phg
foundation
making science
work for health

Citizen generated data and health

Predictive prevention of disease

UNIVERSITY OF
CAMBRIDGE

Authors

Sarah Cook, Tanya Brigden, Sobia Raza, Alison Hall, Laura Blackburn

Acknowledgements

The PHG Foundation is grateful for the insights provided by the individuals who were interviewed for this report. While their contribution has been invaluable, the responsibility for the content and analysis in this report rests with the authors. Full acknowledgements are listed in Appendix 2.

Published by PHG Foundation

intelligence@phgfoundation.org

November 2020

URLs were correct as of October 2020

This report can be downloaded from
www.phgfoundation.org

The PHG Foundation is a health policy think-tank and linked exempt charity of the University of Cambridge. We work to achieve better health through the responsible and evidence based application of biomedical science.

We are a registered company, no. 5823194

UNIVERSITY OF
CAMBRIDGE

Executive summary

As the UK NHS move towards increasing digitisation of data and is collecting, analysing and managing rising volumes of data, there is a new opportunity to consider how forms of data collected outside the health system can contribute to the effective delivery of healthcare.

Citizen generated data (CGD) is a form of data generated by citizens outside formal health systems that can nevertheless provide insights into health and wellbeing, and that could potentially be harnessed for disease monitoring and treatment. The scope of CGD is very broad, and varies in terms of the type of data collected and the means used to collect it. For example, it can include: smartphone sensors and/or apps to measure or track health parameters such as mobility patterns, mood, sleep, diet; at home medical devices such as blood pressure monitors; online activity such as search engine queries and interactions with retailers or social media users; and direct to consumer genetic or physiological tests.

With prevention high on the UK health agenda, CGD collected throughout the life-course by people who are healthy, unaware of increased disease risk, or showing early signs of ill health should also be considered as a source of potentially useful information to inform prevention strategies both at an individual and population level.

The depth and variety of CGD covers the full spectrum of the health sector, from public health and social care to primary and hospital-based care. We outline essential policy considerations in order to support the development of cooperative strategies for optimising the future use of CGD for predictive prevention.

Relationship between citizens and the health system

Using CGD for predictive prevention breaks down the boundaries between individuals as patients and individuals as healthy citizens. In order for CGD to be fully utilised for predictive prevention, the health system will need to re-evaluate its relationship with the citizen, and consider how conversations might take place to inform citizens of the benefits of gathering information about themselves when in good health, and also to address potential harms, such as providing reassurance about long term data protection and safeguards. This adjustment of the health system-citizen relationship also changes the traditional view of what constitutes being patients within the health system, expanding to citizens interacting with the health system both in sickness and in health.

Gathering evidence, system-wide coordination

Smaller one-off projects to investigate the use of CGD for predictive prevention are already in progress, and these should continue in order for the health system to gather knowledge about which approaches are the most promising. However, as more evidence is gathered, the main health system stakeholders – the Department of Health and Social Care, NHS England, NHS Digital and NHSX, Public Health England and the planned National Institute for Health Protection, – should consider adoption of a coordinated strategy to maximise the benefits of using CGD in this way.

Managing expectations

Digital technologies evolve quickly, but early expectations of what they will deliver for healthcare are not always fulfilled. Clear and suitably adaptable systems of evaluation will be essential to manage the constantly shifting technology landscape. Ongoing dialogue will also be needed, not only within the health system but also with citizens (including patients and the general public) to manage expectations of what digital tools can and cannot reasonably deliver. Open discussion about data sharing and consent processes to facilitate use of CGD, including sharing agreements, will be of particular importance.

Societal impacts

People will vary in their capacity to engage with innovations producing CGD linked to the health system. In the short term, this is likely to increase health inequality, as digitally literate citizens are likely to benefit from accessing personalised services where these provide proven benefits. It is important to avoid discrimination in the nature or quality of health service provision for people who do not provide CGD: those who lack digital knowledge and are unable to generate relevant personal data, or who lack the resources to be able to engage with health services, or who have chosen to opt out of generating or sharing their confidential patient data more widely via mechanisms such as the National Data Opt-out. Substantial efforts should be made to mitigate any potential inequalities that might arise from the expansion of CGD for health, and the health system should be transparent with the public about how data is used for prevention and the anticipated long and short-term impacts of data sharing. These actions are needed to demonstrate that actors, systems and processes are sufficiently trustworthy to maintain public confidence.

Commercial sector collaboration

Since the commercial sector is likely to drive most of the technological development relating to lifestyle and wellbeing, it is vital that the health system works effectively with the commercial sector to ensure that mutually beneficial tools are developed. The National Centre of Expertise planned by NHSX to provide specialist commercial and legal advice to NHS organisations entering data agreements would be an ideal vehicle for this.

Conclusion

While CGD has considerable potential to support the health system’s prevention agenda, this potential has yet to be realised. The health system should decide whether it wants to make use of CGD in this way, and if so, how to tackle issues of citizen engagement, data collection, ethics, regulation and constructive collaboration with the technology sector.

Contents

1 Introduction

1.1 Citizen generated data and prevention

This report explores the role of citizen generated data in contributing to the UK National Health System (NHS) goals for predictive prevention for all those with an interest in building personalised preventative health systems and services. It was informed by semi-structured interviews with healthcare experts on the opportunities created by CGD for predictive prevention strategies, and the technical, ethical and practical challenges posed by potential efforts to harness CGD to this end. The report excludes detailed analyses of legal and regulatory issues which inform how data can be used for predictive prevention, such as what counts as 'personal data' and the legal bases of data processing in a given context.

The NHS has been collecting data on individuals since inception, but accelerating digitisation (including the switch to electronic health records, the uptake of new digital health tools, and rapid advances in imaging and genomic technologies) means that data is increasingly captured in a digital format. This makes the data easier to store, access and process, creating fresh opportunities to use this growing wealth of healthcare data to improve care and inform research.

This report considers the health system in its broader context, to include any organisation involved in the delivery of care, preventive strategies, and which collects, manages and stores data from and about individuals in England. These organisations currently include, but are not limited to, the Department of Health and Social Care, NHS England, Public Health England, NHS Digital and NHSX.

Whilst the data collected within the health system has potential to be used much more effectively in predicting patient risk of further disease, treatment outcomes and improving the coordination and quality of care of patients, enabled by digitisation, other sources of data for informing health have remained largely unexplored until relatively recently.

Data generated by citizens outside the health system – referred to as citizen data or citizen generated data (CGD) (Table 1) – is an ever-increasing resource, the value of which for informing health and care is beginning to be recognised. This data can be generated by individuals directly, but also indirectly by their activities and by their interaction with their wider environment.

Unlike most health system generated data, such as test results and clinical diagnoses, CGD is produced by individuals throughout their daily lives and could provide rich datasets for analysis to uncover important insights into the health and wellbeing of individuals, or into wider determinants of health such as the environment. However, there is currently little guidance available on the use of CGD, and few initiatives within the health system actively seeking external sources of data with the potential to provide greater insights into individual and population health.

CGD includes data that is directly related to health and intentionally generated by individuals, such as health apps and wearables, and also data that is indirectly linked to health, produced by interacting with a digitised commercial sector, for example online banking or supermarket loyalty cards [1]. Another rich source of CGD is that created through online search engine queries and social media interactions. This form of CGD as a potential source of health insights has arguably received the most attention from the health system in recent years. In particular, internet derived data has been studied from a public health perspective for disease surveillance [2].

Table 1. Sources of citizen generated data and potential relevance to health

Source of CGD	Types of data collected	Example
Smartphone	Location (gyroscope), images/videos (camera), location (GPS), usage (e.g. frequency of texts/ calls), environmental factors (e.g. pollution levels, energy usage)	Inbuilt smartphone sensors could identify mobility patterns, or an individual's interactions with their support network or environment. These can indicate physical and/or mental health state.
Health apps	Data through smartphone sensors, connected wearables or user input. User input can vary widely but common parameters include: mood, diet, sleep quality and menstrual cycle dates	Users can track any number of variables around their health that could help them in family planning, managing mental health and planning/tracking lifestyle factors such as diet and exercise
Personal medical devices used in the home	Blood pressure, blood glucose, blood oxygen levels	Although these devices will be predominantly used by those managing long term conditions, otherwise healthy individuals may choose to buy them out of interest or concern for their health
Online activity	Search engine queries, social media interactions, online shopping	Words and phrases users input into search engines and social media posts can give indications of their mental wellbeing or symptoms they are experiencing relating to their physical or mental health
Digital interactions with retailers/banks	Purchase/financial transactions	The patterns in frequency of purchases/ banking transactions along with other data such as timing and quantity could be useful for indicating risky behaviours (e.g. unhealthy eating habits or gambling)
Commercially available genomic or biochemical tests	Genomic variants/SNPs, levels of biomolecules within biological samples	Depending on the type of test and what is being looked at, important information could be gained about individuals' underlying and current risks of ill health (e.g. identification of vitamin deficiencies from blood samples)

Data can be actively or passively produced, the former indicating that the citizen has had to actively participate in the generation of their data (e.g. inputting data into an app) and the latter denoting that the data has been generated and collected without the participation of the individual (e.g. an app using GPS on the mobile device).This can sometimes mean that passive data is generated without the knowledge of the individual. Even more pertinent to the use of citizen generated data are the intentions behind the data production. It is therefore important to establish a distinction between data that is produced with the primary purpose of deriving health-related learning i.e. intentional CGD, from that which is produced without the intention that any health-related inferences are made from it i.e. unintentional CGD.

