# Chapter02DataSources

January 22, 2024

# 1 2 | SOURCES OF HEALTH DATA

## 1.1 Python packages used in this notebook

[1]:
```
import pandas
import wbgapi as wb
```

[2]:
```
%config InlineBackend.figure_format = 'retina'
```

[22]:
```
from plotly import express, io
```

[24]:
```
io.templates.default = 'gridon'
```

## 1.2 Introduction

Data has become a vital compass guiding our understanding and decision-making processes. From busy hospitals to the office of research laboratories, and even in our own homes, health data is being generated, collected, and analyzed. But where does all this data come from? What are the sources feeding this complex informational ecosystem?

In this chapter we explore the diverse sources of health data, examining the electronic health records that capture our interactions with the healthcare system, delving into the world of public health surveys that paint a picture of our collective health, and even exploring the data captured by wearable devices that monitor our every step and heartbeat.

We showcase the World Bank as source of open data and see how easy it can be to use Python to access this data.

The chapter concludes with a discussion the importance of these data sources in improving patient care, informing public health policies, and advancing medical research.

## 1.3 Access to health data

Any data science or research projects requires data. Information about the health of individuals and populations is no exception. Access to health data is underscored by a fundamental question: *How do we access health data?*.

Health data, pertaining specifically to access, is often divided into two categories, namely **open data** and **restricted data**. Open data is data that is freely available to the public, whereas restricted data is data that is not freely available to the public.

### 1.3.1 Restricted health data

**Restricted health data** refers to health-related information that has limitations or restrictions on its access, use, or distribution. These restrictions are often in place to protect the privacy and confidentiality of individuals, to ensure data security, and to comply with legal and ethical guidelines.

Restricted health data often includes identifiable or potentially identifiable information. This could be direct identifiers, like names or social security numbers, or indirect identifiers, like dates of birth or geographic information, which could potentially be used to identify an individual when combined with other data.

Examples of restricted health data include certain types of electronic health records (EHRs), detailed insurance claims data, and certain types of research data. For instance, data from clinical trials often has restrictions on its use to protect the privacy of the trial participants.

Access to restricted health data typically requires approval from a data access committee or a similar governing body. This process often involves a review of the proposed use of the data to ensure it is in line with the data use agreement, which outlines the terms and conditions for accessing and using the data.

In addition, those granted access to restricted health data are usually required to implement specific data security measures. This could include encrypting the data, storing it on secure servers, and limiting who has access to the data.

While the restrictions on this type of data can pose challenges for researchers and others who wish to use the data, they play a crucial role in protecting the privacy and confidentiality of individuals, which is of paramount importance in healthcare.

### 1.3.2 Open health data

**Open health data** refers to health-related data that is freely available for anyone to use, reuse, and redistribute. It's a concept that stems from the broader open data movement, which advocates for the availability of data that is free from restrictions on copyright, patents, or other mechanisms of control.

In the context of healthcare, open health data can encompass a wide variety of data types. This includes, but is not limited to, clinical data from electronic health records, demographic and epidemiological data from public health surveys, genomic data from research studies, and health expenditure data from insurance claims.

One of the key characteristics of open health data is that it is machine-readable. This means that the data is structured in a way that can be easily processed by a computer. This is important because it enables the data to be analyzed and used in applications, algorithms, and models.

Open health data has the potential to drive innovation and improve outcomes in healthcare. For example, researchers can use open health data to study patterns and trends in health conditions, healthcare providers can use it to benchmark their performance against others, and policymakers can use it to inform decisions about healthcare services and interventions.

However, while open health data has many benefits, it also presents challenges. One of the main challenges is ensuring the privacy and confidentiality of individuals. Health data often contains sensitive information, and it's important that this information is protected. This means that open

health data must be de-identified or anonymized to remove any information that could be used to identify individuals.

Another challenge is ensuring the quality and accuracy of open health data. The data must be reliable and valid to be useful. This requires robust data collection and management processes, as well as mechanisms for users to provide feedback and corrections.

Open health data is a powerful resource that can drive innovation and improve outcomes in healthcare. However, it's important that it is used responsibly, with careful consideration of privacy, confidentiality, and data quality.

## 1.4 Electronic Health Records

**Electronic Health Records** (EHRs) are digital versions of a patient's paper charts. They are real-time, patient-centered records that make information available instantly and securely to authorized users. EHRs contain a patient's medical history, diagnoses, medications, treatment plans, immunization dates, allergies, radiology images, and laboratory and test results.

The advent of EHRs marked a significant shift in healthcare information management. Before EHRs, patient information was largely stored in paper records, which made it difficult to manage and share information. EHRs have transformed the way that healthcare providers access and use information. They have the potential to provide a more holistic and up-to-date view of a patient's health, improve the quality of care, and enhance patient safety.

