history of lung disease; the AUCs for these subgroups were 0.914 (SD 0.003) and 0.887 (SD 0.002), respectively. Among all the subgroups, the model had the weakest performance in patients below 55 years of age who had no history of lung disease; the AUC for this subgroup was only 0.797 (SD 0.008) for the one-year prospective prediction. By contrast, the model provided the strongest performance for individuals above the age of 55 years with a history of lung disease, which revealed the highest PPV of 14.3% (SD 2.3%). The model exhibited the lowest PPV of 1.0% (SD 0.2%) for individuals less than 55 years of age with no history of lung disease (Table 2). The receiver operating characteristic curves associated with each of these subgroups are summarized in sections S2.1-S2.9 in Multimedia Appendix 1.
Table 2. Discrimination performance (testing set) of the model in the subgroups.

<table>
<thead>
<tr>
<th>Subgroup</th>
<th>Lung cancer group, n</th>
<th>Control, n</th>
<th>Testing AUC(^a) (SD)</th>
<th>Testing sensitivity (SD)</th>
<th>Testing specificity (SD)</th>
<th>PPV(^b) (SD), %</th>
<th>NPV(^c) (SD), %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Whole population</td>
<td>1304</td>
<td>138,640</td>
<td>0.898 (0.002)</td>
<td>0.805 (0.015)</td>
<td>0.825 (0.018)</td>
<td>4.2 (0.3)</td>
<td>99.8 (0)</td>
</tr>
<tr>
<td>Matching age and gender</td>
<td>1304</td>
<td>13,040</td>
<td>0.818 (0.005)</td>
<td>0.647 (0.017)</td>
<td>0.873 (0.023)</td>
<td>34.6 (0.4)</td>
<td>96.0 (0.1)</td>
</tr>
<tr>
<td>Age ≥55 years</td>
<td>1046</td>
<td>43,328</td>
<td>0.869 (0.002)</td>
<td>0.784 (0.011)</td>
<td>0.785 (0.016)</td>
<td>8.1 (0.5)</td>
<td>99.3 (0)</td>
</tr>
<tr>
<td>Age &lt;55 years</td>
<td>258</td>
<td>95,312</td>
<td>0.815 (0.007)</td>
<td>0.620 (0.080)</td>
<td>0.838 (0.054)</td>
<td>1.1 (0.2)</td>
<td>99.9 (0)</td>
</tr>
<tr>
<td>History of lung disease</td>
<td>361</td>
<td>16,596</td>
<td>0.914 (0.005)(^d)</td>
<td>0.829 (0.021)</td>
<td>0.816 (0.021)</td>
<td>9.0 (0.8)</td>
<td>0.995 (0.1)</td>
</tr>
<tr>
<td>No history of lung disease</td>
<td>943</td>
<td>122,044</td>
<td>0.887 (0.002)</td>
<td>0.781 (0.025)</td>
<td>0.827 (0.026)</td>
<td>3.4 (0.5)</td>
<td>99.8 (0.0)</td>
</tr>
<tr>
<td>Age ≥55 years with history of lung disease</td>
<td>318</td>
<td>8184</td>
<td>0.875 (0.005)</td>
<td>0.755 (0.047)</td>
<td>0.819 (0.044)</td>
<td>14.3 (2.3)</td>
<td>98.9 (0.2)</td>
</tr>
<tr>
<td>Age ≥55 years with no history of lung disease</td>
<td>728</td>
<td>35,144</td>
<td>0.865 (0.003)</td>
<td>0.775 (0.019)</td>
<td>0.786 (0.018)</td>
<td>7.0 (0.4)</td>
<td>99.4 (0.0)</td>
</tr>
<tr>
<td>Age &lt;55 years with history of lung disease</td>
<td>43</td>
<td>8,412</td>
<td>0.909 (0.006)</td>
<td>0.777 (0.054)</td>
<td>0.891 (0.036)</td>
<td>3.8 (1.0)</td>
<td>99.9 (0.0)</td>
</tr>
<tr>
<td>Age &lt;55 years with no history of lung disease</td>
<td>215</td>
<td>86,900</td>
<td>0.797 (0.008)</td>
<td>0.533 (0.048)</td>
<td>0.865 (0.026)</td>
<td>1.0 (0.2)</td>
<td>99.9 (0.0)</td>
</tr>
</tbody>
</table>

\(a\) AUC: area under the curve.
\(b\) PPV: positive predictive value.
\(c\) NPV: negative predictive value.
\(d\) Italic text indicates the best performance for the parameter.

Table 3 summarizes the age, gender, diagnosis, and medications associated with both the correctly and incorrectly classified groups from the testing data set. The mean age of the true-positive group was similar to that of the false-positive group and somewhat greater than that of the false-negative group. This tendency was also observed in other subgroups; overall, our results suggest that age and sex are important predictive factors. This is consistent with the t-SNE analysis, in which patients with lung cancer and control patients over 55 years of age were clustered centrally, as compared to the other patients, who were located at the periphery (Figure 3).

The model’s hidden layer outputs of 1000 patients with cancer (red dots) and 9000 control patients (green dots) were visualized using t-SNE (Figure 3). Dark green and red represent old age control patients and patients with cancer, respectively. As shown in the left image, most patients with cancer can be clustered away from the control patients. Some dark red dots are mixed with dark green dots in the upper area. These are the patients that were wrongly predicted to be controls by the model. The center images show that patients aged ≥55 years were clustered in the center of the graph, with the patients with cancer were successfully clustered in the tip area. The right image shows that patients aged <55 years were clustered at the periphery of the graph. Some patients with cancer were also clustered in the tip area, whereas the others were scattered with the control patients.

Occlusion sensitivity analysis further revealed that the specific diagnosis and medication factors were associated with an increased risk of developing lung cancer. Interestingly, “other noninfectious gastroenteritis and colitis” and “other agents for local oral treatment” were associated with the highest risks of developing lung cancer with respect to patient diagnosis and medication, respectively. The top 20 factors identified in the analysis are summarized in Table 4.
Table 3. Prediction analysis of the prospective testing data set (N=139,944).

<table>
<thead>
<tr>
<th>Group</th>
<th>Patients, n</th>
<th>Age (years), mean (SD)</th>
<th>Male gender, n (%)</th>
<th>Mean diagnosis count (SD), n</th>
<th>Mean medication count (SD), n</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>All patients</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>True positive</td>
<td>1052</td>
<td>69.91 (11.58)</td>
<td>617 (58.65)</td>
<td>141.75 (113.31)</td>
<td>210.7 (186.32)</td>
</tr>
<tr>
<td>False positive</td>
<td>22,624</td>
<td>69.19 (12.48)</td>
<td>12,641 (55.87)</td>
<td>114.96 (111.04)</td>
<td>159.14 (171.74)</td>
</tr>
<tr>
<td>True negative</td>
<td>116,016</td>
<td>41.94 (13.14)</td>
<td>53,671 (46.26)</td>
<td>63.08 (67.53)</td>
<td>81.46 (101.84)</td>
</tr>
<tr>
<td>False negative</td>
<td>252</td>
<td>50.96 (10.79)</td>
<td>134 (53.17)</td>
<td>81.37 (95.67)</td>
<td>104.03 (139.98)</td>
</tr>
<tr>
<td><strong>Patients aged ≥55 years</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>True positive</td>
<td>851</td>
<td>72.86 (9.25)</td>
<td>510 (59.93)</td>
<td>146.32 (110.84)</td>
<td>217.88 (181.04)</td>
</tr>
<tr>
<td>False positive</td>
<td>10,989</td>
<td>74.88 (9.66)</td>
<td>6640 (60.42)</td>
<td>124.11 (119.27)</td>
<td>170.8 (179.15)</td>
</tr>
<tr>
<td>True negative</td>
<td>32,339</td>
<td>63.28 (6.58)</td>
<td>13,871 (42.89)</td>
<td>110.24 (97.26)</td>
<td>152.69 (154.96)</td>
</tr>
<tr>
<td>False negative</td>
<td>195</td>
<td>64.62 (6.63)</td>
<td>106 (54.36)</td>
<td>125.98 (132.09)</td>
<td>185.08 (216.55)</td>
</tr>
<tr>
<td><strong>Patients aged &lt;55 years</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>True positive</td>
<td>209</td>
<td>47.87 (6.07)</td>
<td>113 (54.07)</td>
<td>83.3 (87.98)</td>
<td>106.48 (128.64)</td>
</tr>
<tr>
<td>False positive</td>
<td>32,765</td>
<td>46.78 (6.58)</td>
<td>18,422 (56.22)</td>
<td>59.4 (63.22)</td>
<td>74.38 (92.27)</td>
</tr>
<tr>
<td>True negative</td>
<td>62,547</td>
<td>32.45 (7.43)</td>
<td>27,379 (42.56)</td>
<td>48.67 (48.88)</td>
<td>60.74 (71.36)</td>
</tr>
<tr>
<td>False negative</td>
<td>49</td>
<td>36.22 (5.82)</td>
<td>22 (44.90)</td>
<td>63.98 (63.75)</td>
<td>83.88 (115.66)</td>
</tr>
<tr>
<td><strong>Patients with a history of lung disease</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>True positive</td>
<td>300</td>
<td>72.86 (11.18)</td>
<td>182 (60.67)</td>
<td>184.91 (118.07)</td>
<td>278.71 (194.81)</td>
</tr>
<tr>
<td>False positive</td>
<td>2791</td>
<td>75.41 (11.97)</td>
<td>1750 (62.70)</td>
<td>180.66 (140.56)</td>
<td>253.68 (214.05)</td>
</tr>
<tr>
<td>True negative</td>
<td>13,805</td>
<td>49.34 (15.6)</td>
<td>5876 (42.56)</td>
<td>119.33 (102.8)</td>
<td>162.24 (162.85)</td>
</tr>
<tr>
<td>False negative</td>
<td>61</td>
<td>61.41 (12.11)</td>
<td>34 (55.74)</td>
<td>171.72 (155.81)</td>
<td>246.79 (226.86)</td>
</tr>
<tr>
<td><strong>Patients with no history of lung disease</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>True positive</td>
<td>757</td>
<td>68.45 (11.4)</td>
<td>442 (58.39)</td>
<td>120.97 (104.28)</td>
<td>177.03 (172.5)</td>
</tr>
<tr>
<td>False positive</td>
<td>23,328</td>
<td>66.54 (12.25)</td>
<td>12,881 (55.22)</td>
<td>95.23 (94.24)</td>
<td>130.24 (146.34)</td>
</tr>
<tr>
<td>True negative</td>
<td>98,716</td>
<td>40.39 (12.27)</td>
<td>45,805 (46.40)</td>
<td>56.19 (59.51)</td>
<td>71.56 (88.63)</td>
</tr>
<tr>
<td>False negative</td>
<td>186</td>
<td>48.19 (10.32)</td>
<td>93 (50.00)</td>
<td>65.08 (66.98)</td>
<td>81.69 (101.83)</td>
</tr>
<tr>
<td><strong>Patients aged ≥55 years with a history of lung disease</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>True positive</td>
<td>255</td>
<td>74.89 (9.03)</td>
<td>160 (62.75)</td>
<td>188.33 (119.58)</td>
<td>284.4 (193.99)</td>
</tr>
<tr>
<td>False positive</td>
<td>1778</td>
<td>78.53 (9.16)</td>
<td>1205 (67.77)</td>
<td>188.16 (142.99)</td>
<td>263 (215.97)</td>
</tr>
<tr>
<td>True negative</td>
<td>6406</td>
<td>66.38 (7.88)</td>
<td>2669 (41.66)</td>
<td>169.82 (121.41)</td>
<td>239.26 (195.71)</td>
</tr>
<tr>
<td>False negative</td>
<td>63</td>
<td>70.44 (7.81)</td>
<td>35 (55.56)</td>
<td>203.87 (148.87)</td>
<td>308.17 (221.29)</td>
</tr>
<tr>
<td><strong>Patients aged ≥55 years with no history of lung disease</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>True positive</td>
<td>587</td>
<td>71.76 (9.24)</td>
<td>347 (59.11)</td>
<td>126.04 (102.89)</td>
<td>185.01 (166.72)</td>
</tr>
<tr>
<td>False positive</td>
<td>8958</td>
<td>73.86 (9.69)</td>
<td>5,281 (58.95)</td>
<td>104.85 (103.3)</td>
<td>142.56 (154.72)</td>
</tr>
<tr>
<td>True negative</td>
<td>26,186</td>
<td>62.73 (6.27)</td>
<td>11,356 (43.37)</td>
<td>98.04 (87.47)</td>
<td>135.09 (139.76)</td>
</tr>
<tr>
<td>False negative</td>
<td>141</td>
<td>63.47 (6.25)</td>
<td>74 (52.48)</td>
<td>100.89 (103.77)</td>
<td>148.73 (195.18)</td>
</tr>
<tr>
<td><strong>Patients aged &lt;55 years with lung diseases</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>True positive</td>
<td>37</td>
<td>48.89 (6.08)</td>
<td>18 (48.65)</td>
<td>120.46 (100.27)</td>
<td>157.62 (173.25)</td>
</tr>
<tr>
<td>False positive</td>
<td>1080</td>
<td>46.56 (7.56)</td>
<td>653 (60.46)</td>
<td>85.56 (72.24)</td>
<td>109.78 (108.74)</td>
</tr>
<tr>
<td>True negative</td>
<td>7332</td>
<td>37.7 (9.58)</td>
<td>3099 (42.27)</td>
<td>86.84 (75.16)</td>
<td>113.06 (116.51)</td>
</tr>
<tr>
<td>False negative</td>
<td>6</td>
<td>43.33 (9.24)</td>
<td>3 (50.00)</td>
<td>103.67 (98.36)</td>
<td>149.83 (152.85)</td>
</tr>
<tr>
<td><strong>Patients aged &lt;55 years with no history of lung disease</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Group</td>
<td>Patients, n</td>
<td>Age (years), mean (SD)</td>
<td>Male gender, n (%)</td>
<td>Mean diagnosis count (SD), n</td>
<td>Mean medication count (SD), n</td>
</tr>
<tr>
<td>-----------------------</td>
<td>-------------</td>
<td>------------------------</td>
<td>-------------------</td>
<td>-----------------------------</td>
<td>------------------------------</td>
</tr>
<tr>
<td>True positive</td>
<td>172</td>
<td>47.55 (6.07)</td>
<td>95 (55.23)</td>
<td>74.94 (83.33)</td>
<td>94.44 (114.72)</td>
</tr>
<tr>
<td>False positive</td>
<td>30,982</td>
<td>46.56 (6.56)</td>
<td>17,478 (56.41)</td>
<td>55.1 (58.63)</td>
<td>68.47 (84.96)</td>
</tr>
<tr>
<td>True negative</td>
<td>55,918</td>
<td>32.06 (7.25)</td>
<td>24,571 (43.94)</td>
<td>45.68 (45.68)</td>
<td>56.64 (65.81)</td>
</tr>
<tr>
<td>False negative</td>
<td>43</td>
<td>35.65 (5.54)</td>
<td>19 (44.19)</td>
<td>59.88 (56.98)</td>
<td>78.84 (108.63)</td>
</tr>
</tbody>
</table>