Harnessing CGD offers a variety of opportunities that can be applied within multiple current and new models of care, as well as public health interventions. Improved prevention is a key strategic area for the health system, and an area where CGD has particular potential [see 2.2]. Whilst much can be gained from analysing datasets produced and held within the health system to improve patient care, greater consideration is urgently needed on how to best utilise data to prevent people from becoming ill in the first place, or to delay the onset of disease. This requires an understanding of the patterns of behaviours and physiological markers associated with the wider determinants of health and very early signs of ill health. There is therefore an opportunity to explore how CGD could be used to gain insights into citizens' health before they even interact with the health system.

1.2 Predictive prevention

Recent policy documents highlight prevention of disease and ill health as key goals for the health system in the medium to long term [3-6]. These include the 2018 policy paper Prevention is better than cure: our vision to help you live well for longer; the 2019 NHS Long Term Plan; and the 2019 Green Paper Advancing our health: prevention in the 2020s. The latter two documents set out a vision for the health system to not only deliver prevention strategies to citizens through digital technologies, but also to use personal data to tailor those interventions.

Prevention strategies are designed to prevent the onset or worsening of disease caused by modifiable factors. Prevention is therefore achieved through a wide range of interventions that aim to reduce the risk or threat to health. In theory, there are three levels of prevention [7]:

- Primary prevention: stopping people becoming ill in the first place
- Secondary prevention: intervening at an early enough stage that the severity of the disease is diminished or reversed
- Tertiary prevention: improvements in terms of quality of life and symptom management of an established disease

The concept of predictive prevention was first outlined in 2018 by the Department of Health and Social Care as including the use of personal data (appropriately safeguarded and in compliance with data protection regulation) to find new ways of predicting who may become ill, in order to proactively target those at greatest risk with tailored preventative interventions using cutting-edge technology [3]. Whilst predictive prevention is aimed broadly at primary, secondary and tertiary preventative measures, harnessing CGD is particularly novel for primary prevention.

Information about individuals, such as their external environment and health-related behaviours, is required in order to make predictions about their risk of ill health. This information can be gained through various means including the mining of routinely collected data sources/data collection strategies including: healthcare records, national surveys and data collected through local authorities. CGD is a source of currently underused data that has potential as a valuable resource for informing predictive prevention strategies.

1.3 How might citizen generated data support predictive prevention?

There have been numerous commentaries around the possible future use of CGD for informing health and care, and in predicting ill-health. This has been fuelled by growth of the direct-to-consumer wearable and devices market, advances in data analytics including machine learning, and increased interest in connecting disparate data sources (i.e. the Internet of Things), particularly within the commercial sector. The potential of CGD to be used in personalised prevention strategies was highlighted in the Chief Medical Officer for England's 2018 annual report – Better Health 2040 – Better Health Within Reach – for example in conjunction with social, economic, behavioural, biomedical and genomic data, to generate personalised and 'real-time' estimates of disease risk [5].

Currently, people present to the health system because they are experiencing symptoms, managing a long-term condition, or might have concerns about family history of disease. A major challenge in implementing predictive prevention is therefore how to gain information about people who are not yet interacting with the health system – i.e. are not yet patients – and use this information to target interventions aimed at delaying or preventing the onset of ill-health.

Asymptomatic individuals can choose to take part in national screening programmes such as those for breast or bowel cancer, but otherwise will not present to the health system. However, many individuals will go on to experience ill health in the future where the onset of disease could have been delayed or prevented, had tailored interventions been applied earlier; for example, type 2 diabetes or mental health disorders,

Harnessing datasets that are produced by individuals outside of the health system or other routine data collection routes (i.e. local authority data) has great potential value for informing predictive prevention strategies. CGD could help to deliver two important goals of predictive prevention:

1. Identifying those at increased risk of developing disease
2. Providing citizens with more information about themselves so that they can make informed decisions around their health and care

Examples of how CGD could be used include:

- The targeting of tailored preventative digital interventions or advice to individuals
- Guiding high risk individuals to specific health services for advice and support
- Providing an additional source of information to refine screening programmes

2. CGD and the UK health system

The opportunities presented by CGD have already been recognised by some parts of the health system; this section summarises some examples of how CGD is being used to support health care delivery, what information could be inferred from CGD, and further opportunities.

The information presented in chapters 3-7 was generated from our own desk-based research and from interviews with experts. Full methods, including a list of the experts interviewed, are given in Appendices 1 and 2.

2.1 What can be inferred from CGD?

In theory, much of the data that citizens generate about themselves can be directly or indirectly related to their health. However, the insights and predictions that a dataset can provide only become meaningful for health when linked with clinically relevant information. Research is currently being undertaken on how various types of CGD can be used to infer and predict onset of a wide variety of diseases. Examples include:

- Analysing posts on Instagram (a social media platform) to predict the likelihood of users suffering from depression [8]
- Early prediction of relapse in people with schizophrenia by monitoring background mobile phone usage [9]
- Using historical internet search data to indicate the future likelihood of receiving a diagnosis of pancreatic adenocarcinoma [10, 11]

In addition to these internet and phone based, unintentional forms of data production, there are a multitude of options for users of direct to consumer (DTC) apps, wearables and devices (Table 1), to actively engage with to produce data relating to their health. The benefit of these digital tools is that users can directly monitor various aspects of their health and wellbeing (e.g. activity, mood, diet) over time, and use this information to help them reach a goal or understand their behaviours. Most wearable devices and smartphones possess sensors that enable the measurement of the users' heart rate. Resting heart rate (RHR) is a known risk marker for a multitude of cardiovascular and metabolic diseases, and an RHR that is steadily increasing over time could indicate an underlying medical condition [12].

The accuracy of some features found on DTC wearables devices has been contested; for example, a study by Shcherbina and colleagues found that none of the seven popular devices they tested had an error below 20% for measuring a person's energy expenditure [13]. However, the underpinning technology in many devices is continually improving, and many health indicators, such as heart rate, steps and sleep duration, have been found to be reliable in popular devices [14]. Furthermore, as long as the results are consistent in their accuracy (i.e. consistently inaccurate), it is the trends in data and deviations from baseline recorded by these devices that can be most illuminating. In general, however, there is a lack of information and knowledge gathering about the analytical validity and clinical utility of using CGD for prevention; to better understand the true potential of CGD for predictive prevention, more is needed. Existing projects should however contribute additional knowledge in these areas.

The type of data that citizens generate from DTC tools is also beginning to change, providing further opportunities. In particular, advances in sensor and computing technology are enabling the development of more sophisticated DTC tools that blur the boundary between lifestyle and medical devices. For example, the newest version of the Apple Watch, currently FDA cleared and available in the US, is equipped with an ECG sensor that shows the user their heart trace, a test traditionally only available through the health system. This trace can be used to alert the user to potential abnormalities such as atrial fibrillation, a risk factor for stroke and heart-related disorders. An app enabling ECG monitoring for medical use under the current Medical Device Directive is classified as a class IIa medical device [15].

Following the announcement from the Medicines and Healthcare products Regulatory Agency in September 2020 that the EU Medical Devices Regulation will not be implemented in Great Britain, this will remain the case until a new domestic regulatory model for devices is introduced. This potentially creates opportunities for citizen-held ECG sensors to be used to inform diagnosis and monitoring in the future, with decreased involvement of the health system. However, in order to be marketed and used for a medical purpose, manufacturers would have to demonstrate that they have satisfied requirements for clinical evaluation as well as other requirements in the legislation.

These regulatory challenges are likely to become increasingly pressing as more opportunities emerge for different types of CGD to present as predictive markers for a range of risks. Most studies typically examine a single source of data and how it relates to one disease, rather than integrating multiple sources of data to provide multidimensional accounts of individuals' health, referred to as the digital phenotype [16]. The breadth of opportunities for predictive prevention are currently limited in part due to the lack of comprehensive datasets that combine these different sources of CGD with each other and with data held by the health system.

2.2 Current examples of citizen generated data for health

There are a number of examples of how CGD is currently being used within the health system, and more broadly in community and social care. In particular, there has been an increase in interest in remote patient monitoring in the past few years. This is where patients are provided with or purchase tools to record various aspects of their physiology in order to manage symptoms, pre-empt exacerbations and alert carers to intervene when necessary. This may include patients under direct care or in the community, such as the elderly, those managing long term conditions or recovering from surgery. Many of these remote patient monitoring systems include the use of different types of CGD as additional sources of data to integrate in prediction models (e.g. TIHM [17], Box 1).