For instance, EHRs can help reduce errors by providing healthcare providers with accurate and complete information about a patient's health. This can help providers diagnose problems earlier and improve the overall quality of care. EHRs can also facilitate coordination of care among different healthcare providers, as they can easily share information about a patient's health.

While EHRs offer many benefits, they also present challenges. These include issues related to data privacy and security, the interoperability of different EHR systems, and the usability of EHR software. Addressing these challenges is crucial for realizing the full potential of EHRs.

Looking ahead, the future of EHRs is likely to be shaped by advancements in areas like artificial intelligence, machine learning, and predictive analytics. These technologies have the potential to further enhance the capabilities of EHRs, enabling even more personalized and effective care.

EHRs have revolutionized healthcare information management, providing a more efficient and effective way to store, access, and use health information. They offer numerous benefits, from improving the quality of care to enhancing patient safety. However, challenges related to data privacy, interoperability, and usability need to be addressed to fully leverage the potential of EHRs. With advancements in technology, the future of EHRs looks promising, with the potential for even more personalized and effective care.

### 1.4.1 Examples of EHRs

There are numerous EHR software systems available today, each with its own set of features, capabilities, and specialties. Some of the most widely used EHR systems include are listed below.

1. **Epic Systems**: Epic is one of the most well-known EHR systems and is used by many large hospitals and health systems. It offers a wide range of features, including care management, patient engagement, and population health management.

2. **Cerner**: Cerner's EHR system is used by healthcare organizations of all sizes. It offers solutions for clinical, financial, and operational needs.

3. **Allscripts**: Allscripts offers EHR solutions for practices of all sizes. It provides features like practice management, revenue cycle management, and population health management.

4. **Meditech**: Meditech's EHR system is used by a wide range of healthcare organizations, from small rural hospitals to large multi-site health systems. It offers a fully integrated suite of solutions, including clinical, financial, and administrative applications.

## 1.5 Public health surveys

Public health surveys are a critical instrument in the field of public health, providing valuable data on the health status, behaviors, and needs of populations. These surveys are designed to collect information from a representative sample of a population, and the data collected is used to infer conclusions about the health of the entire population.

### 1.5.1 Understanding the importance of public health surveys

Public health surveys play a pivotal role in informing health policies, planning health services, and monitoring and evaluating health interventions. They provide insights into the prevalence and distribution of health conditions, risk factors, health behaviors, and the utilization of health services in a population.

For instance, public health surveys can help identify health disparities among different population groups, track changes in health behaviors over time, and evaluate the impact of public health interventions. The data collected through these surveys can guide decision-making and resource allocation in public health, helping to ensure that efforts are targeted where they are most needed.

### 1.5.2 Examples of public health survey

There are numerous public health surveys conducted on a regular basis. Three are listed below.

1. **Behavioral Risk Factor Surveillance System (BRFSS)**: Conducted by the Centers for Disease Control and Prevention (CDC), the BRFSS is the largest continuously conducted health survey system in the world. It collects data on health-related behaviors, chronic health conditions, and use of preventive services from adults in the United States.

2. **National Health and Nutrition Examination Survey (NHANES)**: Also conducted by the CDC, NHANES combines interviews and physical examinations to assess the health and nutritional status of adults and children in the United States.

3. **Health Survey for England (HSE)**: The HSE is a series of annual surveys designed to monitor trends in the nation's health. It provides data on a range of health and care issues.

### 1.5.3 Challenges and future direction

While public health surveys are a valuable tool, they also present challenges. These include issues related to survey design, data collection, and data analysis. For instance, ensuring that the survey sample is representative of the population is crucial for the validity of the findings. There are also challenges related to response rates and the accuracy of self-reported data.

Looking ahead, advancements in technology offer new opportunities for public health surveys. For example, digital technologies can enable more efficient data collection and analysis, and the use of mobile devices can facilitate the collection of real-time data.

Public health surveys are a vital tool in the field of public health, providing valuable data on the health status, behaviors, and needs of populations. While they present challenges, they also offer immense potential for informing health policies, planning health services, and monitoring and evaluating health interventions. As technology continues to advance, the future of public health surveys looks promising, with the potential for even more efficient and timely data collection and analysis.

## 1.6 Wearable devices

Wearable devices, such as smartwatches, fitness trackers, and health monitors, have become increasingly popular in recent years. These devices offer a new and exciting way to collect health data, providing insights into individual and population health that were previously difficult or impossible to obtain.

Wearable devices can continuously monitor a variety of health-related parameters, including heart rate, sleep patterns, and physical activity. For example, a smartwatch might track a user's steps, heart rate, and sleep, while a wearable glucose monitor can provide real-time information on a diabetic patient's blood glucose levels.