**Figure 3.** Visualization of the hidden layer of the model using t-stochastic neighbor embedding.

**Table 4.** Top 20 factors related to lung cancer learned by the model.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Factor</th>
<th>Lung cancer risk increase (%)</th>
<th>mean (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Other noninfectious gastroenteritis and colitis</td>
<td>1.85 (1.01)</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Other congenital anomalies of the circulatory system</td>
<td>1.84 (2.21)</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Other agents for local oral treatment</td>
<td>1.76 (1.02)</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Antidotes</td>
<td>1.69 (1.55)</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Postinflammatory pulmonary fibrosis</td>
<td>1.69 (1.43)</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Metronidazole</td>
<td>1.69 (1.29)</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Acariasis</td>
<td>1.65 (1.73)</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Antiviral drugs</td>
<td>1.57 (1.03)</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>Orchitis and epididymitis</td>
<td>1.57 (1.48)</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>Pneumococcal pneumonia</td>
<td>1.52 (0.93)</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>Buflomedil</td>
<td>1.44 (1.76)</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>Danazol</td>
<td>1.42 (1.41)</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>Calcineurin inhibitors</td>
<td>1.42 (1.29)</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>Other disorders of the urethra and urinary tract</td>
<td>1.37 (1.34)</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>Angina pectoris</td>
<td>1.35 (1.44)</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>Other nonorganic psychoses</td>
<td>1.35 (1.99)</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>Respiratory conditions due to other and unspecified external agents</td>
<td>1.33 (1.33)</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>Open wound of back</td>
<td>1.33 (2.46)</td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>Hydrazinophthalazine derivatives</td>
<td>1.31 (1.57)</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>Insulin</td>
<td>1.30 (1.51)</td>
<td></td>
</tr>
</tbody>
</table>
Discussion

Principal Findings

In this study, we explored the possibility of predicting lung cancer using a CNN with diagnosis and medication history extracted from EMRs as a data source. Unlike other proposed lung cancer risk models, our model does not rely on self-reported parameters such as smoking/cessation history, family history, socioeconomic status, or BMI. This model could be readily deployed as a means to evaluate centralized health care databases and perform efficient population-based screening. Such an approach has potential to improve the accuracy of current screening methods, as it can identify those most likely to benefit from interventions [21]. In addition, we attempted to include time-related sequential information as reflected in the medical histories as a means to evaluate lung cancer risk. This approach is different from those used in traditional regression analysis, in which personal history is often simplified and limited to binary or categorical variables. We found that the integration of temporal aspects resulted in improvements in the model performance (Table S12 in Multimedia Appendix 1). The capacity for complex integration of multiple variables is one of the strengths of deep neural networks. To generate this model, we used an established computer vision model (Xception) to extract high-level features from the array representing individual clinical case histories; this ensured that the high-level features associated with the clinical information were effectively extracted for risk prediction.

Related Work

Lung cancer prediction models are under investigation with the goal of identifying high-risk populations that might benefit from LDCT screening. A variety of parameters have been used for prediction, including epidemiologic factors (eg, socioeconomic status, BMI, and smoking history), clinical history (eg, family history and individual history of lung disease history), and results of clinical examinations (eg, blood tests, genetic analysis, and imaging results). The PLCom2012 model is the most widely validated, with AUCs of 0.78 to 0.82 [27-30]. Likewise, the Bach model exhibited AUCs of 0.66 to 0.75 on external validation [5,31]. Other models include the Haggart model, which exhibited AUCs of 0.71 to 0.84 [5,9], the Liverpool Lung Project model, with AUCs of 0.67 to 0.82 [32], and the Lung Cancer Risk Assessment Tool, which achieved AUCs of 0.77 to 0.78 [5,33]. Some models used information extracted from patient EMRs. The model proposed by Iyen-Omofoman et al [10] used lung-associated clinical symptoms and social-epidemiologic factors from a general practice database, and they achieved an AUC of 0.88; likewise, Wang et al [13] included 33,788 clinical features from clinical histories and laboratory tests evaluated in an extreme gradient boosting (XGBoost) model to achieve an AUC of 0.88. With these previous studies in mind, our model featured a deep learning approach and achieved a prospective prediction AUC of 0.87 in patients older than 55 years and 0.90 for the entire patient cohort. It is possible to test other machine learning models (eg, support vector machine or random forest) on our data set. However, this study serves as a proof of concept of using CNN with nonimaging medical records. Comparing the performance of this model to that of different machine learning models of practical interest would be an interesting approach for future studies.

We recognize that direct comparisons between models may not be fully appropriate, as the target populations and predicted outcomes can vary. Previous reports suggested that the performance of models is inflated when nonsmokers and younger subjects (<55 years of age) are included in the study groups [34]. Our findings confirm this point, as can be observed from the higher AUCs associated with the younger age cutoffs (Table S3, Multimedia Appendix 1). Although our data set did not directly include reports of smoking history or cessation, we did include a history of lung diseases (eg, chronic bronchitis, COPD, and emphysema) among our parameters; these could easily be considered as surrogate factors for smoking history. Further analysis of this patient subgroup may help us understand and mitigate the possibility of performance inflation.

In the original NLST trial, the PPV for the LDCT was determined to be 3.4% [1]. The high false-positive rate associated with this intervention remains a major concern with respect to LDCT screening. In this study, the highest PPV (14.5%) was observed in patients ≥55 years of age with a history of lung disease. As noted above, an increase in cancer diagnoses might be expected in this patient subgroup, as a history of lung disease may be a direct consequence of smoking. As such, this finding suggested that individuals in this subgroup are suitable candidates for model prescreening in an effort to avoid unnecessary radiation exposure and costs associated with LDCT. In addition, we found that the 55-year age cutoff selected in the original NLST trial was also appropriate for our model, as the predictive performance was higher with this age cutoff compared to that observed at cutoffs at age 50 or 60 years (Table S3, Multimedia Appendix 1).

Predictive Factor Analysis

The inclusion of an age- and gender-matched subgroup was necessary to explore the roles of clinical diagnosis and medication history in the predictions generated by our model; evaluation of this subgroup prevented the confounding effects of age and its correlations to clinical history (eg, older people are typically prescribed more chronic disease-related medications). With this consideration, our model achieved an AUC of 0.818. These findings can be compared to the model proposed by Spitz et al [12], which included gender-, age-, and smoking status–matched patients and achieved an AUC of 0.63 in former smokers. Although the models generated from matched populations tended to display weaker performance than those from nonmatched populations and may not be clinically useful, this result provided us with a more clear-cut evaluation of the specific parameters included in this model. Taken together, our findings suggest that our model is capable of identifying factors that are useful for predicting lung cancer using clinical information available 1 year before the clinical diagnosis is made.

Our model demonstrated the worst performance in young patients without pre-existing lung diseases. This finding suggests that identifying high-risk patients among young and asymptomatic patients is still the most challenging task. Further
studies are required to assess the performance of the model in patients with different staging. One of the major concerns with respect to the use of lung cancer prediction models is that they tend to select individuals who are older and who have multiple comorbidities [35], thus reducing the overall benefit gained from the screening process [36]. This tendency was also observed in our model. This phenomenon cannot be fully avoided, as it simply reflects the high percentage of older patients in the population who are diagnosed with lung cancer. However, when focused on patients younger than 55 years of age, our model maintained excellent discriminative power (the AUC was 0.82, with a mean age of true positives of 47.8 years).

With the current model, the inclusion of younger individuals remains possible; multiple age-stratified thresholds for lung cancer risk could further optimize the clinical benefits of the predictions from this model.

Although deep learning is often considered a “black box,” and it is often challenging to explain the reasoning behind the outcomes, our study used t-SNE and occlusion sensitivity analysis to identify the most critical of the contributing parameters. Our occlusion sensitivity analysis revealed that many of the important factors were those associated with a history of preexisting lung conditions (eg, postinflammatory pulmonary fibrosis and pneumococcal pneumonia) and medications used to treat smoking-related diseases (eg, buflomedil for peripheral arterial disease and angina pectoris, and insulin for insulin resistance of diabetes mellitus) with increased cancer risk (eg, congenital anomalies of the circulatory system [37] and periodontal conditions [38]), and paraneoplastic phenomena (eg, noninfectious gastroenteritis and colitis [39]). This information must be interpreted carefully, as these findings do not imply a causal relationship. For example, the model may predict an increased likelihood of future lung cancer in patients with pre-existing lung disease simply because these patients receive frequent medical attention; thus, there is a higher likelihood that cancer will be detected incidentally. In addition, the sensitivity analysis in this study is only capable of evaluating one factor at a time; this is a major limitation of the explainability of the model, given the fact that our model was designed to integrate complex, high-level features. Finally, we could not explain some of the predictive features identified by this model, such as the associations with terms including antidote, orchitis, and epididymitis. More studies will be required to decode the findings from the CNN and to elucidate the interactions between age, sex, previous diagnoses, and medications.

Although our model achieved excellent discriminative performance, poor calibration was noted, together with the fact that direct numeric output would overestimate the actual risk. This is a known phenomenon associated with modern neural networks [40]. Unlike the traditional logistic regression models, which perform well in calibration because they directly minimize the loss of calibration, modern neural networks tend to perform suboptimally in this regard. This is likely due to the regularization methods (eg, dropout and batch normalization) and the multiple deep layers applied as components of the model architecture [40]. In our study, poor calibration did not limit the use of the model, as individuals were selected based on a predefined threshold identified in the validation data set rather than on the numerical output of the model. As a result, the increased rates reported in Table 4 do not represent the actual cancer risk.

Our model used nonimaging medical information from EMRs; however, we still used CNN as the model backbone. The study design and aims are different from other lung cancer studies that used CNN to analyze computed tomography (CT) scans and determine if a pulmonary nodule is malignant. Their models were used to automatically identify suspicious nodules from CT scans, which were already present, whereas our model attempted to identify patients with high risk of developing lung cancer in the future.

Limitations

There are several limitations to this study. First, the data collection was limited to the NHIRD database of Taiwan; the patient records do not include tissue histology or lung cancer staging data. Patients with small cell lung cancer and mutation-rich non–small cell lung cancer (eg, epidermal growth factor receptor, anaplastic lymphoma kinase, ROS-1) could not be separated. These specific types may have different disease courses and risk factors; therefore, they were usually not included in the traditional screening, and the benefit of receiving screening is undetermined. Our subgroup analysis did include only patients with pre-existing lung diseases, but this did not mitigate the issue entirely. Similarly, the NHIRD database does not include information on patients’ lifestyles or any genetic or laboratory data. A subgroup analysis of patients with lung cancer based on tissue histology and staging might help to develop a prediction model that was tailored to different risk groups. Second, the data set did not contain any information on smoking status, which is clearly an important risk factor associated with lung cancer development. This limitation restricted the external validation and the comparisons that could be made between our model and those described in earlier published studies. The authors believe that self-reported information, such as family history, smoking/cessation history, and duration of symptoms, are valuable pieces of information for lung cancer prediction that are very important and can further improve prediction accuracy. In our study, a history of lung diseases (eg, COPD and emphysema) was used as a proxy for a smoking history; our model performed with excellent discriminative power with respect to this subgroup. Finally, the NHIRD includes primarily Taiwanese people; as such, the target population was fairly homogeneous, with limited ethnic diversity. The identified risk factors may not apply to other populations with other ethnicities. Nonetheless, the methodology used here could be easily applied to other medical databases with more diverse patient populations.

Conclusion

Our CNN model exhibited robust performance with respect to the 1-year prospective prediction of the risk of developing lung cancer. As our model included sequential data on clinical diagnoses and medication history, it was capable of capturing features associated with evolving clinical conditions and as such was able to identify patients at higher risk of developing lung cancer. With appropriate ethical regulation, this model may be...
deployed as a means to analyze medical databases, thus paving the way for efficient population-based screening and digital precision medicine. A future randomized controlled trial will be required to explore the clinical benefit of this model in diverse populations.

Acknowledgments
This research was funded in part by Ministry of Education (MOE) grants MOE 109-6604-001-400 and DP2-110-21121-01-A-01.

Authors' Contributions
MCHY contributed to the data analysis, model construction, interpretation of results, drafting of the manuscript, and literature review. YHW and HCY contributed to the data curation and data preprocessing. KJB contributed to the investigation and the interpretation of the results. HHW contributed to the interpretation of results, conceptualization, supervision, and manuscript editing. YCL contributed to the conceptualization, supervision, manuscript editing, and interpretation of the results. HHW and YCL contributed equally to this article. The corresponding author, YCL, affirms that the manuscript is an honest, accurate, and transparent account of the study being reported; that no important aspects of the study have been omitted; and that any discrepancies from the study as planned (and, if relevant, registered) have been explained.

Conflicts of Interest
None declared.

Multimedia Appendix 1
Supplementary tables and figures.

References


Abbreviations

ATC: Anatomical Therapeutic Chemical
AUC: area under the receiver operating characteristic curve
CNN: convolutional neural network
COPD: chronic obstructive pulmonary disease
CT: computed tomography
EMR: electronic medical record
ICD-9-CM: International Classification of Diseases, Ninth Revision, Clinical Modification
LDCT: low-dose computed tomography
MOE: Ministry of Education
NHIRD: National Health Insurance Research Database
NLST: National Lung Cancer Screening Trial
PPV: positive predictive value
t-SNE: t-distributed stochastic neighbor embedding
WHO: World Health Organization
XGBoost: extreme gradient boosting

©Hsuan-Chia Yang, Yu-Hsiang Wang, Hsuan-Chia Yang, Kuan-Jen Bai, Hsiao-Han Wang, Yu-Chuan (Jack) Li. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 03.08.2021. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research.
Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.
Construction of Genealogical Knowledge Graphs From Obituaries: Multitask Neural Network Extraction System

Kai He¹,²,³, PhD; Lixia Yao⁴, PhD; JiaWei Zhang¹,²,³, MSc; Yufei Li¹,²,³, PhD; Chen Li¹,²,³, PhD

¹School of Computer Science and Technology, Xi’an Jiaotong University, Xi’an, China
²National Engineering Lab for Big Data Analytics, Xi’an Jiaotong University, Xi’an, China
³Shanxi Province Key Laboratory of Satellite and Terrestrial Network Technology Research and Development, Xian, China
⁴Department of Health Sciences Research, Mayo Clinic, Rochester, MN, United States

*these authors contributed equally

Corresponding Author:
Chen Li, PhD
School of Computer Science and Technology
Xi’an Jiaotong University
Xianning West Road, 27th
Xi’an, 0086 710049
China
Phone: 86 158 0290 2703
Email: cli@xjtu.edu.cn

Abstract

Background: Genealogical information, such as that found in family trees, is imperative for biomedical research such as disease heritability and risk prediction. Researchers have used policyholder and their dependent information in medical claims data and emergency contacts in electronic health records (EHRs) to infer family relationships at a large scale. We have previously demonstrated that online obituaries can be a novel data source for building more complete and accurate family trees.