Box 1: Smart home systems

Within the health system
Technology Integrated Health Management (TIHM) for dementia is one of the seven NHS Innovation Test Beds that uses an in-home network of connected devices (via the Internet of Things), along with artificial intelligence, to enable clinicians and caregivers to remotely monitor individuals with dementia. If a problem is detected by the multisensory network – which includes environmental sensors, apps and connected medical devices – an alert is triggered to the care team for follow up. The key aim of this study is to reduce the need for individuals to go into hospital, or be admitted to care homes, thus reducing the pressure on carers.

Outside the health system
These types of platforms are also being employed outside of the health system for independent carers. For example, British Gas [18] are working with Hive, one of the largest connected home providers, on a smart home system that alerts family members when changes in use or in movements around the home are detected. These systems are proving to be useful for secondary and tertiary prevention in those who already have an established disease or are at high risk of suffering adverse events.

There are also many examples of aggregated CGD being used to inform population health, termed digital epidemiology [19]. Various digital surveillance platforms exist, such as HealthMap [20], that use CGD (typically from online sources) in their software models to predict emerging public health threats (e.g. foodborne illness outbreaks) in real time. Within the health system, Public Health England has reported the use of CGD for disease surveillance activities (e.g. an E. coli outbreak in 2017 [21], Box 2). In addition, CGD has been used for population-based modelling by other authoritative bodies closely linked to the health system such as the UK's Food Standards Agency monitoring tweets to help detect norovirus outbreaks [22].

Box 2: Tracing an E. coli outbreak

In 2017, disease surveillance scientists at PHE used data generated from supermarket loyalty card accounts to help trace the source of a large E. coli outbreak and thereby prevent further cases. Once the genomic analysis of the E. coli strain had been confirmed to be the same in all those affected, scientists started to piece together the information they required to determine the source. Interviews with those affected identified a common connection between them all – purchase of burgers from a particular retailer. PHE asked the supermarket to analyse the loyalty card data of those affected and discovered the exact brand of affected burger, resulting in a food recall by the Food Standards Agency.

The COVID-19 pandemic has increased interest in the use of various technologies to monitor and analyse information generated by citizens to help in efforts to monitor and control infections. Prominent among these is the NHS Test and Trace App, which combines mobile phone positioning data from individuals who test positive for the SARS-CoV-2 virus with information about other mobile phone users who have been in close proximity to that individual. These data are used by a risk-scoring algorithm to determine which potential contacts of the infected individual are also are at risk of infection.

Whilst infection control is the area of most rapid innovation in harnessing CGD arising from the pandemic, the necessity to strictly limit in-person access to healthcare services has in itself led to significant progress in the development and implementation of remote digital health tools. For example, remote sensors in hospitals and care homes controlled by artificial intelligence (AI) tools can be used to create ‘ambient intelligence’ and provide information about measures such as mobilisation of patients and residents, or the use of hand sanitising equipment by caregivers [23]. Combined with new awareness of the potential public health benefits of sharing personal data, the pandemic may therefore have paved the way for wider acceptance of CGD and expansion of the use of CGD across various public services, including for personal and population health, and further into fresh environments such as the home or workplace.

Uses of digital epidemiology vary in their predictive power and ability to inform local intervention strategies, therefore, their utility depends on multiple factors, such as data source and application. Most of these aggregated data models do not have the resolution to predict which individuals may be suffering from an illness. However, aggregated data could be used to raise awareness in individuals who are in a geographical area where it has been detected that there is a risk of high disease incidence. This might be particularly beneficial for those more vulnerable to infection such as the very young, the elderly, and those who are immunocompromised.

In the same way, COVID-19 risk algorithms have been used to assess the risk that an individual may be infected with the SARS-CoV-2 virus. The Alipay ‘Health code’ app widely used in China to control the pandemic works in this way, combining medical and travel history to produce a traffic-light based assessment of COVID-19 risk. However, proposals have already been made to expand the use of this app; the government in Hangzhou announced plans for longer-term use to produce ‘citizen health scores’ based on personal data such as step count, sleep duration, and consumption of alcohol and tobacco [24], although citizens reportedly expressed concerns about these intentions.

Box 3: Digital wellbeing service

Good Thinking is a digital wellbeing service rolled out across the city as part of the Healthy London Partnership; it uses data-driven marketing techniques to target advertisements for digital services to people who may be experiencing mental health issues. This targeting is based on people’s use of online search engines and social media platforms, thereby proactively identifying those who may benefit from services and who would not necessarily self-present to the health system. Those who display patterns of searching or social media posts consistent with early predictors of mental health decline are targeted with subtle advertisements around their personal issue. If the user engages with the advertisement, they are filtered through to a digital service containing recommended and approved apps for their specific problem. All this is done without the health system requiring access to raw data or any personal information about the user and the citizen is not aware they are engaging with the health system.

This approach of directing individuals to personalised health services, identified by their online activity, is one of the ways CGD can be employed to target those at risk without them needing to be in contact with the health system first. The Good Thinking programme is being evaluated by academics at King's College London on the clinical, economic, behavioural and digital outcomes of the service. The approach has multiple potential benefits, including:

- Providing individuals with another option for care through digital, guided self-help; enabling early intervention, which may stop individuals progressing on to developing more severe forms of illness
- Reaching individuals who would not present to the health system either because they are not aware there are interventions available for their specific problem or they do not want to seek help from the health service due to a variety of reasons including: concerns around stigmatisation, a history of negative experiences with seeking help, views on medication and preferring not to speak to another person

In fact, in clear contrast to the use of CGD for remote patient monitoring and disease surveillance, there are very few current examples of using CGD to proactively seek out and predict when otherwise healthy individuals may become ill, before they become patients, including in the UK. However, some innovative programmes actively use CGD in predictive prevention models, such as the exemplar outlined in Box 3 [25], which is supported by a partnership between the Mayor of London, the London Councils, Public Health England and the NHS.

Another example in development involves plans to support the NHS Health Check, a programme that aims to identify early signs of non-communicable disease in those age 40-74, with digital technology [26]. The NHS Health Check Programme Digital Exemplar, in partnership with PHE's national Digital, CVD Prevention and Behavioural Insights Teams, is early in its development, currently midway through the second part of the 'discovery' phase, which looks specifically at user needs. The next stage, alpha stage, is where prototypes of the product will be created and tested with user groups. The aim is that through the NHS app, citizens will be able to upload their own data such as that from fitness trackers, which will be used along with other information, such as weight and blood pressure, to provide personalised risk information to enable further targeted screening or intervention.

Whilst these developments are welcomed and encouraged, they are just the beginning in terms of gathering evidence on the wider utility of using CGD, and in exploring the potential that CGD offers for predictive prevention to improve the public's health. Many of these uses are ad hoc, locally driven and delivered, and responsive to local needs, without plans for wider regional or national implementation. Some of the strategies run by PHE and NHS England could be expanded if current pilot projects are successful, but there is currently no coordinated, national strategy for how and when CGD might be used for predictive prevention. The announcement of new plans to close PHE from 2021 and create a new National Institute for Health Protection [27] casts further doubt over the strategic policy direction for using CGD, especially as it is not yet known how wider health improvement efforts will be coordinated.

Nevertheless, interest in the applications for CGD is rising, and these possible uses raise a number of important considerations for the wider health system.

2.3 Further opportunities for the health system

Harnessing datasets produced by citizens outside of the health system offers significant opportunities for a variety of applications. For example, in addition to some of the sources of CGD beginning to be incorporated into remote patient monitoring systems, there is also potential to incorporate CGD to improve pre-existing disease prediction models and others that are in development, such as polygenic risk scores for cardiovascular disease.

There is a widely acknowledged need for the health system to move towards a more prevention-based approach in general, given the current and predicted increase in future demand on the health system caused by an ageing population, increase in non-communicable diseases such as type 2 diabetes, and patients living longer with multiple co-morbidities. However, there are legitimate concerns that the interim period of a gradual move towards a health system built much more on prevention will place additional demands on stretched clinical services if predictive prevention strategies are implemented without robust models of care and supportive patient pathways in place.

There are important questions to be addressed with respect to how new opportunities could be harnessed, and how the health system could establish which datasets are most useful for predictive prevention. There is a distinction between using data to target citizens based on their personal and individualised usage data (e.g. targeted advertisements) compared to using aggregated CGD data from a variety of sources (e.g. for predicting flu outbreaks). The former application of CGD poses substantial additional challenges including system pressures (e.g. increase in number of people attending primary care due to health concerns); public trust and engagement; linkage of data about individuals from multiple sources; a complex regulatory environment; and ethical considerations around inequalities, responsibility and consent.

3. Health system considerations

The health system has ambitious plans regarding predictive prevention, and is beginning to consider the role that CGD could play in helping it to achieve its goals through some exploratory pilot projects and programmes. However, there are various important system issues that will require careful attention to meet these ambitions, including the impacts on care pathways, expectations of technology and how to engage health system workers.