This data can be incredibly valuable. On an individual level, it can help people monitor their own health and fitness, track changes over time, and make informed decisions about their lifestyle. For healthcare providers, the data can provide a more comprehensive view of a patient's health, supplementing traditional clinical data and potentially improving the quality of care.

### 1.6.1 Positive Uses of Wearable Device Data

1. The data collected by wearable devices has a wide range of potential uses. In healthcare, it can be used to monitor patients with chronic conditions, track the progress of rehabilitation programs, or even predict health problems before they occur. For example, a recent study found that data from a smartwatch could be used to predict episodes of atrial fibrillation, a common heart rhythm disorder.
2. In research, wearable device data can provide valuable insights into population health. For instance, researchers have used data from fitness trackers to study patterns of physical activity and sleep in different populations.
3. In public health, wearable device data could be used to track and manage the spread of diseases. During the COVID-19 pandemic, some researchers have been exploring whether data from wearable devices could help detect early signs of the virus.

### 1.6.2 Concerns with Wearable Device Data

1. Despite the potential benefits, the use of wearable device data also raises important concerns. One of the main concerns is privacy. Health data is sensitive, and there are legitimate concerns about how this data is stored, who has access to it, and how it is used.

2. Another concern is data accuracy. While wearable devices can provide a wealth of data, the accuracy of this data can vary. For instance, a fitness tracker might not be as accurate in measuring physical activity as more sophisticated equipment in a laboratory.

3. Finally, there is the issue of data interpretation. The data from wearable devices can be complex and difficult to interpret, and there is a risk that it could be misinterpreted or misused without appropriate expertise.

Wearable devices offer a new and exciting way to collect health data, with a wide range of potential uses in healthcare, research, and public health. However, it's important to address the challenges and concerns associated with this data, including privacy, data accuracy, and data interpretation. With careful management and appropriate safeguards, wearable devices have the potential to revolutionize the way we collect and use health data.

### 1.7 Sources of open health data

There are numerous open online health data repositories that provide access to a wide range of health-related data. These repositories are often maintained by government agencies, research institutions, or other organizations. They typically provide data in a variety of formats, including spreadsheets, databases, and APIs.

Note that while these repositories provide open access to their data, they may still require users to agree to certain terms and conditions to ensure the responsible use of the data.

1. ClinicalStudyDataRequest.com: This is a generalist repository that offers researchers managed access to clinical trial data. It operates on a managed access or gatekeeper model, requiring a research proposal, agreement to data use, and undergoing a review process for data access.

2. HealthData.gov: This is a U.S. government's open data site that provides access to numerous datasets related to health. The data is collected from various agencies and is available for public use.

3. Global Health Observatory data repository: This is the World Health Organization's (WHO) gateway to health-related statistics for more than 1000 indicators for its 194 Member States. The data covers themes such as mortality and burden of diseases, the Millennium Development Goals (child nutrition, child health, maternal and reproductive health, immunization, HIV/AIDS, tuberculosis, malaria, neglected diseases, water and sanitation), noncommunicable diseases and risk factors, epidemic-prone diseases, health systems, environmental health, violence and injuries, equity among others.

4. Humanitarian Data Exchange (HDX): This is an open platform for sharing data across crises and organizations. The site offers access to datasets covering health and other sectors from numerous countries around the world.

5. CDC's National Center for Health Statistics: This is the U.S. government's principal health statistics agency. It provides data on a wide range of health indicators such as birth and death rates, infant mortality, health insurance coverage, vaccination, mental health, and many others.

6. European Union Open Data Portal: This portal provides access to open data published by EU institutions and bodies. The health section includes data on public health, health care,

pharmaceuticals, and other topics.

7. UK's NHS Digital: NHS Digital is the national provider of information, data, and IT systems for health and social care in the UK. It offers a wide range of data and reports on health and social care.

8. The Cancer Imaging Archive (TCIA): TCIA is a service that de-identifies and hosts a large archive of medical images of cancer accessible for public download. The data are organized as "collections"; typically patients' imaging related by a common disease (e.g. lung cancer), image modality or type (MRI, CT, digital histopathology, etc) or research focus. Supporting data related to the images such as patient outcomes, treatment details, genomics, pathology, and expert analyses are also provided when available.

### 1.8 World Bank Open Data

The World Bank Open Data repositories serves as a good example of accessing and using open data. The World Bank Open Data Repositories provides free and open access to data about development in countries around the globe. It is a comprehensive source of data about development, including data on population, health, education, the environment, and much more.

The World Bank provides an application programming interface for their data. Many language, including Python, have packages that can be used to access the World Bank data. Once such Python package is `wbgapi`. More information about this package can be found at the World Bank Blogs page.