Objective: Aiming at supplementing EHR data with family relationships for biomedical research, we built an end-to-end information extraction system using a multitask-based artificial neural network model to construct genealogical knowledge graphs (GKGs) from online obituaries. GKGs are enriched family trees with detailed information including age, gender, death and birth dates, and residence.

Methods: Built on a predefined family relationship map consisting of 4 types of entities (eg, people’s name, residence, birth date, and death date) and 71 types of relationships, we curated a corpus containing 1700 online obituaries from the metropolitan area of Minneapolis and St Paul in Minnesota. We also adopted data augmentation technology to generate additional synthetic data to alleviate the issue of data scarcity for rare family relationships. A multitask-based artificial neural network model was then built to simultaneously detect names, extract relationships between them, and assign attributes (eg, birth dates and death dates, residence, age, and gender) to each individual. In the end, we assemble related GKGs into larger ones by identifying people appearing in multiple obituaries.

Results: Our system achieved satisfying precision (94.79%), recall (91.45%), and F-1 measures (93.09%) on 10-fold cross-validation. We also constructed 12,407 GKGs, with the largest one made up of 4 generations and 30 people.

Conclusions: In this work, we discussed the meaning of GKGs for biomedical research, presented a new version of a corpus with a predefined family relationship map and augmented training data, and proposed a multitask deep neural system to construct and assemble GKGs. The results show our system can extract and demonstrate the potential of enriching EHR data for more genetic research. We share the source codes and system with the entire scientific community on GitHub without the corpus for privacy protection.

(J Med Internet Res 2021;23(8):e25670) doi:10.2196/25670

KEYWORDS

genealogical knowledge graph; EHR; information extraction; genealogy; neural network
Introduction

Anthropologists often use oral interviews, historical records, genetic analysis, and other means to obtain genealogical information and draw family trees. When combined with a detailed medical history and social and economic relationships, family trees are considered the x-ray of the family and have been used by clinicians to assess disease risk, suggest treatments, recommend changes in diet and other lifestyle habits, and determine a diagnosis. In the United States, the Medicare Access and CHIP Reauthorization Act of 2015 [1] and Meaningful Use program [2] have incentivized the growing adoption of electronic health records (EHR) with the goal to improve the quality of health care delivery systems. Consequently, a vast amount of EHR data has become available for research purposes in the past decade. However, most EHR systems today do not capture the family relationships between patients by design. Nor do they capture the death information unless patients die in the health care system or the EHR system is linked to external death registries. Constructing family trees for patients becomes an urgent need to unlock the full potential of EHR data in understanding disease and trait heritability, evaluating individuals’ health risks, and exploring environmental effects on human health.

Early exploratory works have combined EHR data and family trees for biomedical research. For instance, Mayer et al [3] used twin or multiple relationships to assess the concordance rates for muscular dystrophy and fragile X syndrome in the twin cohort. Schalkwyk et al [4] conducted interviews with family members to build 3-generation family trees with medical chronologies and demonstrated their use in deciding the services required for the psychological well-being of all family members. Wang et al [5] combined diagnosis codes and dependent coverage under medical plans to estimate the heritability and familial environmental patterns of 149 diseases and inferred the genetic and environmental correlations between 29 complex diseases [5]. Similarly, Polubriaginof et al [6] built more than 595,000 family trees from emergency contact information in a large EHR system and then estimated the heritability of 500 traits.

Constructing high-quality large family trees has been a challenging task. Historically, only famous politicians, philosophers, scientists, religious groups, or royal families were tracked elaborately by genealogists. For such reason, large databases of family trees rarely existed, despite their research value. Recently, a few studies automated family tree collection using innovative informatics approaches. For instance, Mayer and colleagues [3] used shared dates of birth and last names, in addition to home addresses, billing accounts, and keywords of “twin” and “triplet” in unstructured clinical notes to identify a cohort of 19,226 twins or multiples in an extensive health care delivery system. Wang et al [5] inferred 128,989 nuclear families from a large medical claims database covering one-third of the US population based on dependent coverage. Polubriaginof et al [6] used the emergency contact information of 3,550,598 patients from three large EHR systems in New York City to build 595,000 pedigrees. However, these indirect sources, like dependent coverage and emergency contact, have inherent limitations for inferring genealogical information: they do not differentiate biological from nonbiological relationships and they cover only limited types and numbers of family relationships. More specifically, medical insurance in the United States is limited to a beneficiary’s spouse and dependents up to age 26 years. Most patients only provided one or two emergency contacts rather than their whole families in their medical records. Missing relationships could be substantial.

Inspired by the work of Tourassi et al [7] and Yoon et al [8], we began to explore online obituaries as a novel data source for the automatic extraction of genealogical information. Obituaries generally cover many more family members with more detailed and accurate descriptions of their family relationships. Nowadays, local newspaper and funeral service companies often publish obituaries on internet, making the cost of obtaining obituaries minimal. In our previous work, we developed and evaluated a new method of name entity recognition (NER) for extracting family members’ names and relation classifications (RCs) for classing the pairs of names among family members mentioned in online obituaries [9]. In this work, we advanced our previous work in the following 5 aspects: (1) for the NER task, we processed more entity types, including people’s name, age, residence, and dates of birth and death; (2) for RC, we also matched residence entity and related people (in the previous work, we only extracted the family relationships among people entity); (3) we parsed two kinds of special language patterns in obituaries, last name distributive and name with parentheses; (4) all the triplets of family relationships were integrated to build the enriched family trees with additional rule-based inference; and (5) in terms of training data, we normalized the family relationships (see details in Data section).

Traditionally, NER and RC were considered two separate tasks for information extraction. NER sought to extract named entities mentioned in unstructured text into predefined categories, whereas RC classified the relations between those extracted entity mentions. Researchers built natural language processing (NLP) pipelines with multiple modules to accomplish specific tasks. However, such modular separation suffered from 3 major issues leading to suboptimal results: (1) errors from the NER propagated to RC, (2) it was computationally redundant and time-consuming as the system had to pair up every two named entities to classify their relations, and (3) the pipeline model could not take full advantage of the knowledge inhabitant in the relationships of 2 or more named entities. For instance, if the system detected a live in relationship between two named entities in obituaries, the first entity is likely to be a person’s name and the second entity is likely to be a location.

Thus, we look at multitask models that can simultaneously handle multiple related tasks and optimize their learning abilities by sharing the knowledge learned in all or some of the tasks [10]. In 2008, Collobert [11] introduced a single neural network architecture that solved NLP tasks such as part-of-speech tagging, chunking, named entity recognition, semantic role identification, and semantically similar word grouping using one language model. Recently, there are 3 prevailing solutions for multitask NLP models. The most popular solution establishes a common neural network presentation space for all tasks followed by task-specific classifiers [12,13]. The second

In this work, we first updated our annotated corpus by defining a family relationship map to normalize various family relations (see details in Data section). We also used data augmentation technology to generate more synthetic data (sentences), in order to address the imbalanced training data issue and boost the performance on rare classes [19]. After that, we proposed an end-to-end information extraction system based on a multitasking solution. The end-to-end system included a knowledge inference layer for gender inference based on name and relationship mentioning. In the end, we constructed family trees centered on the deceased. These family trees contained many family members with detailed information, including age, gender, death date, birth date, and residence. We named such enriched family trees genealogical knowledge graphs (GKGs). These GKGs could be linked to external EHR data in Minnesota by personally identifiable information (PII), in a similar way as Sauver et al [20] did. We empirically estimated the upper bound of the mapping precision could be around 80% to 90%. It would significantly enhance the power of EHR data to study disease and trait heritability, evaluate an individual’s health risks, and explore environmental effects on the human health.

### Methods

#### Data

We collected 12,407 obituaries published from October 2008 to September 2018 from 3 funeral services websites and 1 local newspaper in the Twin Cities area, metropolitan Minneapolis–Saint Paul. Our data sources were limited to openly available obituaries. Considering the PII embedded in online obituaries, we decided to take a cautious and conservative position in our work by marking up the last name of any real people with the symbol XX (see more details on privacy protection in the Discussion section). After data cleaning, we randomly sampled 1700 obituaries for annotation. We developed the annotation guideline and trained 3 annotators to annotate each of the 1700 obituaries independently. The interannotator agreement measured by F-1 was 82.80% [9]. Table 1 shows the summary statistics of the annotated corpus. There were two unique language patterns in obituaries, namely last name distributive and name with parentheses (see Table 2 for examples). These patterns might be due to the word limitation when the family paid for publishing an obituary in printed newspapers. They required special treatment, as described in the next session of end-to-end system.

| Table 1. Summary statistics of the annotated corpus. |
|-----------------------------|-----------------------------|-----------------------------|-----------------------------|-----------------------------|
|                         | Count | Deceased person | Count | Special language patterns | Count |
| Sentences                | 28,317 | Full name       | 1551 | Last name distributive     | 4954 |
| Names                    | 27,108 | Age             | 1379 | Name with parentheses      | 7504 |
| Family relationship      | 25,557 | Death date      | 1557 | Spouse’s name              | 5993 |
| Residence                | 7161   | Birth date      | 1368 | Previous last name         | 1511 |
| Name-residence pair       | 7954   | __             | __   | __                         | __   |

*a* All counts are the number of occurrences except for the full name of the deceased. Considering all obituaries have structured metadata giving the full names of the deceased more precisely, we only annotate and extract the first-time mention of a full name of the deceased in an obituary. Spouse’s name and previous last name are 2 categories in the content name with parentheses.

*b* Not applicable.

| Table 2. Examples of unique language patterns in obituaries. |
|-----------------------------|-----------------------------|
| Language pattern            | Example                                    |
| Last name distributive      | He is survived by grandsons Addison and Owen XX. |
| Names with parentheses      | XX is also the last name for Addison.       |
| Previous last name          | Anne was born March 20, 1952, to William and Isabel (Starr) XX. |
| Spouse’s name               | Starr is the maiden name for Isabel XX.     |
|                             | Survived by her sons, Dale (Mary) and Bruce (Diana). |
|                             | Dale’s wife is Mary, and Bruce’s wife is Diana. |

In this work, we made two improvements in the corpus annotations. First, we created a family relationship map that normalized various family relationship mentions to 71 family relationship groups. For example, there were many mentions of “born to (name),” “daughter of (name),” and “son of (name)” in obituaries, which were equivalent to express the parent of...
the deceased. We grouped them into the “parent” relation. Similarly, we treated “married to” the same way as the “spouse (of)” relation. Figure 1 shows the family relationship map, consisting of 8 generations and 71 normalized family relationships. The numbers in parentheses were the number of occurrences of a specific family relationship in our corpus.

**Figure 1.** Family relationship map in the obituary corpus.

It was observed that some family relationships, such as granduncle, uncle-in-law, and half-sister had small numbers of cases that was not sufficient to train a high-performance neural network model. Therefore, we used data augmentation technology [19] to expand the corpus and alleviate the imbalanced data issue. We first introduced $w_i$, the weight of relation $i$:

$$w_i = \frac{c_i}{n}$$

Where $c_i$ stood for the count of annotated sentences with relation $i$, and $n$ was 71, the count of all family relationship groups defined in the family relationship map. For each family relationship $i$, the number of training sentences to be generated, $g_i$, was computed as follows:

$$g_i = \frac{g \cdot c_i}{N}$$

Where $N$ was the total number of all human annotated sentences and was the user-defined ratio for data augmentation. Essentially we generated more synthetic sentences to ensure each family relationship had no less than 200 examples, with the constraint that the count ratios of all family relationships remain as close as possible to those in the original training data. After deciding $g_i$, 2 steps were performed to generate extra sentences. First, we randomly chose $g_i$ sentences from the raw
corpus and replaced one of the raw family relationship tokens in these sentences with relationship word $i$. Second, we randomly chose one of the following operations introduced by Wei and Zou [19] to generate the final augmented sentences:

- Synonym replacement: randomly replace $n$ non-stop words with their synonyms in the sentence
- Random insertion: randomly insert a word’s synonym before or after the chosen non-stop word in the sentence
- Random swap: randomly swap 2 non-stop words in the sentence
- Random deletion: randomly remove a non-stop word in the sentence

The 4 entity types of interest in this work, name, residence, birth date and death date, are exempt from the changes. It should also be noted that the generated sentences could not be guaranteed to be grammatically and semantically correct. However, for neural network models, such sentences, when created with appropriate $\alpha$, were demonstrated to improve models’ generalizability as noisy training data.

End-to-End System

Figure 2 illustrates our end-to-end system. It took a list of segmented sentences in an obituary as the input and generated a GKG centered around a deceased person. Its core was a multitask system that combined common parameter sharing across different modules and custom tagging schemes. The multitask solution promised better performance, as it used more supervision information and understood data from different views [21]. The 4 modules were (1) named entity recognition and relation classification through a joint training model and customized tagging scheme, (2) matching locations to people’s name, (3) a parser for resolving last name distributive, and (4) a parser for resolving names with parentheses. These 4 modules shared the same model parameters, as they were trained jointly using one common weighted loss function. Among these modules, module 2 needed the extracted names and locations from module 1 as inputs. Module 5 was added as an independent rule-based layer for gender inference and age, date of death, and birth inference. Eventually, the results of these modules were combined to construct the GKGs.