3.1 Laying the groundwork for health benefits

For the UK health system to harness the opportunities presented by using CGD for predictive prevention, a number of issues require careful attention:

Consideration 1: In the context of limited resources, and continually expanding volumes and sources of CGD, the health system should undertake a scoping and prioritisation exercise to (1) map out the landscape of CGD, including the types of data, sources and accessibility, and (2) determine which types of CGD will be most useful for the purposes of predictive prevention. This could include consideration of specific applications or areas of unmet need that it wants to address, and the anticipated benefits and challenges of harnessing CGD for this purpose.

Consideration 2: Based on this scoping and prioritisation exercise, if there is evidence to support engaging with CGD in a more systematic and coordinated way, the health system should develop a wider strategy for using CGD for personalised prevention. This would include considerations around the wider impact and implications of using CGD for patients, citizens and the health system, and draw on learnings from the COVID-19 pandemic.

Consideration 3: The health system should aim to generate evidence of analytical validity and clinical utility for CGD applications developed within the health system. These and other projects that offer potential clinical utility could be used as exemplars for establishing the feasibility of using CGD for specific conditions and for further implementation. Initially, these are likely to be small scale projects with clearly defined aims. A coordinated and larger regional/national approach will be required by different stakeholders to support implementation of CGD predictive prevention approaches on a system-wide scale.

Consideration 4: When putting pathways and structures in place to facilitate use of CGD for prevention, the health system should consider which stakeholders within the health system are responsible for using CGD in this way, and in what context.

Consideration 5: The health system should foster internal and external collaborations to share and/or link CGD datasets for the benefit of developing predictive prevention strategies. This could be achieved via a data repository that links these datasets, creating an evolving system where the types of data that are most useful can be recognised, and therefore captured and combined with other datasets about individuals to make meaningful predictions. If this was adopted, substantial work would be needed to ensure that such data collection and integration is compliant with relevant data protection legislation and guidance. A high-level, coordinated and systematic approach involving collaboration between the health system and external stakeholders will be needed to achieve this.

3.2 Additional demands on services

The expansion of predictive prevention efforts using CGD have the potential to exacerbate existing pressure on health services if citizens who are yet to experience ill health are being proactively sought out. One element of potential concern is how risk is communicated to the targeted citizen. For example, if an individual is aware that they have been actively sought out by the health system, they might reasonably wish to discuss this with their GP, thereby increasing demands on primary care services. Another issue is how predictive prevention strategies might fit within the current paradigm of screening models; how they are categorised will affect the way they are evaluated, and therefore decisions will need to be made about how to approach these new models.

The increasing number of direct to consumer devices, particularly those that include medically relevant features, raise potential challenges for the health system as it will have no control over what is being developed in the commercial space. Whilst the Medical Devices Regulation and In Vitro Diagnostics Regulation ensure that products that fall under these regulations accurately and reliably measure what they claim to, these Regulations do not require evidence of the clinical significance of results. Therefore, DTC digital lifestyle products that incorporate medically relevant features (e.g. ECG on a smartwatch) raise concerns around what the impact could be on the health system and how it would manage patients who hold this type of data. Enabling citizens to generate and hold data about themselves that has previously only been available in a clinical context could result in many more healthy people unnecessarily seeking advice and medical assistance due to false positive results. Whilst some people with undetected health conditions could benefit from such technology, for the most part these new technologies could also induce unnecessary health anxiety, expanding the number of ‘worried well’ in the healthy population.

Consideration 6: It should be clear how predictive prevention models harnessing technologies generating CGD will be evaluated (i.e. as screening programmes or new models of care). Such applications should be evaluated by the appropriate authority and criteria for their impacts on care pathways, including whether new care pathways should be established to accommodate these strategies.

3.3 High expectations of digital tools

The use of CGD for predictive prevention also raises the question of whether current expectations for what the implementation of digital tools and interventions can achieve is too high. Trying to evoke behaviour change through directing individuals to apps or wearables to self-manage or reduce their risk of disease will, inevitably, only be effective for some people, as not everyone will be willing and able to use digital tools. Others will require additional and alternative support meaning there needs to be an infrastructure built to provide this as part of the implementation of new or changing existing patient pathways. Vulnerable populations in particular may not have the motivations, resources or capabilities necessary to benefit from digital interventions unaided.

A possible counter argument is that the more data the health system collects about how individuals respond to digital interventions, the better it will be able to provide effective, preventative options on a personalised basis. Developing predictive prevention strategies would therefore be an iterative, evolving process, enabling the tailoring of different interventions that will work for the targeted individual.

3.4 Engaging the health system workforce

Using CGD for prevention will require health system professionals to have a different relationship with data, and also with patients. Engagement strategies need to consider not only what CGD could be used for, but also how this might affect working practices. At present, concerns and lack of clarity around information governance often result in health system professionals being unsure about what data they can and cannot use, and what consent is needed. If not addressed, these sorts of issues could similarly result in healthcare professionals being overly-cautious, or reluctant to engage with CGD at all due to uncertainty surrounding their responsibilities once they have done so. As reported in the Topol review on the future of the health system workforce, specific training in the areas of data science and digital delivery of care are required across the health system in order to realise the potential of these technologies [28]. Training needs to include more than just information on the potential advantages of CGD, but also clear guidance on consent processes, and incorporating CGD data into clinical decision making and how this potentially impacts on professional responsibility.

Consideration 7: The health system, in particular NHS England, PHE and successor bodies charged with health improvement, should develop guidance for the workforce around harnessing citizen generated data for predictive prevention strategies. The workforce should also be trained on the opportunities that CGD presents, and how and when to use this data appropriately in clinical decision-making. It is likely that initially this will apply to a limited number of approved devices for specific situations.

Consideration 8: Healthcare professionals should understand and apply appropriate forms of consent for the implementation of different types of CGD. The use of CGD might require changes to existing consent processes, due to difficulties in foreseeing all future uses and the potential for secondary uses of data.

3.5 Health system responsibilities

Healthcare professionals and institutions have a legal duty of care to their patients. However, questions arise as to whether the health system has a duty of care to warn or to act in situations where it has access to CGD collected outside direct clinical care that could help to identify those at risk. These individuals may not be patients or under an established duty of care.

Commercial companies providing social media platforms have started demonstrating an awareness of how the data they hold can be used for prevention, and are taking measures to minimise harm. For example, Facebook has started to use ‘proactive detection’ AI technology to scan online posts for suicidal indications [29], and Instagram have introduced an anonymous reporting tool that redirects anyone searching for a banned hashtag to a support system, and allows individuals to report posts that are related to self-harm or indicate poor mental health [30]. Instagram then provides the author of the post with a variety of options such as talking to friends or local helplines. This raises questions as to whether any formal or ethical responsibilities or obligations arise as a result for companies holding such data, and if so, what threshold for intervention would be most appropriate?

Similar considerations might apply to health systems if they were to collect similar data or access it from third parties. The commodification and commercialisation of health, and the influx of direct to consumer health technologies has led to the blurring of the boundaries surrounding what constitutes a ‘patient’. It is foreseeable that there will be instances where the health system has information on high risk of disease of a currently healthy individual, or has access to CGD from those who are not formal patients but who have some interaction with the health system. Although duties owed to non-patients by health systems are currently limited, it is possible that the duty could be extended to non-patients where there is an expected therapeutic benefit and healthcare professionals take some degree of responsibility for oversight. An analogous debate might be that surrounding the duty to feedback results to participants in observational research.

However, extending the duty of care to non-patients could be very burdensome, and exacerbate liability concerns. In addition, it needs to be made clear to citizens that more accurate risk prediction will not necessarily mean that any and all disease can be identified and prevented.

Consideration 9: The nature of the responsibilities owed by the health system to people who are not under an established duty of care should be clarified. It is also important that public expectations about what the health system can deliver in this situation are realistic and managed appropriately: in particular, management of expectations around the limits of prediction.

4. Engaging citizens

Whilst there is much activity and hype around using digital technologies to empower people to manage their own health, so far efforts have mostly focused on what the health system can provide to the individual (e.g. NHS Apps Library). Predictive prevention proposes a complementary model where the individual is proactively targeted to take part in self-help preventative interventions. This targeting can be based on information that individuals choose to provide to the health system in order for it to be able to help them in the future or from data sources such as online or retailer records.

This new model for prevention will require change in how the health system responds to citizens, a change in the expectations and engagement of citizens, and will pose some unique ethical challenges. Securing and maintaining public trust, and understanding the parameters of what the public find acceptable will be vital, if potentially difficult.

Broadly, there are two distinct challenges around getting public support for predictive prevention strategies:

1. Encouraging those who are not yet ill to produce or collect the data that could be useful for predicting their future health.
2. Engaging with individuals to share that data with the health system for potential future use.