[6]:
```
# Dowlnoad all the World Bank indicators as a pandas dataframe assigned to the␣
↪variable series
series = wb.series.Series()
```

[7]:
```
# View the first 5 rows of the series dataframe
series.head()
```

[7]:
```
AG.AGR.TRAC.NO                        Agricultural machinery, tractors
AG.CON.FERT.PT.ZS    Fertilizer consumption (% of fertilizer produc…
AG.CON.FERT.ZS       Fertilizer consumption (kilograms per hectare …
AG.LND.AGRI.K2                                Agricultural land (sq. km)
AG.LND.AGRI.ZS                      Agricultural land (% of land area)
Name: SeriesName, dtype: object
```

[8]:
```
# Dowload all the World Bank topics as a pandas dataframe assigned to the␣
↪variable topics
topics = wb.topic.Series()
```

[9]:
```
# List the first 5 rows of the topics dataframe
topics.head()
```

[9]:
```
1    Agriculture & Rural Development
2                  Aid Effectiveness
3                   Economy & Growth
```

```
4                          Education
5                   Energy & Mining
Name: TopicName, dtype: object
```

[10]:
```
# Display the number of rows and columns in the series dataframe
series.shape
```

[10]:
```
(1478,)
```

[33]:
```
# Download the 15 most recent data points for the US population as a pandas␣
 ↪dataframe assigned to the variable us_population
us_population = wb.data.DataFrame('SP.POP.TOTL', 'USA', mrv=15) # The dataframe␣
 ↪object that is returned is in wide format
```

[34]:
```
# Transpose the us_population dataframe
us_population = us_population.T
```

[35]:
```
# View the first 5 rows of the us_population dataframe
us_population.head()
```

[35]:
```
economy          USA
YR2007   301231207.0
YR2008   304093966.0
YR2009   306771529.0
YR2010   309327143.0
YR2011   311583481.0
```

[36]:
```
# Remove the first two letters from each index
us_population.index = us_population.index.str[2:]
```

[37]:
```
# Set the index as a datetime index
us_population.index = pandas.to_datetime(us_population.index)
```

[38]:
```
# Display the first 5 rows of the us_population dataframe
us_population.head()
```

[38]:
```
economy             USA
2007-01-01  301231207.0
2008-01-01  304093966.0
2009-01-01  306771529.0
2010-01-01  309327143.0
2011-01-01  311583481.0
```

[44]:
```
# Create a line plot of the US population using the plotly express line function
express.line(
    us_population,
    title='US Population',
```

```
    labels={'index': 'Year', 'value': 'Population'}
).update_layout(
    showlegend=False
).show()
```

### 1.9 The benefits of open health data

Data has become a powerful tool that can drive significant changes in various sectors, including healthcare. Open health data, which refers to health-related data that is freely available for anyone to use, reuse, and redistribute, is playing an increasingly important role in healthcare. It has the potential to improve patient care, inform public health policies, and advance medical research.

When it comes to patient care, open health data can provide healthcare providers with a wealth of information that can help them make better decisions. For example, data from electronic health records can give providers a comprehensive view of a patient's health history, helping them diagnose conditions more accurately and prescribe treatments that are more likely to be effective. Similarly, data from wearable devices can provide real-time insights into a patient's health status, enabling providers to monitor patients remotely and intervene promptly when necessary. By providing a more complete and up-to-date picture of a patient's health, open health data can enhance the quality of care and improve patient outcomes.

Open health data can also play a crucial role in informing public health policies. By providing insights into the health status, behaviors, and needs of populations, it can help policymakers identify public health issues, understand their causes, and develop strategies to address them. For instance, data from public health surveys can reveal trends in health behaviors, such as smoking or physical activity, which can inform the development of public health campaigns. Similarly, data on the prevalence and distribution of diseases can help policymakers allocate resources more effectively, targeting areas with the greatest need.

In medical research, open health data is a valuable resource that can fuel new discoveries and innovations. Researchers can use this data to study patterns and trends in health conditions, identify risk factors for diseases, and investigate the effectiveness of treatments. For example, genomic data can help researchers understand the genetic basis of diseases, paving the way for the development of personalized treatments. Similarly, clinical trial data can provide insights into the safety and efficacy of new drugs and interventions. By making health data openly available, we can accelerate the pace of medical research and bring about advances that can benefit patients around the world.

While the potential benefits of open health data are immense, it's important to use this data responsibly. Issues related to data privacy, security, and quality need to be addressed to ensure that the use of open health data is ethical and that the data is reliable. With appropriate safeguards in place, open health data can be a powerful tool that can transform healthcare, benefiting patients, providers, policymakers, and researchers alike.

Open health data is a transformative force in healthcare. It has the potential to improve patient care, inform public health policies, and advance medical research. As we continue to generate and collect more health-related data, the role of open health data in healthcare will only become more important. The future of healthcare is data-driven, and open health data will be at the forefront of this transformation.