Figure 2. End-to-end extraction system to parse obituaries and generate genealogical knowledge graphs.
Module 1: Joint NER and RC

This module aimed to extract family members’ names, relationships, and additional attributes of people (residence, age, death date, birth date). Gender was usually not explicitly mentioned in the obituaries, so we inferred the gender in module 5. We adopted a customized tagging scheme (shown in Figure 3) when annotating the training data. Each tag consisted of 2 parts. The first part indicated the type of an entity, and the second part illustrated the position of the word in the entity. As shown in Figure 3, “sister_B,” “sister_I,” “sister_E,” and “Age_S” indicated the beginning, the inside, and the end of a sister entity and a single-word entity of age, respectively. In the system, the deceased was the default baseline entity for all family relationship triplets. In the sentence shown in Figure 3, for example, “Robert” was the name of the deceased person (we knew it from the obituary metadata and the context of the entire obituary). After annotation, we obtained three triplets (Robert, sister, Eva Katherine XX), (Robert, brother, Stanley), and (Robert, brother, Terry XX). The calculating process was as follows:

For each input token $x_i$, we used BERT [22] as a common encoder to obtain each hidden representation $h_{1\text{common}}^i$. Then $h_{1\text{common}}^i$ were sent into one LSTM classifier to obtain each tag $T_i \in T_1$, where $T_i$ was the result set of module 1, and $w_j$ and $b_j$ were parameters for training.

Module 2: Matching Locations to People

After identifying the residence entities (eg, Rochester in Figure 3), we need to match them with specific people. To do so, we used 3 inputs, all extracted names $T_1\text{name} \in T_1$, all extracted residences $T_1\text{residence} \in T_1$, and common representation $h_{1\text{common}}$. This module followed by a co-reference solution [23]. We defined the process as follows:

$$[\text{Robert}, 92, \text{of Rochester, is survived by his sister, [Eva Katherine XX] and two brothers, [Stanley] and [Terry Johnson].}]$$

Module 3: Judging Last Name Distributive

We identified 2 special language patterns in obituaries, last name distributive and names with parentheses, as shown in Table 2. Resolving these language patterns was helpful for extracting and constructing high-quality GKGs. The task of module 3 was to decide for each token in an input sentence if the last name distributive existed by assigning each token with a binary tag of yes OR no. When we cotrained module 3 with other modules, these tags would concatenate with other modules’ tags for joint training. In the sentence in Figure 3, for example, “Stanley” and “Terry” shared the same last name of “Johnson.” Therefore, in module 3, “Stanley” was assigned a label “brother_S yes” and “Terry Johnson” was given 2 tags “brother_B yes” and “brother_E yes.” This way, the module would extract 2 full names, Stanley Johnson and Terry Johnson, instead of Stanley and Terry Johnson. The detailed computing process was as follows:

where $w_j$ and $b_j$ were parameters for training, $T_4i \in T_4$ was the result for each name, and $T_4$ was the result set of module 3.

Module 4: Recognizing Names With Parentheses

Module 4 was a 3-class classifier to determine whether there was a parenthesis in a name, and if so, whether it referred to a previous last name or the name of spouse. The computing process was the same as module 3, which took the input of $h_{1\text{common}}$ and output the tags of 3 classes (“no parenthesis,” “previous last name,” and “spouse’s name”).

Module 5: Rule-Based Inference Layer

This module aimed to infer age, death date, and birth date for the deceased and gender for both the deceased and their family members. First, if an obituary mentioned any 2 attributes out of age, birth date, and death date for the deceased, we calculated the third one. Second, we used both family relationship keyword and name to infer gender. If a family relationship keyword (eg, son, daughter, nephew) suggested gender, we would add the gender tag accordingly. Otherwise, when the family relationship keyword (eg, spouse and parent) did not tell the gender, we used an external human name knowledge base to match the most likely gender with names. For instance, “Tom” and “Emily” indicated male and female, separately.

After constructing the GKGs from each obituary by modules 1 to 5, we assembled the extracted GKGs into bigger ones by matching PII, including people’s names, residence, birth date, death date, and family relationship.
**Joint Training Loss**

We minimized the negative log likelihood loss of the generated tags for the first 4 modules (module 5 is a rule-based inference layer that did not require training). For module $k$ ($k=1, 2, 3, 4$), the loss function was defined as follows:

$$L_k = -\sum_{i=1}^{B} \log p(y^i_k | p^i, I_k)$$

Where $B$ was the batch size, $l_i$ was the length of input sentence sentences, $y^i_k$ and $p^i$ were the true tag and the normalized probability of the predicted tag for an input token $I$, and $w_i$ was a hyperparameter. $P(O)$ was the indicator function that determined which part of equation 10 was used to calculate the loss. If the current tag was not “O” (other), the hyperparameter would decide the weight of the loss function. It was defined as follows:

$$w_i = \begin{cases} 1 & \text{if true tag was not “O”} \\ \frac{1}{P(O)} & \text{otherwise} \end{cases}$$

In the end, we combined all four loss functions $L_1, L_2, L_3$ and $L_4$ together, using different weighting parameters $\lambda_i$ into the final loss function, which was optimized for the entire training as follows:

$$L = \sum_{k=1}^{4} \lambda_k L_k$$

**Evaluation Metrics**

We performed 10-fold cross-validation by randomly selecting 10% of the annotated data for validation and the remaining for training. It is worth noting that the augmented data were only used for training models. Extracted GKGs consists of outputs from modules 1 to 5. They were measured by averaged performance of all modules except module 5 due to this rule-based inference module lacking a gold standard. For modules 1 to 4, we used precision, recall, and F-1 measure for evaluation, which were computed as follows:

$$\text{Precision} = \frac{TP}{TP + FP}$$
$$\text{Recall} = \frac{TP}{TP + FN}$$
$$\text{F-1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

In module 1, the outputs were entity mentions with extra entity and relation types. We defined an extracted mention as true positive instances only if the mention’s boundary, entity type, and relation tags were exactly matched with the gold annotation. The instances of false positive were predicted mentions that do not precisely match with gold annotation boundaries, entity, or relation types. False negative instances were those existing in the gold annotation but not recognized by the model.

In module 2, true positive instances were defined as pairs of name and location that matched exactly. If either name or location was wrong, the pair would be considered a false positive. False negative referred to the name-location pairs missed by our system.

Module 3 and module 4 were formulated as generic classification tasks, so we used common definitions of false negative, false positive, and true positive. For all modules, evaluation metrics were precision, recall, and F-1 measure.

**Results**

Table 3 illustrates the performance of modules 1 to 4 with ablation experiments in terms of macroaveraged and microaveraged precision, recall, and F-1 measure (without data augmentation). A macroaverager is the arithmetic average of the computed metrics for all classes and a microaverage sums up all true positive, true negative, false positive, and false negative instances before computing the final precision, recall, or F-1 measure for all classes. Macroaveraged metrics are often used for evaluation, particularly when there are extremely imbalanced classes, as no single class should largely dominate the results.

As shown in Table 3, we can see that module 1, which included in module 2. This phenomenon was named negative transfer. It meant that although module 2 significantly benefited module 2. It should be noticed that module 2 seemed not helpful in improving the overall performance of each module. For module 1, the macroaveraged and microaveraged F-1 measure dropped by 1.41% (compare the first and third row of the macroaveraged precision) and 1.03% (compare the third and third row of the microaveraged precision) after introducing module 2 into the end-to-end system. Other modules had similar effects when included in module 2. This phenomenon was named negative transfer.

It should be noticed that module 2 seemed not helpful in improving the overall performance of each module. For module 1, the macroaveraged and microaveraged F-1 measure dropped by 1.41% (compare the first and third row of the macroaveraged section of Table 3) and 1.03% (compare the third and third row of the microaveraged section) after introducing module 2 into the end-to-end system. Other modules had similar effects when included in module 2. This phenomenon was named negative transfer. It meant that although module 2 significantly benefited module 2. It should be noticed that module 2 seemed not helpful in improving the overall performance of each module. For module 1, the macroaveraged and microaveraged F-1 measure dropped by 1.41% (compare the first and third row of the macroaveraged precision) and 1.03% (compare the third and third row of the microaveraged precision) after introducing module 2 into the end-to-end system. Other modules had similar effects when included in module 2. This phenomenon was named negative transfer. It meant that although module 2 significantly benefited module 2. It should be noticed that module 2 seemed not helpful in improving the overall performance of each module. For module 1, the macroaveraged and microaveraged F-1 measure dropped by 1.41% (compare the first and third row of the macroaveraged precision) and 1.03% (compare the third and third row of the microaveraged precision).
In our system, the solution for avoiding the negative transfer was that module 1, 3, 4 would be co-trained and module 2 would be separated from the whole system for training. In such a way, each module could benefit the most from the joint training method.

Table 3. Model performance of each module with ablation experiments.

<table>
<thead>
<tr>
<th>Module and ablation test</th>
<th>Macroaveraged performance</th>
<th>Microaveraged performance</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P (%)</td>
<td>R (%)</td>
</tr>
<tr>
<td>Module 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Baseline</td>
<td>81.68</td>
<td>79.93</td>
</tr>
<tr>
<td>Joint training (module 2, 3, &amp; 4) + negative transfer</td>
<td>82.07</td>
<td>81.99</td>
</tr>
<tr>
<td>Joint training (module 3 &amp; 4)</td>
<td>83.85</td>
<td>83.05</td>
</tr>
<tr>
<td>Module 2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Baseline</td>
<td>83.17</td>
<td>68.43</td>
</tr>
<tr>
<td>Joint training (module 1, 3, &amp; 4)</td>
<td>84.27</td>
<td>73.60</td>
</tr>
<tr>
<td>Module 3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Baseline</td>
<td>89.64</td>
<td>92.01</td>
</tr>
<tr>
<td>Joint training (module 1, 2, &amp; 4) + negative transfer</td>
<td>91.48</td>
<td>91.12</td>
</tr>
<tr>
<td>Joint training (module 1 &amp; 4)</td>
<td>92.38</td>
<td>93.01</td>
</tr>
<tr>
<td>Module 4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Baseline</td>
<td>90.65</td>
<td>94.74</td>
</tr>
<tr>
<td>Joint training (module 1, 2, &amp; 3) + negative transfer</td>
<td>92.34</td>
<td>95.76</td>
</tr>
<tr>
<td>Joint training (module 1 &amp; 3)</td>
<td>93.41</td>
<td>96.06</td>
</tr>
</tbody>
</table>

aP: precision.
bR: recall.
cF1: F-1 measure.
dThe microaveraged and macroaveraged performances are the same for module 2 and module 3 because they are both binary classification tasks. All results shown are from the curated corpus without data augmentation.

We also adopted data augmentation technology to expand our corpus, aiming to improve the relation extraction performance for family relations (module 1) with too few training examples. By synonym replacement, random insertion, random swap, and random deletion, we augmented the training data to ensure every relation had no less than 200 training examples. However, the automated data augmentation method introduced new noise. We tested a different augmentation ratio (α) to find the best balance. As shown in Figure 4, when the augmentation ratio was set to 40%, the extra synthetic data in training benefited our model most. It was worth noting that the augmentation data were only used in training for module 1, and we still evaluated our system with real, nonsynthetic test data. Figure 4 shown that the best macroaveraged and microaveraged F-1 measures achieved 89.14% and 95.55%, respectively, for module 1. With augmented module 1, our whole system achieved the best macroaveraged performances, 92.59% (precision), 90.05% (recall), and 91.30% (F-1 measure), and the best microaveraged metrics were 94.79% (precision), 91.45% (recall), and 93.09% (F-1 measure). These results confirmed that data augmentation technology can alleviate the problem of imbalanced data.

After extracting GKGs from all obituaries, we assembled them into bigger ones by matching available PII, including name, gender, age, residence, and birth date. Considering obituaries usually provide detailed PII for the deceased but not for their family members and relatives, we did fuzzy matching for the relatives. That is, if the mentioning of 2 people in 2 different obituaries are likely to refer to the same person based on 1 or more shared piece of PII, we would assemble 2 GKGs into 1. In the end, we had 319 GKGs assembled into 149 bigger GKGs after processing all 12,407 downloaded obituaries. Among those 319 obituaries, 22.3% (71/319) had 1 shared PII item, 8.5% (27/319) had 2, and 69.3% (221/319) had more than 2. We manually evaluated those 149 assembled GKGs and confirmed that 71.8% (107/149) were correct, 12.1% (18/149) were wrong, and 16.1% (24/149) were uncertain. We acknowledge that this rule-based matching method is limitedly useful for the selected geographic location of the Twin Cities area in Minnesota. It might be more error prone to apply to the entire country or other densely populated areas with high population mobility. So we did not include the assembly function in the end-to-end system but kept it as an additional resource for cautious users.
Figure 4. Comparing the F-1 measures of raw corpus and augmented corpus.

Figure 5 shows one example of assembled GKG from 3 obituaries. It contained 4 generations and 30 people. Figure 6 is the corresponding gold standard result conducted from manual validation. It can be seen that the assembled GKG missed the state name Minnesota for Dorothy and Patrick’s residences and one family member, Joe, who was Lynne’s husband (missing parts are shown in dashed boxes). In the original obituary, the sentence mentioning Lynne and Joe’s relation was “...he proposed...they began 54 years of happy life.” and our system failed to capture this subtle language. The successful assembly of multiple obituaries also demonstrated the feasibility of linking family relations extracted from obituaries to EHRs to support genetic research like linkage analysis and disease risk prediction. Meanwhile, it should be noticed that even though obituaries inherently contained rich genealogical information and the system extracts the GKGs with high accuracy, the GKGs should not always be equated to pedigrees used by genealogists. Although it is common to declare blood or nonblood relationships in obituaries due to data specialty (detail analysis for the slippery slope of genealogy issue shown in the Discussion section), we cannot guarantee people always declare the difference of blood or nonblood and always list all of their family members for various reasons.
Discussion

Principal Findings

In this work, we proposed an end-to-end system to construct GKGs from online obituaries, aiming at supplementing EHR data for genetic research. This system achieves microaveraged precision of 94.79%, recall of 91.45%, and F-1 measure of 93.09% after data augmentation technology. The work exploits the large availability of obituaries on the internet, which are consistent with the vital records and census records and more reliable and comprehensive than dependent information from medical insurance and emergency contact in EHR systems [5,6].

We demonstrate an efficient system to automatically build large GKGs from 10 years’ online obituaries in the Twin City area, Minnesota. Furthermore, by identifying individuals, we explore integrating related GKGs into bigger GKGs and manually validating the integrated results. The results show the feasibility of identifying individuals by extracted information, including residence, age, gender, birth, and death dates. We compute similarities between GKGs to further merge them into more complete GKGs. In the future, the similarity computing techniques could assist mapping the GKGs to the EHRs.

In this work, we use publicly available obituaries. The Association of Internet Researchers, in partnership with their Ethical Working Committee, formulated general principals to guide online research [26]. While this document presents the overarching ethical considerations relevant to social media–based research, a comprehensive determination of ethical...
principles and best practices has yet to be developed. Furthermore, debate continues as to whether some forms of social media–based research, namely analysis of existing textual archives (strictly speaking, online obituaries are not social media, but they have similar characteristics as a data source for biomedical research), fall within the parameters of human subject research or constitute an alternative form of humanistic inquiry [27]. Considering the PII embedded in online obituaries, we decided to take a cautious and conservative position in our work by marking up the last name of any real people mentioned in the paper.