4.1 Engaging healthy people with data collection issues

It is likely that a mixture of different sources of CGD will be helpful for predicting ill health in individuals. Some CGD sources will be relatively ubiquitous across the population, such as energy consumption in the home or the use of internet search engines, and therefore will not require concerted effort from individuals to generate. However, other types of CGD such as that produced by health apps and wearables require individuals to actively engage with the product to create data.

Motivations to actively monitor some aspects of health using DTC digital tools will vary between individuals, ranging from wanting to improve fitness, to managing a health concern. When an individual has a specific reason for monitoring themselves, the results of doing so are instantaneous. Arguably, individual monitoring that may enable future warning of impending ill health might have limited appeal for many, due to the lack of instant utility. This could help explain the high abandonment rate of wearables. A 2014 survey found that over a third of those who own a wearable stopped using it after six months [31], which has led some to consider them a fad and question their utility in terms of collecting clinically relevant CGD.

For some types of data, this may not be a barrier. As sensors become more sophisticated, passive and engrained in our daily lives, citizens may produce more health relevant data without having to be actively engaged. For example, most wearables and smartphones are able to measure heart rate using simple sensors. As increases in resting heart rate can indicate risk of a multitude of diseases this parameter could be a valuable one to incorporate in predictive prevention strategies. However, whilst the ease of ubiquitous sensors that do not require active interaction is often convenient, users might forget that their data is being collected creating concerns surrounding the ongoing validity of consent. This is particularly the case for vulnerable groups, such as the elderly.

Prevention models face difficulties in general, being based on interventions that attempt to evoke behaviour change or specific actions in order to reduce the risk of ill health. This can be especially hard for people who may not yet be experiencing any obvious negative impacts from this ill health in their daily lives. Predictive prevention is expected to have long term benefits, such as a reduction in risk, that are harder to quantify and communicate and often require more effort on the part of individuals. This is in contrast to other primary prevention approaches such as immunisations, where specific diseases are prevented as a result of the intervention.

Careful consideration needs to be given to the nature of the relationship between the health system and citizens, particularly healthy citizens, to facilitate the collection and access of data to support preventive strategies at a population and individual level. Health systems should aim to be transparent about the purposes for data collection and retention and envisaged secondary uses. This will include initiatives around how to engage individuals without overt disease to collect relevant data, and how to promote a shift in the perception of the health system from one that is focused on the management and treatment of acute and long-term illness to one that helps to keep people healthier, longer.

4.2 Encouraging appropriate data sharing

Engaging the public from the very beginning of the development process for initiatives that use CGD for prevention will be vital, encouraging them to contribute their ideas around strategies that use different alternative data sources for predictions about their health. This is important to maximise constructive engagement between citizens and new initiatives, and to avoid negative public reactions such as that around care.data, the unsuccessful attempt to create a large linked data repository of various health data sources in 2014 [32].

As the sources and types of CGD that exist are extremely diverse and heterogeneous in nature, there will be different modes of data sharing and access. In some cases, people may actively want to share their DTC or app data with clinicians; in other scenarios, the data might be gathered indirectly through third parties, or online activity.

The sharing of personal data with healthcare professionals for direct care – treating, monitoring or management of chronic conditions – potentially has real-time, tangible benefits to individuals, and privacy concerns associated with sharing data will need to be weighed against these benefits. If these benefits are not perceived as being immediate and clear cut, for example improving control over symptoms, healthy citizens may be less likely to share their personal data. This could be further compounded by fears or concerns around how the data might be used or linked with other datasets.

For example, someone receiving disability benefit might be concerned that their activity data could be used against them for removing their entitlement to such benefits.

There is undoubtedly potential for predictive prevention to be thought of as ‘Big Brother’ in nature i.e. an all-seeing, controlling presence, which may discourage personal data sharing. The narrative around the use of data for malevolent purposes in the UK, through the print and visual media, could also be detrimental; in combination with highly publicised scandals around using data to manipulate citizen voting strategies – for example Cambridge Analytica [33, 34] – it serves to undermine public trust in the use of CGD in the future, even if the intended use is legitimate and made transparent.

Assuming there is legitimate intention behind the use of CGD, such as citing predictive prevention, questions remain as to whether citizens should be made aware if their data is being used to target them for receiving behaviour change interventions. This is particularly important when considering using CGD that is produced online or via interactions with retailers where the individual does not need to be actively involved in data sharing. This points to a larger issue around the opacity of user agreements on online and digital platforms.

Furthermore, using online data to monitor individuals without their knowledge potentially raises a range of ethical challenges that will need to be addressed, including concerns around potential loss of privacy, the need for public trust, and whether the health system has identified a threshold to act on data received. There are additional problems with targeting individuals for behaviour change interventions, as successful behaviour change relies upon individuals wanting to change their behaviour and the health system being able to react to individual motivations and drivers. A key challenge is that we do not know enough about why people have unhealthy behaviours – they are likely fuelled by a complex interaction of social, physical and emotional factors – placing limits on the usefulness of CGD. It is possible, however, that CGD could provide helpful insight into the factors influencing these behaviours.

Consideration 10: The health system should have ongoing and transparent dialogue with the public to understand what data is considered acceptable to be used for what purposes. Clear communication is required, consisting of simple messaging around how CGD will be used and what benefits will be derived for the population and individuals. In order to foster trust and demonstrate trustworthiness, consent models will need to be tailored to the mode of data access and extent of data sharing.

4.3 Understanding the variation in individuals’ abilities to take responsibility for their health

Public health policy surrounding use of digital technologies employs the rhetoric of ‘empowerment’, to ‘activate’ and ‘engage’ people to gain ‘greater control’ of their own health. This is driven by a move towards patient autonomy as well as the incentives of better health outcomes and lower costs. However, devolving responsibility for health on to individuals, regardless of their circumstances, risks unfairly blaming them for adverse outcomes and downplays structural and social determinants of health.

A multitude of contextual factors such as individual capabilities and resources can prevent individuals adopting a ‘healthy’ lifestyle. They may be less likely to engage with predictive prevention if they feel accountable and ashamed of their lifestyle, or if they lack the requisite resources, have impaired abilities (e.g. as a result of comorbidities and/ or mental health problems) or motivation due to other competing priorities. For example, a demanding work schedule might prevent someone from having the time and inclination to exercise and cook fresh food.

Too much focus on individual responsibility risks marginalising those groups with reduced capacity to understand or respond to health improvement efforts, such as those with lower health literacy in poorly resourced areas. The need for an emphasis on shared responsibility, and to support individuals in making healthy decisions through using assets from communities and the formal health system has been recognised [35], and efforts have been made by the NHS to understand barriers to health and how these might be addressed. Initiatives such as Widening Digital Participation acknowledge the importance of individual context and circumstances and have focused on building individuals’ digital skills [36].

It should be noted that an individual’s ability to take responsibility for their health is in part influenced by wider societal factors, the focus of the next section of this report. Consideration should be given to infrastructure changes that might need to accompany models of care incorporating predictive prevention, including providing additional support and resources (through social prescribing, for example) to help individuals build capabilities that enable them to make lifestyle changes.

5. Wider societal impacts

In addition to individual and health system considerations, using CGD for the purposes of predictive prevention is likely to generate concerns surrounding wider societal impacts, with particular challenges around widening health inequalities, the loss of privacy, and the need for transparency.

5.1 The potential for widening health inequalities

CGD could reduce health inequalities through increasing access to health information and a steadily decreasing digital divide. However, it also has undeniable potential to exacerbate such inequalities, disadvantaging specific groups. These groups might include those who are unable or unwilling to use digital devices or enact changes required to prevent disease due to financial constraints, physiological or psychological barriers, or lack of interest in optimising health through preventative measures.
It is likely that benefits derived from CGD will be primarily gained by those who prioritise their health and have the resources, skills and literacy to access and operate digital tools, and to respond appropriately to opportunities to reduce risk and improve health suggested by such tools. In addition, demographic biases resulting from disparities in the populations that produce large volumes of CGD (including those used to develop tools for predictive prevention) could lead to some groups missing out on clinically relevant insights that could help them prevent disease.

Digital health initiatives are unlikely to benefit all people equally, as those with higher levels of education are more likely to search online for health information [37], can use more accurate search terms, and better identify relevant and reputable sources. This may be exacerbated in instances of mental health, where often symptoms are difficult to identify and articulate.

Consideration 11: Consideration should be given as to how to minimise health inequalities when designing CGD-based interventions. A crucial first step will be to gain a better understanding of how demographic factors (including socio-economic status) impact the different levels of engagement with, preferences for, and sharing of CGD.

5.2 Minimising privacy loss arising from sharing CGD

Digital tools and devices allow individuals to capture large quantities of data about themselves, which when shared with the health system could enable useful insights into individuals and public health. However, the potential for this data to be linked to other datasets, examined through advanced analytics and almost instantaneously distributed anywhere in the world also raises privacy concerns.

People are often prepared to accept some degree of privacy loss when there is clear benefit for them, for example having medical records transferred from a GP to a consultant with expertise in a specific condition. Similarly, consumers consent to terms of use which permit companies to sell users' data to third parties in exchange for a device or product they want, trading privacy for access.

As outlined in section 2.1 and Table 1, the sources of CGD will differ, depending on whether it is being actively generated by individuals or more passively through routine daily activities. In instances when data is not produced with health being the primary intention, e.g. social media activity, individuals may be unaware if it is being used for this purpose. Despite being collected for the shared goal of improved health, the pervasive nature of this type of data capture could engender concerns about surveillance and 'being watched'. Individuals could be harmed by merely feeling that they are being observed, even if no one is reviewing the data.

Whilst studies show that those with chronic conditions view the benefit of rapid access to information as outweighing privacy concerns [38], the trade-off may be less appealing for predictive prevention where there is no clear short-term benefit for the individual. It was suggested that as the reduction in privacy is largely inevitable, it could in this context be viewed more positively as an 'investment' by the individual in their future health. Exchanging data in order to use digital health tools could result in consumer benefits such as receiving more targeted health information earlier and/or increased opportunities for interventions.

Additionally, interviewees noted that a tension exists where individual data has implications for the health and wellbeing of others, especially when the risks involved in using that data are minimal, for example when processing aggregated de-identified data. Privacy is undoubtedly an important principle, but we accept the suspension of our privacy in other situations e.g. CCTV surveillance. Interviewees discussed the potential to create a system where the suspension of privacy is clearly stated along with reasons and benefits. Whilst obligatory data sharing may be too extreme, it highlights the importance of engendering trust and demonstrating trustworthiness so that citizens feel comfortable sharing their data for their own benefit and the benefit of future healthcare services. Privacy policies are often inaccessible and opaque, and while privacy concerns can be addressed by clarifying policies around use of CGD, this may be challenging where data is accessed via commercial companies who hold proprietary interests over these data.

Consideration 12: Privacy concerns around use of CGD may be addressed in part by clear and concise privacy policies outlining when, how and the purposes for which CGD can be collected, used, modified, retained or disposed.

5.3 Transparency and public trust

High levels of public trust in the health system underpins citizen engagement with data collection (section 5.1), data sharing (section 5.2) and addressing concerns around privacy (section 6.2). Transparency is important as a means of holding relevant stakeholders accountable and engendering public trust, hard to gain and easily lost. Widespread engagement with the prevention agenda will be supported by clarity around who is using CGD, for what purposes, and which parties will benefit.

There are inherent dangers in presenting system benefits as benefits for individuals; to claim that healthy behaviours are to the benefit of individuals in all instances may be misleading. Healthy behaviours are always good for the system as a whole, and are also of biological benefit to the individual, but there are circumstances where people cannot or do not want to enact them, and they should not be forced to in order to maximise system benefit. Instead the health system should encourage and support individuals looking to change their behaviours.

One of the advantages of personalised data is being able to more accurately assess what intervention is useful for a specific individual, including an option for ‘no intervention’, and identify circumstances and determinants of health that may render individuals unwilling or unable to engage with healthy individuals. Benefits to individuals can be achieved if these are addressed.

The health system ought to make clear from the outset the system benefits of implementing predictive prevention – such as increased efficiency and reduced costs – as well as potential benefits for individuals, including those that are system-related, for example more accurate and efficient delivery of healthcare services. Transparency will also help the public develop realistic expectations of the short- and long-term benefits and potential risks arising from using CGD for prevention. This includes the fact that individual benefits might be limited initially, even when data are shared. However, building public trust can help to foster a positive altruistic ecosystem where future generations benefit from prior data sharing.

Transparency is particularly important in instances where the individual is unaware that data is being collected. The uses, secondary uses (i.e. those uses that are not the primary purpose for which the data was collected) and limits to the use of different types of CGD should be communicated clearly to citizens to reduce privacy concerns about ‘being watched’.

One suggestion is that an opt-out system similar to the national data opt-out [39], could be an important tool to gain public trust, and to protect individuals who do not want their data to be used for predictive prevention or other purposes beyond their own individual care and treatment. Those most vulnerable and in need are likely to be the most suspicious and feel at risk in this system, concerned that their data will be used to discriminate against them. An opt-out would allow them to appease this concern, and those who are hesitant may over time become more confident about sharing their data as evidence is generated that identifies and explains the effects of participating.

Consideration 13: There should be a transparent dialogue with the public regarding the drivers for using CGD for predictive prevention, as well as the anticipated benefits and risks for the population and individuals.

Consideration 14: Models should be developed by which individuals can opt-out of their CGD being used for secondary purposes, particularly where it is identifiable. This will be dictated by the mode of data sharing.

6. Aligning commercial sector and health system needs

The generation of CGD occurs via a range of different devices, digital interactions with retailers or via online activity. These devices and services are made or operated by a large range of different commercial companies. The data is collected and stored by these individual organisations, therefore siloed and not available in a standardised format. Interviewees discussed various challenges around harnessing the data collected and held largely within the private sector.

6.1 Gaining access to data from various sources

Gaining access to CGD will depend largely on the source and the purpose of use. Some business models are dependent on gathering data about users (i.e. to sell to third parties); others may collect data as a by-product or for internal analysis to improve their services. Therefore, gaining access to individual level data from commercial companies could be a challenge for the health system. However, depending on the application, there are three models of how the health system could gain access to the datasets it requires for predictive prevention:

1. The individual becomes the primary data holder and gives permission to the health system to access it.
2. The health system bypasses the individual and initiates data sharing agreements directly with the private companies that hold the data.
3. The CGD is openly accessible, e.g. social media feeds.

It is likely that all three approaches will be useful for different applications.

The health system could enter into agreements with commercial providers to analyse the data collected, for example through DTC tests, apps or wearables, and direct at-risk individuals to healthcare services. These would require clarity about the legal bases for data processing and be supported by contractual agreements. Progress is being made in health system thinking in this area: the Department of Health and Social care recently provided an update on its framework to realise the benefits of data to patients and the NHS, which includes the announcement of a National Centre of Expertise, sitting within NHSX, that will 'provide specialist advice and guidance to the NHS on agreements for use of data' [40].

The rights to access and data portability under the General Data Protection Regulation (GDPR) are a way to obtain personal data from commercial providers, but to be within the scope of the GDPR the data must be 'personal data'.

Various organisations have championed the goal of putting individuals in control of their data and enabling them to access and control their data from multiple sources in one platform (e.g. HubofAllThings [41], MyData [42] and DECODE [43]). OpenBanking, where customers can link together all their banking data in one place in order to help them organise and understand their finances, is a possible exemplar of how health data systems could be modelled [44].

To further facilitate meaningful data gathering from social media and internet searches, more open-source data and/or increased data sharing is required. Data sharing agreements could be sought with the organisations that hold such data provided that this is done transparently and is acceptable to the public.

Consideration 15: The health system should consider how to work with the commercial sector to align their respective interests in order to facilitate appropriate sharing of CGD whilst being compliant with the GDPR, demonstrating trustworthiness and fostering and maintaining public trust. The National Centre for Expertise within NHSX could facilitate this process.

6.2 Production of mutually beneficial digital tools

The DTC market for digital tools will inevitably continue to grow in coming years, increasing the amount and type of data that consumers can collect about themselves. The tools under development are not aligned with health system needs – in this case, providing actionable information about the user in terms of risk prediction.

The health system has little control over what tools and devices are being marketed and developed by industry and it is likely to continue this way. In the DTC lifestyle market, the motivation for development is typically centred around what is technically possible and commercially viable rather than what will be most useful for the health system. However, to encourage the development of mutually beneficial tools that enable the collection of data for predictive prevention there are routes that the health system can take to actively engage with industrial partners so that devices are created to address the areas of greatest clinical need, but also reflect what consumers might want.

Consideration 16: The health system should set up a process of horizon-scanning to identify commercially available tools that have future potential in terms of utility for CGD predictive prevention strategies.

6.3 Data quality and structure

The challenges that exist with data produced within the health system such as adhering to quality and safety standards will also apply to data produced by commercial providers. The health system employs the Fast Healthcare Interoperability Resource standards, which are internationally agreed data standards to enable the sharing of clinical information between healthcare organisations and systems which are supported by guidance from NHS Digital on digital, data and technology standards [45]. However, given that much CGD is not necessarily created with the primary purpose of health care impact, standards for data collection, quality and safety will differ and will not necessarily meet those that apply to the health system.

Furthermore, much CGD will be unstructured data (e.g. free text and images) and might therefore pose additional challenges including data cleaning in order to analyse it, which could take considerable time and effort. The lack of standardisation of some types of CGD is a potential challenge for the health system to manage this type of data. Once the types of CGD that are useful have been recognised, data standards can be built to ensure that these data meet a specific clinical need.

Consideration 17: The health system should consider if and how it wants to collaborate and engage with commercial organisations to develop data standards agreements for CGD that is being collected for predictive prevention strategies using commercially available tools.