As a novel data source, obituaries are informative for constructing family trees. It is hard to obtain such rich genealogical information from other data sources, but there are caveats to their use as genealogical data. First, semantic ambiguity occurs in obituaries as it occurs in many other types of human writing. For example, it is not uncommon to see statements like “...survived by two sons, Marshal and Paul XX and daughter Daisy, and four grandchildren Denny, Gary, Cecil, and Alina.” In this case, it is impossible to tell the exact parents for each of the 4 grandchildren Denny, Gary, Cecil, and Alina. All we know is that their parents are Marshal XX, Paul XX, and Daisy. Additional data sources like birth certificate registries can be helpful in this case.

A second point worth discussing is the slippery slope of genealogy. Compared with medical insurance and emergency contact information [5,6], a statement of nonblood relationship is more common in obituary data due to their specificity. As shown in Figure 1, for child relationship the ratio for nonblood versus blood is 483:5472 (there are 25 mentions of child-in-law, 99 of daughter-in-law, 105 of son-in-law, 151 of stepchild, 48 of stepdaughter, and 55 of stepson compared with 2489 cases of child, 1399 of daughter, and 1584 of son). A similar ratio can be observed in nonblood parent relationship. This advantage could be helpful for alleviating the problem of the slippery slope of genealogy. However, it is still worth mentioning that not all people make such distinctions in obituaries.

In addition, Figure 7 displayed the related statistics aimed at showing potential data bias. We plotted the distribution of age (at death), average number of mentioned family members, and marital status of the deceased for all GKGs extracted from 12,407 downloaded obituaries. As shown in Figure 7, the age distribution of the deceased is consistent with public health data (73.9% of the deceased died at the ages of 70 to 100 years). The average numbers of mentioned family members seem similar for different age groups; only those died in the 0 to 10 and 100 to 110 age groups had relatively smaller family size ($\leq 15$); 87.6% of the GKGs indicated that the deceased was married at least once. We did not interpret the results too deeply because we did not have a good understanding of the sample bias. Meanwhile, it was noticed that people who had complete and/or affluent families tended to publish obituaries. Although these data biases would not affect the performance of our extraction system, the fact that extracted GKGs may be biased should be considered when researchers are using them in other research.

**Figure 7.** Left: distribution of average numbers of mentioned family members. Right: age and marital status of the deceased person in 12,407 extracted genealogical knowledge graphs.

Technically, the data used in the research are very imbalanced, in which 14 rare relationships have fewer than 10 instances. We adopted the augmentation technology to enhance system performance. For example, in the relationships half-sister, grandchild-in-law, and grandson-in-law, their F-1 measures increased from 20.0%, 30.0%, and 35.71% to 66.67%, 50.0%, and 71.43%, respectively. Next step, we plan to experiment with additional few-shot (extremely imbalanced)–based information extraction and mate learning to improve the system [28,29].

In our end-to-end solution, the performance of module 2 was obviously inferior to the other modules. Besides the error propagation problem (module 2 need the results from module 1), the task of module 2 was a semantic matching resolution problem, which is still challenging in the NLP community. In addition, we currently have curated an obituary corpus in English to train the neural network models. To expand to other languages, a new corpus in those specific languages and new gender inference rules would need to be curated. There is some cross-language transfer research in the NLP community which suggests neural models trained on an English corpus can help to build NLP models in other languages by reducing training data and training time. Sometimes such transfers even provide more robust models with better performance [30,31].
In our end-to-end solution, module 2 currently is the bottleneck. This module suffered significantly from negative transfer. Generally speaking, when a task or domain was joined with data of no relatedness or similarity, the added data would become noise rather than useful information. It remains challenging to quantitatively measure the relatedness or similarity among different tasks or domains [32]. Therefore, most transfer learning solutions rely on empirical methods and do not account for negative transfer effects. In this work, we considered module 2, which matched locations to people, as strongly related to other modules that extracted locations or people and paired them. Unfortunately, the experiment results showed negative transfer still occurred. One possible explanation was about the different natures of tasks in modules 1, 2, 3, and 4. Module 2 was a classification task with 2 entity mentions as the input and a class tag as the output. All other modules were sequence tagging tasks, where the whole sentence was the input and tags for all tokens of an input sentence were output. Another possible reason was that the task of module 2 was much more challenging than the others. Modules 1, 3, and 4 all had a higher than 90% microaveraged F-1 measure when we tested them individually, while module 2 had a 75.08% microaveraged F-1 measure. In addition, module 2 needed inputs from module 1. The errors of module 1 would propagate to module 2. How to improve module 2 and alleviate its negative transfer and error propagation is what we plan to focus on methodologically in the future.

Besides the performance benefits shown in the Result section, the multitask solution is also faster to train. We use a single V100 GPU in this study. For the traditional pipeline model, one round 10-fold cross-validation experiment costs about 240 hours in total. However, the multitask model with all 4 modules together takes only 150 hours. For module 1, the training process took about 70 epochs to achieve an F-1 measure of 80% when being trained independently. The multitask method takes less than 5 epochs to achieve the same level of F-1 measure.

Limitations
The first limitation of our work is the existing potential data bias. Our data are collected from online obituary websites. In such conditions, people who had intact and/or affluent families tended to publish obituaries. The second limitation is that our system is mainly for English obituaries. Modules 2 and 3 are designed for 2 English writing patterns.

Conclusions
GKGs have great potential to enhance many medical research fields, especially combined with EHR data. We believe a high-quality, large-scale genealogical information database will have significant research meaning. In this work, we presented a new corpus with a predefined family relationship map and augmented training data and proposed a multitask deep neural system to construct and assemble GKGs. With the data augmentation technology, the system achieved microaveraged precision, recall, and F-1 measure of 94.79%, 91.45%, and 93.09%, respectively, and macroaveraged precision, recall, and F-1 measure of 92.59%, 90.05%, 91.30%, respectively. Based on such promising results, we developed PII-matching rules to assemble large GKGs, demonstrating the potential of linking GKGs to EHRs. The system is capable of generating a large number of GKGs to support related research, like genetic research, linkage analysis, and disease risk prediction. We share the source codes and system with the entire scientific community on GitHub, without the corpus for privacy protection [33].

In the future, we will improve the performance of our system further and match GKGs with more medical information, like EHR databases. With the massive obituary data freely available on the internet or other textual data that contain genealogical information, our ultimate goal is to accelerate large-scale disease heritability research and clinical genetics research.

Acknowledgments
This work has been supported by grant 2018YFC0910404 from the National Key Research and Development Program of China, grant 61772409 from the National Natural Science Foundation of China; grant 61721002 from the Innovative Research Group of the National Natural Science Foundation of China, and grant IRT_17R86 from the Innovation Research Team of the Ministry of Education, Project of China Knowledge Centre for Engineering Science and Technology.

Conflicts of Interest
None declared.

References


33. Genealogical knowledge graph. URL: https://github.com/KaiHe-better/Genealogical-Knowledge-Graph [accessed 2021-07-19]

Abbreviations

- **EHR**: electronic health records
- **GKG**: genealogical knowledge graph
- **LSTM**: long short-term memory
- **NER**: name entity recognition
- **NLP**: natural language processing
- **PII**: personally identifiable information
- **RC**: relation classification
Evaluating Epidemiological Risk by Using Open Contact Tracing Data: Correlational Study

Stefano Piotto\textsuperscript{1,2}, PhD; Luigi Di Biasi\textsuperscript{3}, MSc; Francesco Marraffino\textsuperscript{1}, PhD; Simona Concilio\textsuperscript{1,2}, PhD

\textsuperscript{1}Department of Pharmacy, University of Salerno, Fisciano, Italy
\textsuperscript{2}Bionam Research Center for Biomaterials, University of Salerno, Fisciano, Italy
\textsuperscript{3}Department of Computer Sciences, University of Salerno, Fisciano, Italy

Abstract

Background: During the 2020s, there has been extensive debate about the possibility of using contact tracing (CT) to contain the SARS-CoV-2 pandemic, and concerns have been raised about data security and privacy. Little has been said about the effectiveness of CT. In this paper, we present a real data analysis of a CT experiment that was conducted in Italy for 8 months and involved more than 100,000 CT app users.

Objective: We aimed to discuss the technical and health aspects of using a centralized approach. We also aimed to show the correlation between the acquired contact data and the number of SARS-CoV-2–positive cases. Finally, we aimed to analyze CT data to define population behaviors and show the potential applications of real CT data.

Methods: We collected, analyzed, and evaluated CT data on the duration, persistence, and frequency of contacts over several months of observation. A statistical test was conducted to determine whether there was a correlation between indices of behavior that were calculated from the data and the number of new SARS-CoV-2 infections in the population (new SARS-CoV-2–positive cases).

Results: We found evidence of a correlation between a weighted measure of contacts and the number of new SARS-CoV-2–positive cases (Pearson coefficient=0.86), thereby paving the road to better and more accurate data analyses and spread predictions.

Conclusions: Our data have been used to determine the most relevant epidemiological parameters and can be used to develop an agent-based system for simulating the effects of restrictions and vaccinations. Further, we demonstrated our system's ability to identify the physical locations where the probability of infection is the highest. All the data we collected are available to the scientific community for further analysis.

Keywords

SARS-CoV-2; COVID-19; contact tracing; Bluetooth Low Energy; transmission dynamics; infection spread; mobile apps; mHealth; digital apps; mobile phone

Introduction

In China, during December 2019, SARS-CoV-2 was identified as a novel beta coronavirus. At the time of writing this paper (December 2020), SARS-CoV-2 has caused almost 60 million confirmed human infections worldwide and more than 1 million deaths since its discovery \cite{1,2}. The disease caused by SARS-CoV-2 is called COVID-19, and the disease was declared a global pandemic on March 11, 2020 \cite{3}. Containment measures are the first and most crucial step for rapidly halting an outbreak that could otherwise become an epidemic or even turn into a pandemic, such as the COVID-19 outbreak \cite{4}.
Notable examples of disease epidemics with a high occurrence of superspreading events (SSEs) are the SARS-CoV (severe acute respiratory syndrome coronavirus; 2002-2003) and MERS-CoV (Middle East respiratory syndrome coronavirus; since 2013) epidemics [5-9]. The basic reproduction number ($R_0$) is a key measure of transmissibility. It is defined as the number of infected contacts that 1 infected individual generates on average during their infectious period. An $R_0$ value of >1 means that a virus will continue its propagation among susceptible hosts. In contrast, an $R_0$ of <1 means that it is certain that epidemic spread will stop [10,11]. The SARS-CoV and MERS-CoV have an $R_0$ of around 3 [12]. For SARS-CoV-2, the estimated $R_0$ ranges between 2 and 3 [9,13]. However, it is unknown as to what extent SSEs are involved in the spread of SARS-CoV-2 infection.

Lockdown was the most widespread pandemic containment response, and it was introduced at different levels by most affected countries. As already predicted by mathematical models [14] and proven by trends that were updated at the time of writing this paper, the contagion's spread resumed rapidly when lockdown countermeasures were lifted. Rapid and automatic contact tracing (CT) is an essential intervention for contagion containment [15-19]; however, user localization poses a privacy risk and reduces compliance rates [20]. According to the World Health Organization, CT involves the following three steps: the identification of a contact (identifying those that a confirmed positive patient had contact with based on the transmission modalities of the pathogen of interest), the listing of contacts (keeping a record of individuals who possibly had contact with infected patients and informing these individuals), and contact follow-up [21]. CT has a dual purpose—treating people who have possibly been exposed to infectious diseases and stopping the transmission chain to contain an epidemic. Due to the prevalence of smartphones, CT has the potential to become a powerful intervention; the vast majority of smartphone users carry their smartphone devices with them throughout the day, and smartphones can generate detailed GPS location information. However, due to the availability of users' location data, there is growing concern about the infringement of an individual's right to privacy. An alternative is using other contact monitoring technologies that are based on proximity assessments rather than those based on location information [22]. It is important to note that this study does not constitute an endorsement or rejection of CT based on potential data security risks or privacy limitations. This study intends to assess whether and to what extent the acquisition of contact data helps with assessing the spread of SARS-CoV-2.

Technologies such as Bluetooth Low Energy allow for the evaluation of the distance between users without locating them and thus help with addressing the privacy issue. The number of CT apps that have been introduced since the beginning of the SARS-CoV-2 pandemic is considerable [23,24] and reflects governments' interest in automating the tracing of people who have had recent contact with individuals who tested positive for COVID-19. An app that uses a centralized approach was developed by the academic spin-off company of the University of Salerno—SoftMining (SM). The app [25] was supported by government agencies such as the Campania Region and was validated by more than 120,000 users; the app had peaks of more than 15,000 active daily users.

CT is a fundamental intervention for acquiring population data, which show how different population groups can behave differently. Such behaviors result in different risks of infection among group members. In Multimedia Appendix 1, we describe how CT data were acquired via the Bluetooth Low Energy technology of the SM-COVID-19 app and how data were clustered to obtain different mobility and behavior groups. In this paper, we discuss how we used Italian National Institute of Health data on contagion trends in Italy [26] to estimate a more precise number of SARS-CoV-2—positive cases that was less influenced by the number of tests performed on the population. In addition, we show the link between the acquired CT data and the number of new SARS-CoV-2—positive cases. This allowed us to define an epidemiological risk function that was based on the number of, frequency of, and distance between contacts. The risk function expresses the probability that an individual will become ill as a function of their age within a given period of time. This study aims to evaluate whether the use of CT can support the containment of an epidemic. The data acquired from CT were analyzed and correlated with data on the progression of SARS-CoV-2 infection.

This study was not conducted for commercial purposes; it was conducted for the purposes of academic research and aims to make CT data available to the scientific community for future research.

**Methods**

**CT Data Acquisition**

During the CT phase, the SM-COVID-19 app analyzed the environment and, at regular intervals, sent data on the duration of a contact and the instantaneous and average distances (over the time) of a contact to the server. App users could voluntarily decide to share location data as well. If they did, the server also received latitude, longitude, precision, and smartphone provider data. We provide the full description of the data acquisition procedures in Multimedia Appendices 1 and 2. The developed technologies allowed for high precision in distance calculations (less than 0.5 m under optimal conditions and after device calibration) and were implemented via the SM-COVID-19 app, which is available on Android and iOS smartphones (via TestFlight; Apple Inc). Daily data were anonymized and saved for further use, as described in Multimedia Appendix 3, in accordance with the General Data Protection Regulation. Anonymity was also guaranteed when the GPS localization function was enabled, as data were stored randomly in the database; the database did not present an individual user's location in a precise way. The app only used random 128-bit proximity IDs, and only the user's device kept track of the device IDs. The app’s functions were conducted and maintained with a back-end server, on which arbitrary identifiers were stored. Users could not be identified directly with app data, as only the app's random identifiers were stored on the server.
Social Mobility Analysis

The data set obtained from the SM-COVID-19 app in the period of April to November 2020 was analyzed. The data set's structure is described in Multimedia Appendix 2. Reported data from August 1 to August 30, 2020, were obtained to analyze mobility data from a period when no lockdown measures were in place. Such data are useful for tracking movements in real situations. We removed users with less than 15 days of activity from our analysis to exclude users who may have deactivated the app. The cleaned data set was clustered. Before the clustering process, the t-distributed stochastic neighbor embedding machine learning algorithm was applied to the data set to reduce its dimensionality to 2. The clustering was carried out by using the Ward linkage method. This method allows the user to select the number of clusters arbitrarily. We analyzed the distribution of data for different numbers of clusters (2-10 clusters); the optimal distribution was obtained with 5 clusters. The average number of daily contacts and the SDs for the clusters are reported in Table 1. SDs were high, since every cluster had many users with 0-contact days among those with low- and high-contact days. As shown in Table 1, the population was divided into clusters of approximately the same size. However, cluster 5 was larger and included users who had a larger number of contacts. This cluster accounted for the population with the highest number of contacts and included users with the highest number of contacts and the highest mobility.

Table 1. The cluster data of active users for the period of August 1 to August 30, 2020.

<table>
<thead>
<tr>
<th>Cluster number</th>
<th>Number of daily contacts based on Bluetooth Low Energy technology, mean (SD)a</th>
<th>Percentage of active users</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>23.40 (38.55)</td>
<td>14</td>
</tr>
<tr>
<td>2</td>
<td>12.05 (22.62)</td>
<td>19</td>
</tr>
<tr>
<td>3</td>
<td>41.95 (75.79)</td>
<td>17</td>
</tr>
<tr>
<td>4</td>
<td>69.91 (103.76)</td>
<td>20</td>
</tr>
<tr>
<td>5</td>
<td>121.48 (145.05)</td>
<td>30</td>
</tr>
</tbody>
</table>

aThe average number of daily contacts for each cluster and SDs were calculated based on all cluster data (ie, from days 1 to 30).

Data Availability

All data can be made available upon request from the authors or the SM-COVID-19 team [27].

Results

Statistical Analysis and Estimates of the Real Number of SARS-CoV-2–Positive Cases

For our statistical analysis, we relied on official data on the daily SARS-CoV-2–related trends in Italy, which were released by the Italian National Institute of Health and aggregated by the Department of Civil Protection of the Presidency of the Council of Ministers [17]. We estimated the possible number of real infections that may have occurred during the epidemic in Italy. We obtained the daily number of newly performed tests based on the total number of tests performed. This was calculated by using equation 1 in Multimedia Appendix 4. The method for estimating the number of new daily SARS-CoV-2–positive cases is detailed in Multimedia Appendix 4. We performed data smoothing via sliding-window averaging to reduce each day’s variability, which was the result of the cumulative regional data's intrinsic variability. The SARS-CoV-2–related trends over a given period were roughly linear; there were no sudden peaks. Additionally, the averaging process performed allowed us to smoothen the curves, which were in line with these trends. Equation 2 in Multimedia Appendix 4 was used to define the ratio between the number of daily tests and the number of daily reported SARS-CoV-2–positive cases. The estimated number of new SARS-CoV-2–positive cases (EP[k]) for each day was calculated with equation 3 in Multimedia Appendix 4. With our method, we estimated a correction for the number of real SARS-CoV-2–positive cases that occurred during the pandemic period. We observed that around 224,000 cases were not diagnosed, and of these cases, nearly around 81,000 were missed in the period of March to May 2020. The difference between the official number of cases and the estimated number of new SARS-CoV-2–positive cases is shown in Figure 1.

https://www.jmir.org/2021/8/e28947
Correlation Between CT Data and Contagion Trends

The correct number of daily SARS-CoV-2–positive cases was calculated to perform correlation analyses with the data obtained from CT. The data distributed by the ISS, due to how the data were structured, showed considerable fluctuations based on the number of tests performed. It was also possible to observe a weekly trend in the number of SARS-CoV-2–positive cases recorded due to the reduced number of tests performed during weekends. Such data therefore presented fluctuations that could alter the analysis. Data smoothing via sliding-window averaging also provided an additional element for alleviating the issue with fluctuations.

We then examined whether the contact index (CI) and the alpha index ($\alpha$) correlated with the number of daily new SARS-CoV-2–positive cases. These two parameters are indices of effective contacts and account for the distance between two users who come into contact with each other and the contact’s duration. These parameters and the related equations are described in detail in Multimedia Appendix 5 [4,28,29]. These parameters were necessary, since not all of the contacts recorded by the app involved people who could effectively transmit the virus. CI$_k$ is a value that indicates a user’s risk of infection on day $k$ based on the number of effective contacts that the user had on the same day. CI$_k$ was calculated with equation 4 in Multimedia Appendix 5 [4,28,29]. $\alpha_k$ is a risk index, and it is based on data from the previous $k-14$ days (excluding day $k$). $\alpha_k$ reflects a user’s behavior. The optimization of these parameters will be the subject of future studies.

The SM-COVID-19 data set lists the CI and $\alpha$ values for each day and every user. Therefore, to evaluate daily trends, we calculated the total CI and $\alpha$ values for each day ($k$) by summing each individual users' values. As such, it was possible to evaluate the trends for CI and $\alpha$ values and exclude users who deactivated the app for a given period. The values were smoothed by using a sliding window of 7 days. In Figure 2, we show the temporal evolution of CI values over 160 days. For visualization, in Figure 2, we report the logarithm of the number of new SARS-CoV-2–positive cases. There is an evident, rough correlation between the CI and the number of new SARS-CoV-2–positive cases. For each CI$_k$ and $\alpha_k$ value, we calculated the Pearson correlation coefficient based on the estimated number of SARS-CoV-2–positive cases to assess how the number of contacts varied before and after a confirmation of COVID-19 positivity. It was very interesting to note that the correlation coefficient for CI$_k$ reached its maximum at $k+7$ days. The high correlations observed in the subsequent days correlated with SARS-CoV-2 incubation times, and COVID-19 positivity occurred in the days following an effective contact. The $\alpha_k$ value reached its maximum at $k+5$ days. The differences between the $\alpha$ and CI values’ correlation coefficients (ie, their correlation with the number of new SARS-CoV-2–positive cases) were attributable to the different calculation methods that were used for the two parameters, as the $\alpha$ value accounts for the risk of infection in the 14 days before day $k$. The correlation between CI values and the number of new SARS-CoV-2–positive cases is shown in Figure 3. We reported the correlation data that corresponded to the period of June to October 2020 because of the high availability of more consistent CT data. This correlation was also monitored for the previous studied period (March to May 2020) to confirm that the obtained values were not the result of artifacts or autocorrelations.
Discussion

The analysis of the collected data allowed us to determine the aspects of CT that are essential for the evaluation of the progression of the SARS-CoV-2 pandemic. These essential aspects were identified via the estimation of the real number of new SARS-CoV-2–positive cases and the correlation of the number and frequency of contacts with the probability of infection.

Estimation of the Total Number of People Who Tested Positive for SARS-CoV-2

At the beginning of the pandemic in Italy, during the period of March to May 2020, the substantial underestimation of the total number of people who tested positive for SARS-CoV-2 in Italy...
was a likely scenario. This was undoubtedly due to the reduced number of tests that were performed during the first phase of the SARS-CoV-2 pandemic and the lack of an adequate response for tracing infections. One method for estimating a realistic number of SARS-CoV-2–positive cases is to use the ratio between the number of tests carried out and the number of SARS-CoV-2–positive cases detected every day. We chose this ratio because as the number of tests carried out increases, this number eventually plateaus. These data are collected throughout the country and are therefore subject to regional and local variability. It has been assumed that the ratio between the number of positive cases and the number of tests performed varies slowly over time in the absence of hospitalization problems. This ratio has been used to estimate the actual number of SARS-CoV-2–positive cases, which is always greater than or equal to the official number of cases. As shown in Figure 1, the difference between the official number of daily new SARS-CoV-2–positive cases and the estimated number of cases was higher during the initial phases of the pandemic (ie, during the period of March to May 2020). During this period, according to our analysis, at least 81,000 patients with SARS-CoV-2 infection were not diagnosed with COVID-19. As already mentioned, calculating the real number of new SARS-CoV-2–positive cases was necessary because the data provided by the Istituto Superiore di Sanità (Italian Superior Institute of Health) varied according to the number of tests performed each day. In the initial stages of the pandemic, the number of tests was remarkably low due to the lack of adequate diagnostic tools.

**Ethical and Practical Issues of CT Apps**

CT apps have generated much discussion, particularly discussions regarding privacy and such apps’ susceptibility to attacks. Considerations of data security and possible privacy violations are certainly essential elements and have resulted in the creation of numerous solutions that have been adopted at the national level. This paper does not aim not to take a position on the security and privacy of CT apps, although the developers of SM-COVID-19 have considered these aspects. Rather, we are concerned with assessing whether CT apps, that is, those that can be developed based on currently available technology, can impact communities’ health. Several apps have been adopted at a national level by multiple countries. However, during our research, we did not find any information on the availability of data collected by these apps. CT data provide useful information on various aspects of the SARS-CoV-2 pandemic (eg, the pandemic course) and the behavior and mobility of app users, thereby allowing researchers to map the frequency of contacts and identify high-risk areas. Our CT data set allowed us to analyze data and identify different classes of behavior among the population.

The SM-COVID-19 app uses a centralized model [23,24]. However, despite using a centralized model, users’ privacy is completely protected via anonymization, as per the General Data Protection Regulation. The advantage of using a centralized model is that data stored on the server can be anonymized via aggregation and used by public authorities as a source of important aggregate information about the number of contacts in the population, the app’s effectiveness in tracing and alerting contacts, and the aggregate number of people who could potentially develop symptoms. Unlike a decentralized model, a centralized model provides access to CT data, thereby making these data available for analysis and the improvement of epidemiological models. As already stated by Ferretti et al [19], the control of the SARS-CoV-2 epidemic via manual CT is impossible, as CT introduces a time lag resulting from the need to notify individuals about having contact with infected individuals. Such lag exacerbates the spread the infection, which is already remarkable given the infectivity of SARS-CoV-2 and the high percentage of transmission by presymptomatic individuals. The use of this app model, in which individuals are immediately notified about having contact with people who tested positive for SARS-CoV-2, would be sufficient for stopping the epidemic if the app is used by an adequate number of people [30] and would provide valuable data for creating accurate and valid predictive and epidemiological models. The choice of using a centralized model allows for the reconstruction of the chains of contagion transmission and the rapid propagation of risk indices (calculated with mathematical models)—operations that are difficult to implement when tracing data are only kept on devices.

By using data from August 2020, during which no lockdown measures or restrictions on mobility were in place and only partial restrictions were placed on gatherings, it was possible to identify 5 different behavior classes (or mobility classes). Table 1 shows the data from the clustering process. The five groups had approximately the same population size except for cluster 5, which had the largest number of people and included individuals with the highest mobility. The high amount of deviation in cluster 5 shows how users in this class alternated between experiencing days with 0 contacts (ie, no mobility; eg, days when they could be working from home) and experiencing days with a very high number of contacts (eg, due to a commute or due to work involving contact with the public). From these clusters, it is impossible to define the reasons behind a given number of contacts, but this is irrelevant as long as similar behaviors are present among the users belonging to a certain cluster. However, this clustering process provided interesting insights; it showed that there are classes of people with very low mobility (eg, older people) and classes of people with high mobility who experience a high number of contacts (eg, working in a hospital, supermarket, etc). This information can be even more useful when using a localized approach, such as using GPS data, as such data would help with providing more appropriate definitions for categories. The contacts registered by the app allowed us to trace the frequency of contacts and the trend in the number of contacts for a given period, a single user, a cluster, or the whole data set.

**Correlation Between CT and the Total Number of New SARS-CoV-2–Positive Cases**

CT data correlated with the growth in the number of new SARS-CoV-2–positive cases, and the highest correlation was observed 5 to 7 days after day k. This observation is in line with the hypothesis that an increase in the number of contacts is linked to an increased risk of infection. The most interesting element of the correlation is the time gap. The differences in the correlation values were probably related to the incubation
period of SARS-CoV-2. Consequently, a contact that occurs on day k will not result in COVID-19 positivity on day k but on day k+n. This time gap is in line with the estimated incubation time for SARS-CoV-2 [4,28], and our analysis shows the effectiveness of using CT data to predict the number of new SARS-CoV-2–positive cases. This high correlation means that CT data can be used to develop new and more accurate epidemiological models and predictive tools.

Although a distributed approach that involves the use of a central advertising server makes it possible to alert individuals in direct contacts (the first contact between a newly infected individual and another person) about an eventual infection, flooding operations are necessary on CT networks to warn individuals about contacts of level 2 or higher. The decentralized model provides only 1 degree of separation from a CT app user who tested positive for COVID-19 (user A). To obtain data on a longer chain of contacts, which would have a decreasing risk gradient, it would be necessary for user B (a user in user A’s contact chain) to publish their identifier so that user C (a user who had contact with user B but not with user A) is alerted. This could prove particularly dangerous when an asymptomatic or low-symptomatic individual who has not been tested for SARS-CoV-2 infection could infect another person and even cause another person’s death. [31] In such a situation, decentralized CT would fail. On the other hand, the centralized model allows for the instant tracing of all contacts, regardless of the degree of separation. This would result in the more effective containment of the contagion, since all individuals in a contact chain that are deemed to be at risk for infection would be notified immediately about the danger. In this model, voluntary data input by individuals involved in first-degree contacts for informing those involved in second-degree contacts would not be required whenever the former was notified about having contact with a person who tested positive for COVID-19. Similar conclusions were reached by Aleta et al [30], who proved the effectiveness of using an automatic and extensive CT system to contain the spread of SARS-CoV-2 when lockdown measures are lifted. The work of Aleta et al [30] confirmed the usefulness of CT data collected from the population and provided an excellent basis for improving predictions and reducing the social and economic impact of SARS-CoV-2 prior to the effective vaccination of the entire population. At the time of writing this paper, we did not find any other available data sets with real CT data.