6.4 Evidence required for different applications of CGD

There is evidence to suggest that, at a population level, internet search data is a useful tool to correlate epidemiological parameters such as disease incidence with online behaviour [46]. However, there is yet to be a breakthrough study showing the efficacy of this approach at the level of an individual without substantial limitations. There have been criticisms that academic studies predicting impending ill health in individuals using this type of CGD may present findings in a way that is misleading, resulting in over interpretation of the impact and significance of the predictive accuracy [47].

Inevitably, there will be different levels of evidence required for different applications of CGD and evidence needs to be fit for purpose. For example, results from digital self-help tools may be subject to less stringent scrutiny than those from digital tools which could influence clinical decision making, including leading to additional testing. Guidance will be required on the types of evidence appropriate for different kinds of prediction models, which will partly be based on the potential harms that could arise from inaccurate predictions.

Even if the predictive models that are used on CGD are relatively accurate, sufficient evidence about the viability and efficacy of preventative interventions will also be essential. Data will need to be collected on the demographics of those who benefit from predictive prevention strategies. More clarity is needed as to what constitutes a potential ‘benefit’ and what the specific outcomes are from each strategy.

The NICE Evidence Standards Framework for Digital Technology [48] are a welcome form of guidance for developers and commissioners to understand what types of evidence are required for which types of digital tools. Similar evidence standards should be considered for enabling informed decision making around the utility of different forms of CGD in different situations.

7. Conclusions

Citizen generated data is a potentially valuable resource that could have a significant role to play in the delivery of the health system's prevention agenda. It could facilitate our understanding of disease causation through providing valuable insights into the wider determinants of health, potentially leading to evidence-based changes in policy relating to socioeconomic factors, occupational environments and local infrastructures. The variety of types of CGD and variations in how it is generated and collected is therefore an opportunity to understand what good health looks like for individuals, but also poses a challenge in terms of focusing efforts on how best to utilise this potentially rich resource to benefit future health.

7.1 A co-ordinated approach

All organisations who have a stake in using CGD for predictive prevention should co-operate to develop a co-ordinated strategy to interact with, and use, data from healthy individuals. This will include the Department of Health and Social Care, NHS England, Public Health England (and successor body or bodies) and local authorities. This will require a cultural shift in the nature of the relationship between health system and citizen – should citizens share data with the health system, it can better support individuals to stay well, not just treat them when they are unwell. Citizens will therefore need to consider the health system in a different light – as a resource to support them through a healthy life, not just to be called upon when needed to treat ill health. The engagement efforts required to achieve such a major shift in perceptions and behaviours will be crucial in order to ensure the success of CGD predictive prevention strategies, and to enhance the sustainability of the health system through a more general shift towards prevention, but will also take time, investment and patience to implement.

7.2 Focusing on the value of CGD

Identifying the additional value of CGD for specific applications and diseases, what kinds of data will be useful and where CGD can help to optimise preventative efforts will be crucial going forward. Depending on what is being proposed, potential applications will primarily impinge on different parts of the health system. The potential benefits and harms of engagement will thus be heavily dependent on context. The health system already has a small number of projects looking at how CGD could be used for prevention. The potential in these cases has been recognised, but raises questions about what the health system should do next to optimise efforts. The utility in discrete areas needs to be established in order to avoid overburdening healthcare practitioners or wasting resources.

To fully realise these benefits will require continued public engagement in the prevention agenda and trust in the health system, through long-term strategic planning and the establishment of an ongoing dialogue to provide clarity on the risks, limitations and potential benefits of using CGD for prevention. Attempts must be made to alleviate ethical concerns including fears around potential threats to privacy and narratives about personal responsibility for health. In parallel, the health system will also need to consider if and how it wants to further engage with using CGD for prevention for the benefit of all citizens, not just the technologically-savvy or the 'worried well'. Initiatives recognising the benefit of suitably supporting those who need assistance to make healthy choices may help to address some inequalities that would otherwise be exacerbated as digital health becomes more established.

7.3 Building infrastructure

Depending on the strategy the health system takes for harnessing CGD, if it chooses to do so, the broader infrastructure for capturing, storing and analysing CGD must be considered within the context of continuing digitisation developments within the health system. Adequate IT infrastructure – something that NHS is striving towards – alongside investment into platforms for the analysis of complex, diverse datasets will be required. The role of the professional health system bioinformatician could be expanded to include the analysis of CGD alongside existing clinical data. It is likely that to realise the full potential of CGD, advanced analytics, such as machine learning, will be required to make predictions based on large, complex, multifaceted datasets. The combination of datasets that can be pieced together to create digital blueprints of individuals enabling the analyser to understand the individual and therefore tailor interventions to them, has great promise, but also raises significant challenges. In order to mitigate some of the concerns about data integration and discrimination and stigmatisation, the health system will need to work closely with regulators to develop proportionate, responsive regulation and governance.

7.4 The role of the technology sector

There are opportunities for the health system to work in partnership with technology developers to ensure a joined-up approach that works for mutual benefit. This is likely to involve collaborative and novel approaches and clear agreements on data sharing. In addition, it will be necessary to support technology development from within the health system and to put measures in place to manage data.

Should the health system decide not to engage further with CGD for personalised prevention, the danger is that the pace of change within the technology sector will leave the health system behind, and that technologies will not be developed in a way that also benefits the delivery of healthcare. The health system could be left to manage the potential harms of DTC health devices, for example patients using health monitoring tools that have not been validated for medical applications but which might indicate a health problem, leading that individual to visit their GP who will need to decide the best course of action. Failure to collaborate with the technology sector could result in the health system losing the potential benefits of cooperative development, influencing strategic direction and also utilising the skills and expertise that the private sector can offer, such as digital expertise and skills in machine learning/AI. By working with the technology sector, the health system has an opportunity to influence the pace of change and development and to help bring innovative CGD approaches into healthcare for the benefit of all.

7.5 A data driven future?

While initial efforts have suggested that there is potential in using CGD for predictive prevention, this potential is yet to be realised. The health system will need to develop a coordinated approach with internal and external stakeholders about whether it wants to make use of CGD in this way, and if it decides to do so, to engage with larger-scale collection and use of CGD. The benefits of using CGD could also extend further, providing valuable insights into the wider determinants of health, potentially leading to evidence-based changes in policy relating to socioeconomic factors, occupational environments and local infrastructures.

Making the most of these opportunities will require a consistent engagement with citizens around the processes and principles of using CGD for prevention, but also to build trust between citizens and the health system. The health system will also need to engage with issues around data collection and access, management and storage, ethical and regulatory issues and collaboration with the technology sector. These developments should be considered a long-term investment, not just in the health of citizens, but also in the 'health' of the system and its ability to deliver effective, personalised healthcare well into the 21st century.

Appendices

Appendix 1: Methods

This report is an analysis informed by desk-based research using a combination of public sources of information including NHS England and UK government policy documents, grey literature and international peer-reviewed literature.

In addition, we identified and interviewed seven experts from different parts of the health system (Appendix 2) who are either actively engaged with CGD, have a specialist interest in CGD, or are working in a role where CGD could be of interest or have an impact on their way of working. The experts were chosen to help us understand the health system's view on using CGD for predictive prevention and how the health system could implement preventive strategies. Following the initial seven interviews, two additional experts also gave us insights into the ethical challenges that might arise through the increased collection and use of CGD, and to explore some of the key concerns that had arisen through our research and in the initial wave of interviews.

The interviews were semi-structured with a targeted list of questions aimed at understanding the interviewees' perspectives on the opportunities, barriers and challenges to harnessing the potential of CGD, with particular attention drawn to predictive prevention strategies. The interviewees also had the opportunity to raise further points as part of a general discussion during the interview.

Sections 3-7 reflect key themes that arose during the interviews, providing a synthesis of interviewees' answers alongside our own analysis and conclusions from internal research. The conclusions of this report are based on our internal analysis, complemented by the expert insights and advice of the interviewees.

We wish to thank all these experts for their valuable contributions to this work.

Appendix 2: Interviewees

- Dr Felix Greaves – Deputy Director, Science and Strategic Information, Public Health England
- Dr Indra Joshi – Clinical Lead for Digital Health and AI, NHS England
- Dr Tom Foley – Senior Clinical Lead for Data, NHS Digital
- Leanne Summers – Digital Strategy Delivery Lead, NHS England
- Dr Paul Southern – Chief Clinical Information Officer, Bradford Teaching Hospitals NHS Foundation Trust
- Kay Pagan – Chief Nursing Information Officer, Bradford Teaching Hospitals NHS Foundation Trust
- Dr Lamiece Hassan - HDRUK/UKRI Research Fellow, University of Manchester
- Dr Brent Mittelstadt - Senior Research Fellow in data ethics, Oxford Internet Institute, University of Oxford
- Dr Chiara Garattini – Lead User Researcher, NHS Digital

References

1. Cook, S., Raza, S. What is citizen generated data? PHG Foundation. 2018; Available from: http://www.phgfoundation.org/briefing/what-is-citizen-generated-data.
2. O'Shea, J. Digital disease detection: A systematic review of event-based internet biosurveillance systems. Int J Med Inform. 2017. 101: pp. 15-22.
3. Department of Health and Social Care Prevention is better than cure: our vision to help you live well for longer. Gov.uk. 2018; Available from: https://www.gov.uk/government/publications/prevention-is-better-than-cure-our-vision-to-help-you-live-well-for-longer.
4. NHS Long Term Plan. NHS England. 2019; Available from: https://www.longtermplan.nhs.uk/.
5. Davies, S. Chief Medical Officer annual report 2018: better health within reach. Gov.uk. 2018; Available from: https://www.gov.uk/government/publications/chief-medical-officer-annual-report-2018-better-health-within-reach.
6. Cabinet Office, Department of Health and Social Care Advancing our health: prevention in the 2020s. Gov.uk. 2019; Available from: https://www.gov.uk/government/consultations/advancing-our-health-prevention-in-the-2020s/advancing-our-health-prevention-in-the-2020s-consultation-document.
7. Picture of America: Prevention. Centers for Disease Control and Prevention 2019; Available from: https://www.cdc.gov/pictureofamerica/pdfs/Picture_of_America_Prevention.pdf.
8. Reece, A. G., Danforth, C. M. Instagram photos reveal predictive markers of depression. EPJ Data Sci. 2017. 6(15).
9. Barnett, I., Torous, J., Staples, P., et al. Relapse prediction in schizophrenia through digital phenotyping: a pilot study. Neuropsychopharmacology. 2018. 43(8): pp. 1660-1666.
10. Paparrizos, J., White, R. W., Horvitz, E. Screening for Pancreatic Adenocarcinoma Using Signals From Web Search Logs: Feasibility Study and Results. J Oncol Pract. 2016. 12(8): pp. 737-44.
11. Merchant, R. M., Asch, D. A., Crutchley, P., et al. Evaluating the predictability of medical conditions from social media posts. PLoS One. 2019. 14(6): p. e0215476.
12. Bohm, M., Reil, J. C., Deedwania, P., et al. Resting heart rate: risk indicator and emerging risk factor in cardiovascular disease. Am J Med. 2015. 128(3): pp. 219-28.
13. Shcherbina, A., Mattsson, C. M., Waggott, D., et al. Accuracy in Wrist-Worn, Sensor-Based Measurements of Heart Rate and Energy Expenditure in a Diverse Cohort. J Pers Med. 2017. 7(2).
14. Xie, J., Wen, D., Liang, L., et al. Evaluating the Validity of Current Mainstream Wearable Devices in Fitness Tracking Under Various Physical Activities: Comparative Study. JMIR Mhealth Uhealth. 2018. 6(4): p. e94.
15. Manual on Borderline and Classification in the Community regulatory framework for medical Devices. European Commission. 2017; Available from: https://ec.europa.eu/docsroom/documents/26785/attachments/1/translations/.
16. Jain, S. H., Powers, B. W., Hawkins, J. B., et al. The digital phenotype. Nat Biotechnol. 2015. 33(5): pp. 462-3.

17. Rostill, H., Nilforooshan, R., Morgan, A., et al. Technology integrated health management for dementia. Br J Community Nurs. 2018. 23(10): pp. 502-508.

18. Hive Link. British Gas. 2019; Available from: https://www.britishgas.co.uk/smart-home/hive-link.html.

19. Salathe, M. Digital epidemiology: what is it, and where is it going? Life Sci Soc Policy. 2018. 14(1): p. 1.

20. HealthMap. DiseaseDaily.org. 2019; Available from: https://www.healthmap.org/en/.

21. Grant, K., Byrne, L. Disease Detectives: Using supermarket loyalty cards to trace an E Coli outbreak. Public Health England. 2018; Available from: https://publichealthmatters.blog.gov.uk/2018/09/26/disease-detectives-using-supermarket-loyalty-cards-to-trace-an-e-coli-outbreak/.

22. Poppy, G. Twitter - norovirus model. In: Food Standards Agency presentation. Newton.ac.uk. 2015; Available from: https://gateway.newton.ac.uk/sites/default/files/asset/doc/1608/Poppy.pdf.

23. Haque, A., Milstein, A., Fei-Fei, L. Illuminating the dark spaces of healthcare with ambient intelligence. Nature. 2020. 585(7824): pp. 193-202.

24. Davidson, H. Chinese city plans to turn coronavirus app into permanent health tracker. The Guardian. 2020; Available from: theguardian.com/world/2020/may/26/chinese-city-plans-to-turn-coronavirus-app-into-permanent-health-tracker.

25. Good thinking digital wellbeing service. Healthy London Partnership. 2019; Available from: healthylondon.org/good-thinking-digital-wellbeing.

26. NHS Health Check Programme Digital Exemplar. NHS Health Check. 2019; Available from: https://www.healthcheck.nhs.uk/nhs-health-check-digital-exemplar/.

27. Government creates new National Institute for Health Protection. Department of Health and Social Care. 2020; Available from: gov.uk/government/news/government-creates-new-national-institute-for-health-protection.

28. Topol, E. The Topol Review: Preparing the healthcare workforce to deliver the digital future. Health Education England. 2019; Available from: https://topol.hee.nhs.uk/.

29. Card, C. How Facebook AI helps suicide prevention. Facebook. 2018; Available from: https://newsroom.fb.com/news/2018/09/inside-feed-suicide-prevention-and-ai/.

30. Instagram help centre - report something - self-injury. Instagram, Inc. . 2019; Available from: https://help.instagram.com/553490068054878.

31. Endeavour Partners Inside Wearables: How the science of human behavior change offers the secret to long-term engagement. Medium.com. 2017; Available from: https://medium.com/@endeavourprtnrs/inside-wearable-how-the-science-of-human-behavior-change-offers-the-secret-to-long-term-engagement-a15b3c7d4cf3.

32. van Staa, T. P., Goldacre, B., Buchan, I., et al. Big health data: the need to earn public trust. BMJ. 2016. 354: p. i3636.

33. Editorial. Cambridge Analytica controversy must spur researchers to update data ethics. Nature. 2018. (555): pp. 559-60.

34. Feng-Gu, E. Cambridge Analytica: You Can Have My Money but Not My Vote. Practical Ethics. 2018; Available from: http://blog.practicalethics.ox.ac.uk/2018/03/guest-post-cambridge-analytica-you-can-have-my-money-but-not-my-vote/.

35. Ham, C., Charles, A., Wellings, D. Shared responsibility for health: the cultural change we need. The King's Fund. 2018; Available from: https://www.kingsfund.org.uk/publications/shared-responsibility-health.

36. Widening Digital Participation. NHS Digital. 2019; Available from: https://digital.nhs.uk/about-nhs-digital/our-work/transforming-health-and-care-through-technology/empower-the-person-formerly-domain-a/widening-digital-participation.

37. Cotten, S. R., Gupta, S. S. Characteristics of online and offline health information seekers and factors that discriminate between them. Social Science & Medicine. 2004. 59(9).

38. Hale, TM, Kvedar, JC. Privacy and Security Concerns in Telehealth. Virtual Mentor. 2014. 16(12): pp. 981-985.

39. National data opt-out operational policy guidance document. NHS Digital. 2019; Available from: https://digital.nhs.uk/services/national-data-opt-out-programme/operational-policy-guidance-document.

40. Department of Health and Social Care Creating the right framework to realise the benefits for patients and the NHS where data underpins innovation. Gov.uk. 2019; Available from: https://www.gov.uk/government/publications/creating-the-right-framework-to-realise-the-benefits-of-health-data/creating-the-right-framework-to-realise-the-benefits-for-patients-and-the-nhs-where-data-underpins-innovation.

41. Hub of All Things. Medium.com. 2019; Available from: https://www.hubofallthings.com/.

42. MyData Global. mydata.org. 2019; Available from: https://mydata.org/.

43. Decode Project. Nesta. 2019; Available from: https://decodeproject.eu/.

44. Open Banking. Open Banking Ltd. 2019; Available from: https://www.openbanking.org.uk/.

45. BETA - NHS digital, data and technology standards framework. NHS Digital. 2019; Available from: https://digital.nhs.uk/about-nhs-digital/our-work/nhs-digital-data-and-technology-standards/framework.

46. Wehner, M. R., Nead, K. T., Linos, E. Correlation Among Cancer Incidence and Mortality Rates and Internet Searches in the United States. JAMA Dermatol. 2017. 153(9): pp. 911-914.

47. Gigerenzer, G. Can search engine data predict pancreatic cancer? BMJ. 2017. 358: p. j3159.

48. Evidence standards framework for digital health technologies. National Institute for Health and Care Excellence. 2019; Available from: https://www.nice.org.uk/about/what-we-do/our-programmes/evidence-standards-framework-for-digital-health-technologies.

phg
foundation
making science
work for health