**Geolocation**

CT data can be beneficial for evaluating SARS-CoV-2 propagation data. The data set that was made available by the app is particularly interesting because, due to its structure, it can be used as the basis for tracing SSEs. SSEs are generally defined as outbreaks in which a small number of individuals infect a large number of secondary individuals (ie, well-above the expected average number of individuals) [32]. The CT data that allowed us to define behavioral clusters for the population can also help with determining the SARS-CoV-2 pandemic’s potential for generating SSEs. Although lower than those of the SARS-CoV and MERS-CoV pandemics, the SARS-CoV-2 pandemic’s potential for generating SSEs is significant. In the absence of interventions such as social distancing, this potential would be even more significant. When developing disease control measures, people should focus on the rapid CT and quarantining of infected individuals and policies for physical distancing or targeted shutdowns to prevent the occurrence of SSEs. Having the ability to predict a pandemic’s potential for generating SSEs would be vital in preventing outbreaks, and it would considerably reduce a contagion’s overall $R_0$ value. The use of GPS data that are made anonymous with an appropriate protocol would enable researchers to use a rapid localized approach to significantly reducing the risk of contagion spread in certain areas and act in a targeted and localized manner. This type of information can prove very useful for planning the possible containment of a contagion in defined areas. The tests we performed that used GPS data showed the potential of this approach. For these tests, CT data that were acquired during the lockdown period (April 14 to May 3, 2020) from SM-COVID-19 users who had explicitly activated GPS tracing and whose GPS coordinates included the Campania Region were used (Figure 4). The simulations showed that a higher number of alerts were generated in locations that corresponded to the outbreaks that occurred during the lockdown (Figure 5). This type of voluntarily provided information can be a handy tool for confining and preventing contagion spread.
Conclusions

The high correlation between CT data and the number of recorded SARS-CoV-2–positive cases (with a delay of 5-7 days) was remarkable. The number of registered contacts and the number of new SARS-CoV-2–positive cases showed the same weekly trend fluctuations, which not only depended on the number of tests but also on the different mobility abilities of people. Moreover, there was a time lag between the two factors, and this was the result of the incubation time of SARS-CoV-2. This time lag can be used to estimate the real incubation time of SARS-CoV-2. Further, this correlation can be extremely useful for defining and predicting infection trends and can be used to improve predictive models that only use health authorities’ data. Regardless of the effectiveness of CT, the collected data provided a powerful tool for improving predictive and epidemiological models and could be integrated...
into different types of analyses to improve the accuracy and efficiency of predictions based on real data.

This study lays the foundation for our upcoming papers. In future papers, we will show how CT data were implemented in a CT simulator to turn it into a real data-based contagion spread simulator, which provided us with data on the mobility of the different clusters that were defined in this study. The agents' mobility data will be used to determine the risk of infection, identify epidemiological parameters, and simulate the spread of SARS-CoV-2 in different contexts. The SM-COVID-19 data set is open and free for use by the scientific community. This paper does not represent a policy pronouncement, as this would not be a scientific objective. We believe that our study may prompt informed discussions of the possible risks and likely benefits of our approach to using CT data. For these reasons, all collected data are available for further analysis.

Acknowledgments

The authors thank the Campania Region for supporting our scientific research and the development of the app. We express our gratitude to Cesare Pianese and Luca Canepa for the numerous opportunities to discuss ethical and privacy issues. We also thank the entire SM-COVID-19 team [33] for supporting and distributing the data used in this work.

Conflicts of Interest

SP and LDB are members of the academic spin-off company SM, and they were involved in the development of the SM-COVID-19 app.

Multimedia Appendix 1

Acquisition of SM-COVID-19 app data on contacts.

[DOCX File, 41 KB - jmir_v23i8e28947_app1.docx]

Multimedia Appendix 2

Open data format.

[DOCX File, 15 KB - jmir_v23i8e28947_app2.docx]

Multimedia Appendix 3

Dumping and anonymization.

[DOCX File, 15 KB - jmir_v23i8e28947_app3.docx]

Multimedia Appendix 4

Statistical analysis and estimates of the real number of SARS-CoV-2–positive cases.

[DOCX File, 14 KB - jmir_v23i8e28947_app4.docx]

Multimedia Appendix 5

Contact index and alpha values.

[DOCX File, 20 KB - jmir_v23i8e28947_app5.docx]

References


Abbreviations

CI: contact index  
CT: contact tracing  
MERS-CoV: Middle East respiratory syndrome coronavirus  
SARS-CoV: severe acute respiratory syndrome coronavirus  
SM: SoftMining  
SSE: superspreading event
Association Between Public Opinion and Malaysian Government Communication Strategies About the COVID-19 Crisis: Content Analysis of Image Repair Strategies in Social Media

Nasaai Masngut¹*, MSocSci; Emma Mohamad²,³*, PhD

¹Malaysian Nuclear Agency, Bangi, Malaysia
²Centre for Research in Media and Communication, Faculty of Social Sciences and Humanities, Universiti Kebangsaan Malaysia, Bangi, Malaysia
³UKM x UNICEF Communication for Development Centre in Health, Faculty of Social Sciences and Humanities, Universiti Kebangsaan Malaysia, Bangi, Malaysia

*all authors contributed equally

Corresponding Author:
Emma Mohamad, PhD
Centre for Research in Media and Communication
Faculty of Social Sciences and Humanities
Universiti Kebangsaan Malaysia
Jalan Nik Ahmed Kamil
Bangi, 43600
Malaysia
Phone: 60 389215456
Email: emmamohamad@ukm.edu.my

Abstract

Background: The COVID-19 health crisis has posed an unprecedented challenge for governments worldwide to manage and communicate about the pandemic effectively, while maintaining public trust. Good leadership image in times of a health emergency is paramount to ensure public confidence in governments’ abilities to manage the crisis.

Objective: The aim of this study was to identify types of image repair strategies utilized by the Malaysian government in their communication about COVID-19 in the media and analyze public responses to these messages on social media.

Methods: Content analysis was employed to analyze 120 media statements and 382 comments retrieved from Facebook pages of 2 mainstream newspapers—Berita Harian and The Star. These media statements and comments were collected within a span of 6 weeks prior to and during the first implementation of Movement Control Order by the Malaysian Government. The media statements were analyzed according to Image Repair Theory to categorize strategies employed in government communications related to COVID-19 crisis. Public opinion responses were measured using modified lexicon-based sentiment analysis to categorize positive, negative, and neutral statements.

Results: The Malaysian government employed all 5 Image Repair Theory strategies in their communications in both newspapers. The strategy most utilized was reducing offensiveness (75/120, 62.5%), followed by corrective action (30/120, 25.0%), evading responsibilities (10/120, 8.3%), denial (4/120, 3.3%), and mortification (1/120, 0.8%). This study also found multiple substrategies in government media statements including denial, shifting blame, provocation, defeasibility, accident, good intention, bolstering, minimization, differentiation, transcendence, attacking accuser, resolve problem, prevent recurrence, admit wrongdoing, and apologize. This study also found that 64.7% of public opinion was positive in response to media statements made by the Malaysian government and also revealed a significant positive association (P=.04) between image repair strategies utilized by the Malaysian government and public opinion.

Conclusions: Communication in the media may assist the government in fostering positive support from the public. Suitable image repair strategies could garner positive public responses and help build trust in times of crisis.

(J Med Internet Res 2021;23(8):e28074) doi:10.2196/28074
KEYWORDS
COVID-19; crisis; health communication; image repair; Malaysian government; sentiment; communication; content analysis; public opinion; social media; strategy

Introduction

Background

A public health crisis is often a threat to health infrastructure because it can cripple an existing health care system on a larger scale. As such, a global pandemic such as the COVID-19 pandemic, could potentially be damaging to a national health system [1,2] and the stability of a country. Much of the success of a public health campaign is determined by how far the public trust their government and health institution. During the 2018 Ebola outbreak in the Democratic Republic of Congo, the federal government faced a critical challenge in disseminating relevant information, obtaining local cooperation, and conducting mass inoculation programs. This was due to a weak social system and diminished public trust. The civil unrest and domestic terrorism in the Democratic Republic of Congo made it more difficult for the government to combat the Ebola virus [3].

In Malaysia, the sudden outbreak of severe acute respiratory syndrome (SARS) in 2003 initially proved too much for the government to handle. Constant communication on SARS in the mass media led to a mild panic among the public [4]. However, the Ministry of Health played a vital role in mitigating the situation by providing constant situational reports and mobilizing a crisis management team. The Ministry also solidified their role as the main source of information to the public through daily updates on SARS and active public engagements [4]. Communication from health authorities is important to ensure that the public take necessary actions to control the spread of infectious diseases [5,6] by framing issues that are deemed critical for public’s safety [7].

While keeping the public informed remains the main goal of crisis communication, earning public trust is another matter. Gaining public trust is often achieved by using credible means [8], which can effectively improve message reception. However, social media has become an important source of information regardless of its authenticity and reliability [9]. The spread of misinformation has proved to be dangerous in times of crisis because it can easily undermine health efforts and sow the seeds of doubt among the public. Earning trust is vital to allow the government to implement necessary measures and obtain cooperation from the public.

COVID-19 was first detected in Malaysia in January 2020. Since then, the Malaysian Government through the Ministry of Health has been consistent in delivering daily updates on COVID-19 to the public. As part of their integrated effort in combatting COVID-19, the Malaysian government has activated its Crisis Preparedness and Response Center, which is a part of an overall strategy to overcome the pandemic. This center acted as a central command for early outbreak coordination and management control over the pandemic [10].

Even though the Malaysian government was generally commended for their swift action during the initial stage of local COVID-19 transmission, there were several challenges to secure public’s trust at the time. This was largely due to an unexpected political upheaval after the Prime Minister’s resignation in early March 2020. The sudden government transition itself puts Malaysia in a fragile state because of uncertainty about government stability and the capability of the newly formed government to combat COVID-19 [11]. In addition, a premature statement made by the Health Minister, suggesting that warm water might kill COVID-19 virus, was ridiculed, which did not help to foster public trust [12]. This controversial statement was later criticized by many medical practitioners who refuted the claim. In another occasion, the minister was teased for his lack of general knowledge when he accidentally claimed that 500 countries had participated in a recent World Health Organization video conference [13]. Controversial statements such as these could damage government’s reputation and jeopardize public’s trust in the government’s ability to manage the pandemic. Furthermore, several statements from the government were found to be contradictory [14], which led to escalating confusion among the public. Even though providing contradictory information in early phases of a crisis is not uncommon [15], such inconsistency may lead to further distrust among the public.

Therefore, we examined how the Malaysian government managed crisis communications on COVID-19 in the media. Specifically, this study identifies image repair strategies utilized by the Malaysian government in the media and examines how society react to these strategies by observing the directions of public sentiments through social media responses. Findings from this research could be useful in planning better crisis communication strategies.

Measures to Mitigate COVID-19

Daily updates on COVID-19 were given by the Director General of Health, while the Minister of Defense was in charge of communicating standard operating procedures and public measures to mitigate the spread of the virus. These updates were given through separate press conferences and published in different forms of mass media, including Telegram, a freeware instant messaging app that could reach millions of people rapidly. In addition, to ensure information on COVID-19 was disseminated through proper channels, a dedicated website was set up under the purview of Ministry of Health [16]. This website is a centralized platform for daily updates on cases as well as any new developments on COVID-19, in addition to the daily press conferences. To step up their effort in curbing COVID-19, the Malaysian government enacted a Movement Control Order on March 18, 2020 for a period of 2 weeks, then extended the order by 5 weeks (for a total 7 weeks). The Movement Control Order was executed under the purview of the National Security Council under the Infectious Disease Control and Prevention Act 1988, as well as the Police Act 1967 [17]. This special measure allowed the government to enforce various movement restrictions on general population, in order to reduce virus transmission in the population.
Key economic areas and activities came to halt, along with education sector, religious houses, and sports and cultural activities. However, movement restriction exceptions were given for essential services. The implementation of Movement Control Order later transitioned to a Conditional Movement Control Order, which eased economic restriction while maintaining movement restriction. When the local case transmission showed improvement, the government introduced a Recovery Movement Control Order, which allowed gradual reopening of most key sectors. Systematic implementation of the different levels of Movement Control Order demonstrated the Malaysian government’s commitment in tackling COVID-19 nationwide. The Malaysian government also implemented MySejahtera, a contact-tracing system mobile device app. MySejahtera provided a database of COVID-19 cases in communities and charts of patients’ prior movements, to help warn the public about hotspots [18]. Some 6.2 million users downloaded the MySejahtera app by July 2020, indicating a high-level voluntary usage of the tracking app among the public [19]. The app also received an award at the Ministry of Health Innovation Day 2020 that recognized its robust features and contribution in tackling the pandemic [20]

Image Repair Framework

This study utilized Image Repair Theory [21] to analyze government communications in the media. The theory proposed 5 main strategies and 15 substrategies of image repair. The main strategies proposed by the Image Repair Theory are denial, evasion of responsibility, reducing offensiveness, corrective action, and mortification. These strategies can be utilized independently or collectively to improve public perception. Image repair strategies are frequently used by organizations in crisis to gain favorable responses from the public. Studies [22-24] have also showed how different image repair strategies were able to help improve government’s image in crisis situations.

One study [25] suggested that a majority of the public form their perception of risk of a public health crisis from the media they consume. Various media frames may be used to deliver public health messages to help contextualize the message and urge the public to take action. Among the commonly utilized frames are risk magnitude, self-efficacy, episodic framing and economic uncertainty, which are used to communicate symptoms, likeliness to contract a disease, or protective measures the public may undertake [7]. Restoring damaged perceptions may require effective framing contexts to deliver key messages to the public. This is because the media often use sensationalism to ensure engagement from the audience [8] and this may jeopardize individual’s or organization’s image and reputation. It has also been suggested that studying the trend of public attention in the media may also help authorities determine appropriate frames to deliver key messages effectively [26]. Even though this study did not specifically investigate framing in the media, it is important to note that media messages can play a huge role in influencing public behavior [27].

While the theory [21] categorizes image repair strategies, it does not propose that one strategy is better than the other. However, it is useful to look at how these strategies are utilized and how the public responds to them in different context. For example, a study [22] found that despite different strategies used to improve Chinese government’s image during the SARS crisis, the effort was unsuccessful due to frequent contradictory messages. In another study [23], the Chinese government was successful at restoring public’s confidence using denial, bolstering and corrective action strategies in their communications. Similar strategies were also utilized by the Saudi Arabia government when faced by accusations of terrorism back in 2003 [24].

Although prior studies have recognized the importance of image repair in various crises, investigation of public sentiment on image repair strategies has been limited. Therefore, to address this knowledge gap, we explored the following research questions: (1) What are the image repair strategies employed by the Malaysian government in their communications on COVID-19 in the media? To investigate which image repair strategies utilized in the media by the Malaysian government on COVID-19, it is pertinent to analyze media statements by government officials. These statements may contain either one or more strategies outlined in Benoit’s image repair theory. (2) What is the public sentiment towards COVID-19 media statements from the Malaysian government? Government communications on COVID-19 may bring about different kinds of sentiments from the public. As the direction of sentiment may become an initial indicator of the effectiveness of image repair efforts, this study will scrutinize public’s responses toward COVID-19 media statements by the Malaysian government. (3) What is the relationship between image repair strategies and the direction of public sentiment on COVID-19 media statements by the Malaysian government? By examining the relationship between image repair strategies and the direction of public sentiment, this study aims to describe the effectiveness of the Malaysian government’s effort in communicating about COVID-19.

Methods

Data Collection

This study selected 2 mainstream newspapers—Berita Harian and The Star—for data collection. Statements made by the Malaysian government were retrieved via online Facebook platforms (Berita Harian Online and The Star Facebook page). These 2 newspapers have a broad influence, with 5.7 million (Berita Harian) and 1.2 million (The Star) page followers in Malaysia. Due to their prominence and influential nature, media statements from selected officials (ie, the Prime Minister, Ministers, Director General of Health and Inspector General of Police) were chosen. Additionally, responses to questions from several ministries were also included. Online comments from the public on statements made by these government officials were also collected to answer the second research question.

Sample Period

Key official media statements published in Berita Harian Online and The Star Facebook page pertaining to COVID-19 from March 4, 2020 (2 weeks before the implementation of Movement Control Order) to April 30, 2020 (6 weeks into the implementation of the Movement Control Order) were collected.
Online comments from the statements were also extracted from the same period.

Sample Size and Sample Data Collection
To investigate the first research question, 120 media statements were taken as a sample. Subsequently, we estimated [28] a sample size of 382 out of 59,941 online comments to examine the direction of public sentiment; this sample has a confidence level of 95% and margin of error of 5%. Sampling with stratification (50% from Berita Harian Online and 50% from The Star Facebook page), by selecting top comments labeled as “most relevant” by Facebook algorithm from each media statement, was used. The algorithm sorted and ranked comments based on the highest quality of comments that are relevant to users [29]. Only comments with Malaysian names were taken as sample [30]. These comments were then included for content analysis based on image repair framework.

Content Analysis and Coding Scheme
Content analysis was employed to identify image repair strategies utilized in the 120 online media statements by the Malaysian government and 382 comments from the public. Coding scheme for research question 1 was adapted from a previous study [31], in which 5 main strategies and 15 substrategies were used. The strategies are denial, evading responsibility, reducing offensiveness, corrective action, and mortification. The substrategies were simple denial, shifting blame, provocation, defensability, accident, good intention, bolstering, minimization, differentiation, transcendence, attacking accuser, fixing problem, preventing recurrence, admitting fault, and apologizing.

To answer research question 2, this study categorized sentiments by employing Valence Aware Dictionary for Sentiment Reasoning (VADER) text analysis technique [32]. Through this particular sentiment analysis method, public comments were categorized into positive, negative or neutral. Despite the language limitation of the VADER technique, which only catered to English language [33], this study was able to identify and replicate its algorithm to the Malay language for when analyzing comments in Bahasa Malaysia. To correspond with VADER analytical algorithm, 4 main principles were applied toward analyzing social media text in online comments: punctuation, capitalization, intensifiers, and conjunction [33]. Results obtained from research question 1 and research question 2 were then to measure relationships outlined in research question 3.

Validity and Intercoder Reliability
Validity for coding instrument in content analysis can be described as face validity, construct validity, content validity, and criteria validity [34]. The coding instrument for the study has been face validated by a researcher based at the Universiti Kebangsaan Malaysia.

Two coders analyzed the media statements and public comments. To ensure intercoder reliability, both coders were trained using a coding scheme. Intercoder reliability [35] was calculated as $R = 2M / (N1 + N2)$, where $M$ is the total mutually agreed coding result and $N1$ and $N2$ are numbers of decisions made by coder 1 and coder 2. Each coder analyzed 10 media statements that were not part of the sample ($R=0.74$) and 10 public comments ($R=0.74$). An $R$ value equal or more than 0.70 is accepted as reliable [36].

Statistical Analysis
A statistical analysis was performed using SPSS software (version 22; IBM Corp). For descriptive analysis, frequency tables on media source, publication phase, statement source and image repair strategies (including substrategies) were created to address research question 1. This analysis revealed highest frequency of image repair strategies and substrategies utilized by the Malaysian government. A frequency table was also generated to categorize and reveal direction of public sentiments (positive, negative or neutral) to answer research question 2. Chi-square tests were used to determine the association between COVID-19 image repair strategies by the Malaysian government and the direction of public sentiment from online comments to answer research question 3.

To measure the effect size for cross-tabulated data, Pallant [37] has suggested that for a table larger than 2 by 2 Cramer $V$ takes into account degrees of freedom, with $V<0.07$ small, $V=0.08-0.21$ medium, and $V>0.35$ large suggested.

Results
The reducing offensiveness strategy was the most utilized strategy in COVID-19 media statements (75/120, 62.5%). It was used more than the corrective action strategy, which was the second-most used (30/120, 25.0%) (Table 1).

For image repair substrategies, a total of 454 statements were coded (Multimedia Appendix 1). The bolstering positive quality substrategy (reducing offensiveness strategy) appeared the most in the media statements (108/454, 23.7%), while the fixing problem substrategy (under corrective action strategy) came in second (104/454, 22.9%). Other frequently occurring substrategies were preventing recurrence (102/454, 22.4%) and minimization (38/454, 5.3%).

---

**Table 1**

<table>
<thead>
<tr>
<th>Strategy</th>
<th>Frequency</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reducing offensiveness</td>
<td>75</td>
<td>62.5%</td>
</tr>
<tr>
<td>Corrective action</td>
<td>30</td>
<td>25.0%</td>
</tr>
<tr>
<td>Preventing recurrence</td>
<td>102</td>
<td>22.4%</td>
</tr>
<tr>
<td>Minimization</td>
<td>38</td>
<td>5.3%</td>
</tr>
</tbody>
</table>
Table 1. Image repair strategy frequency.

<table>
<thead>
<tr>
<th>Image repair strategy</th>
<th>Instances, n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Denial</td>
<td>10 (8.3)</td>
</tr>
<tr>
<td>Evade responsibility</td>
<td>4 (3.3)</td>
</tr>
<tr>
<td>Reduce offensiveness</td>
<td>75 (62.5)</td>
</tr>
<tr>
<td>Corrective action</td>
<td>30 (25.0)</td>
</tr>
<tr>
<td>Mortification</td>
<td>1 (0.8)</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>120 (100.0)</strong></td>
</tr>
</tbody>
</table>

Findings (Multimedia Appendix 2) also revealed that most media statements were made by the Director General of Health (24/120, 20.0%), followed closely by the Prime Minister and Senior Minister of Defense, (both 22/120, 18.3%). In terms of media sources, Berita Harian published more media statements from the Government on COVID-19 (87/120, 72.5%) than The Star (33/120, 27.5%). This study also observed that most media statements were published during the Movement Control Order (105/120, 87.5%) as opposed to before the Movement Control Order (15/120, 12.5%) (Multimedia Appendix 3).

A sample of Facebook comments (382/59,941) from the 120 media statements showed more positive (247/382, 64.7%), than negative (86/382, 22.5%) or neutral (49/382, 12.8%) sentiments. Berita Harian had slightly more positive sentiments (134/382, 70.9%) than The Star (112/382, 58.3%). Positive sentiments were higher in both media; however, there were more negative sentiments (Berita Harian: 34/382, 18.0%; The Star: 52/382, 27.1%) than neutral sentiments (Berita Harian: 22/382, 11.1%; The Star: 27/382, 14.6%) recorded.

The denial strategy received the most positive sentiments (33/47, 70.2%) (Table 2), while the corrective action strategy came second (42/60, 70.0%). The reducing offensiveness strategy, which was most utilized in Government statements, was the third highest in receiving positive sentiments (169/264, 64.0%). There was a significant association between image repair strategies of the Malaysian government COVID-19 media statements and the direction of public sentiment with a small effect (n=382, $\chi^2 = 0.146, P=0.039$ Cramer $V=0.039$).

### Table 2. Direction of sentiment.

<table>
<thead>
<tr>
<th>Direction of sentiment</th>
<th>Strategy, n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Denial (n=47)</td>
</tr>
<tr>
<td></td>
<td>Evade responsibility (n=10)</td>
</tr>
<tr>
<td></td>
<td>Reduce offensiveness (n=264)</td>
</tr>
<tr>
<td></td>
<td>Corrective action (n=60)</td>
</tr>
<tr>
<td></td>
<td>Mortification (n=1)</td>
</tr>
<tr>
<td>Positive</td>
<td>33 (70.2)</td>
</tr>
<tr>
<td>Negative</td>
<td>6 (12.8)</td>
</tr>
<tr>
<td>Neutral</td>
<td>8 (17.0)</td>
</tr>
<tr>
<td></td>
<td>2 (20.0)</td>
</tr>
<tr>
<td></td>
<td>169 (64.0)</td>
</tr>
<tr>
<td></td>
<td>42 (70.0)</td>
</tr>
<tr>
<td></td>
<td>1 (100.0)</td>
</tr>
<tr>
<td></td>
<td>4 (40.0)</td>
</tr>
<tr>
<td></td>
<td>66 (25.0)</td>
</tr>
<tr>
<td></td>
<td>10 (16.7)</td>
</tr>
<tr>
<td></td>
<td>0 (0)</td>
</tr>
<tr>
<td></td>
<td>4 (40.0)</td>
</tr>
<tr>
<td></td>
<td>29 (11.0)</td>
</tr>
<tr>
<td></td>
<td>8 (13.3)</td>
</tr>
<tr>
<td></td>
<td>0 (0)</td>
</tr>
</tbody>
</table>

### Discussion

#### Principal Results

A range of image repair strategies were employed by the Malaysian government in their COVID-19 media statements and could be categorized according to the 5 strategies suggested by Image Repair Theory [21]. In particular, the reduce offensiveness strategy was the most utilized strategy in communicating COVID-19 in the media, and the majority of these media statements employed the bolstering positive quality substrategy. There was also a significant association ($P=.04$) between image repair strategies and the direction of public sentiment. Although the reducing offensiveness strategy was most utilized, results showed that the denial strategy received the highest positive sentiments, which was followed by the corrective action strategy, and then the reducing offensiveness strategy. Nonetheless, overall sentiment towards government’s messaging were positive.

Although the image repair theory does not stipulate crisis management strategies employed by the Malaysian government, the theory provided a framework to analyze their statements in the media [38]. An understanding of how a crisis is framed by the media helps to identify strategies that might work in future crisis communications.

This study revealed that the strategy utilized most by the Malaysian government in their COVID-19 media statements was reducing offensiveness. An example of reducing offensiveness was a statement by the Health Minister explaining the need to minimize visiting hours at the hospitals as a proactive measure to reduce the risk of COVID-19 infection. A previous study [39] revealed a similar result; reducing offensiveness in the wake of Russian tourism ban to Egypt due to safety and security concern was proven to be effective and has resulted in Russian flight resuming its operations to Cairo in April 2018 after a 2-year ban.

### Implications, Limitations, and Future Work

This study contributes to current literature on COVID-19 in Malaysia [40,41] as well as understanding COVID-19 crisis communications by the Malaysian government based on the framework suggested by Benoit [21]. It adds to the body of knowledge on image repair strategies and public opinion, which may be useful in a health crisis with this global magnitude.
Government efforts to mitigate the spread of COVID-19, such as enacting the Movement Control Order, have had a big impact on population, such as closure of school and non-essential businesses. Therefore, communication of these types of orders must be done correctly to not instigate fear and panic among the public [42]. Negative sentiment from the public could hamper government efforts and cause distrust in the health care delivery system.

A prior study found that, in Malaysia, television and internet news portals are primarily used to access information on COVID-19 [40] and suggested that health authorities should pay considerable attention to the use of appropriate media channels and sources to allow for more effective dissemination of critical information to the public. By identifying too which image repair strategies the Malaysian public responded well, the findings of our study provides insight into information framing that can receive positive responses from the public. We suggest using the denial, corrective action, and reducing offensiveness in television and internet news portals to communicate about crises.

The government should evaluate strengths and limitations of a country in addressing a health crisis [43]. A previous study [44] highlighted several challenges in communicating in a crisis including misinformation, lack of guidance, and information leakage. It has also been suggested that social media caused more confusion, rather than consolidating public effort against the pandemic [45]. Therefore, information in the media must focus on improving trust, building social solidarity, and reducing chaos, while educating the public on prevention measures and reducing burden on the health system. With suitable and effective image repair strategies, the government could minimize public uncertainty and mitigate the spread of false information.

One of the limitations of this study pertains to the sampling period. As the media statements and comments were captured in a specific time span, the results obtained are only applicable and true to the specific time range. Additionally, the study only took public opinions from Facebook comments into account. Results may have differed if the study extrapolated samples in other social media such as Twitter or Instagram. In addition, the consistent appearance of well-liked individuals such as the Director General of Health, may have contributed to the overall positive responses. The Director General of Health was considered a national hero and an exemplary leader in crisis, and his appearances in front of the camera to deliver daily updates on COVID-19 in a calm and composed manner has earned him the image of a rationale leader, providing assurance and tranquility to the public [46]. An investigation into the roles of frequent media appearances and leadership figures in times of crisis, as well as its influences toward public acceptance should be explored in future studies.

**Conclusions**

This study provided comprehensive insight into image repair strategies in the media by the Malaysian government and how members of the public reacted in response to these strategies. The findings of this study could be useful to advise future crisis communication planning, particularly in a health crisis.

**Conflicts of Interest**

None declared.

Multimedia Appendix 1

Image repair substrategy frequency distribution.

[DOCX File, 13 KB - imir_v23i8e28074_app1.docx ]

Multimedia Appendix 2

Source frequency distribution.

[DOCX File, 13 KB - imir_v23i8e28074_app2.docx ]

Multimedia Appendix 3

Publication phase distribution.

[DOCX File, 13 KB - imir_v23i8e28074_app3.docx ]

References


Abbreviations

SARS: severe acute respiratory syndrome

VADER: Valence Aware Dictionary for Sentiment Reasoning

Please cite as:
Masngut N, Mohamad E
Association Between Public Opinion and Malaysian Government Communication Strategies About the COVID-19 Crisis: Content Analysis of Image Repair Strategies in Social Media
J Med Internet Res 2021;23(8):e28074
URL: https://www.jmir.org/2021/8/e28074
doi:10.2196/28074
PMID: