



ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Second cycle degree programme (LM) in Statistical Sciences

85278 – Methods and tools for Health Statistics

**85277 – Methods and tools for Official Statistics: Population
and Health Statistics**

a.a. 2022/2023

Presentation of the Course

Fabrizio Carinci

fabrizio.carinci@unibo.it

Learning outcomes

- By the end of the course, the student should have gained a fundamental understanding of the **objectives, theory and application of statistical methods** for the production of health indicators.
- The learning process will include an updated **overview of the aims and definitions** adopted internationally to monitor population health and health care, particularly at EU level.
- The course will provide detailed explanations on the statistical methods applied for the continuous improvement of health and social policies, with practical examples showing **how to produce indicators** from large databases using R software.

Course contents

- This course is intended to support the student towards a **fundamental understanding** of the societal value of health information and the methods and tools used to report, evaluate and continuously improve policies.
- In particular, the course will present solutions to current challenges involving the **use of large scale routine databases available at national and international level**. Practical cases of data analysis will be presented using relevant statistical software. Issues in the correct communication of health statistics will be also discussed.

Course contents

At the end of the course, the student should be able:

- to **calculate and interpret** health indicators used in regional, national and international reports (in particular, the EU European Core Health Indicators, European Sustainable Development Indicators and State of Health in the EU): from life expectancy to quality of care, access and efficiency measures.
- to **apply advanced techniques for health systems performance evaluation**: from risk adjustment and standardization through the use of multivariate models (generalized linear models, generalized estimating equations and multilevel models), to modern approaches using person-centered statistical models (risk prediction and stratification for population health management).
- to **apply principles of study design** (experimental vs observational) and analytical techniques (e.g. propensity scores, difference-in-difference) to **plan and evaluate health interventions and policies**, considering social determinants of health and prevention.

Course contents

Three “streams”

1. Regional, national and international health statistics
2. Risk stratification and standardization
3. Statistical methods and tools to evaluate health interventions

Stream 1

Regional, national and international health statistics

- International data sources, projects, classification and coding systems (Health status and quality of life, Surveys, Reports and Health Databases - WHO, European Commission/EUROSTAT, OECD Health at a Glance).
- Standardized health care data sources in the Italian National Health System. Covid-19 dashboards.
- Theory and applications of health systems performance assessment. The “Triple Aim” and the future of health statistics: Patient Reported Outcome Measures (PROMs) for value-based health care.
- *R Labs: analysis of health indicators*

Stream 2

Risk stratification and standardization

- Standardization methods in AHRQ and OECD indicators.
Multivariate models for complex data: GEE logistic regression.
- Healthcare performance intelligence: modern approaches for international comparisons
- *R Labs: risk adjustment and standardization techniques*

Stream 3

Statistical methods and tools to plan and evaluate health policies

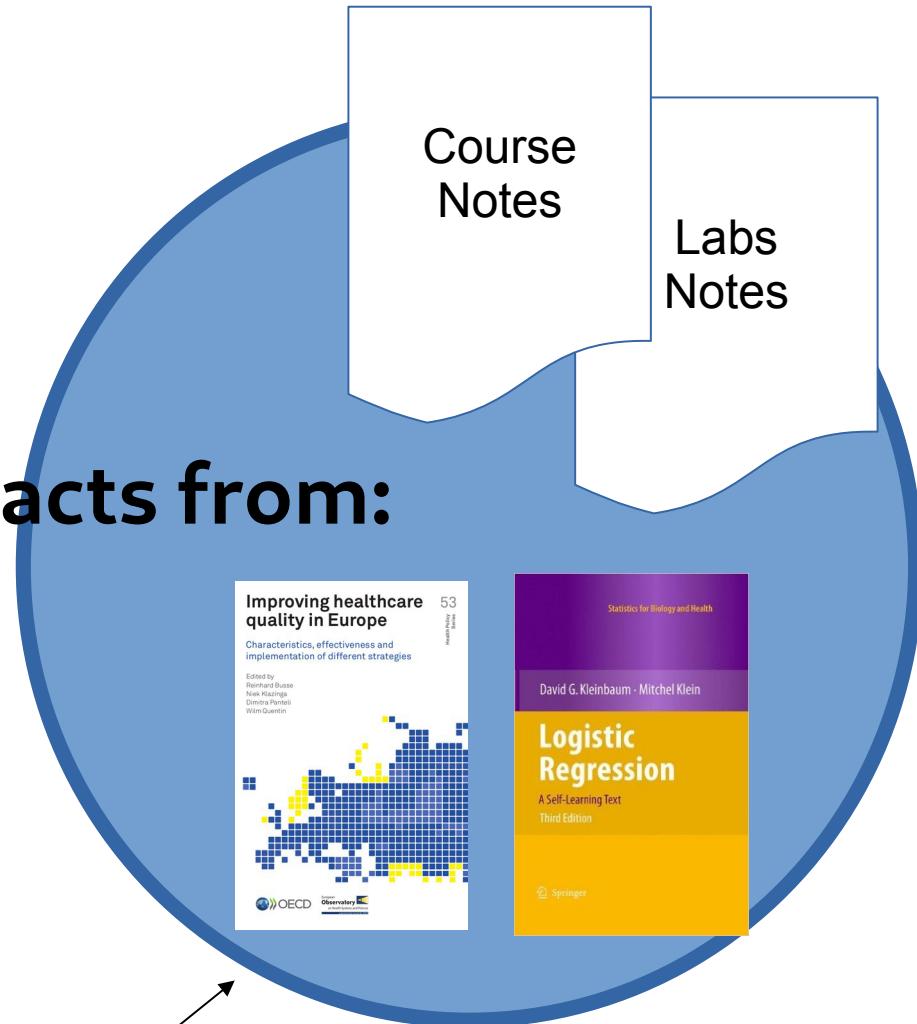
- Study design (experimental vs observational, cluster clinical trials, etc) and related statistical techniques (propensity scores, difference-in-difference, etc).
- *R Labs: propensity scores and difference-in-difference.*

Assessment methods

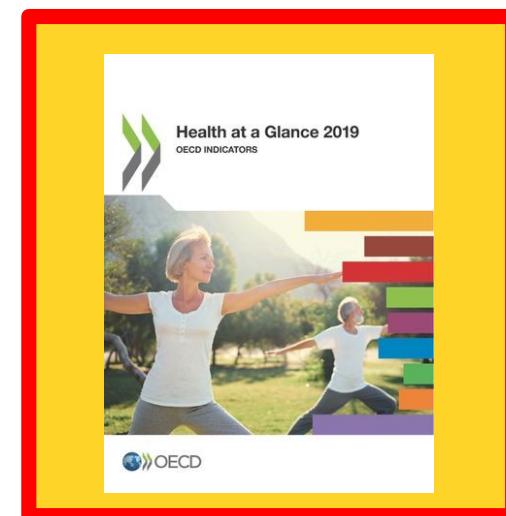
- The exam will evaluate the achievements of the student through a direct assessment of the following learning outcomes:
 - ability to identify the most suitable methods required to carry out the analysis and interpretation of health-related data
 - ability to conduct a critical appraisal of the methodology adopted in a research project or scientific publications
- The exam will consist in the presentation of a **small project of data analysis** chosen by the student, using datasets available in the public domain, followed by an **oral assessment** concerning the statistical techniques and the underlying theory related to the specific study program. An understanding of the fundamental *societal goals* relevant to population health and health policies will be also requested for the examination.

Materials for the final assessment

Extracts from:



Methods



Reports

**+ selected
papers
(applications)**



Who am I

fabrizio.carinci@unibo.it

Fabrizio Carinci

2023-today	Executive Analyst, Coordinator of the National Portal for the Transparency of Health Services, AGENAS, Rome
2017-today	Adjunct Professor, University of Bologna, Coordinator of the EUBIROD network
2020-2022	Scientific Director, COVID-19 National Portal, AGENAS, Rome
2018-2020	Principal Epidemiologist, Italian National Observatory of Patient Safety, AGENAS Visiting Professor, University of Surrey, UK, Co-Investigator EU HEALTHPROS
2015-2017	Full Professor of Health Systems and Policy, University of Surrey, UK National Delegate, OECD WP on Health Care Quality and Outcomes
2005-2014	Senior Expert, Consultant , Coordinator EU Projects BIRO, EUBIROD National Agency for Regional Health Services (AGENAS), Italian Ministry of Health (SIVEAS), Regione Umbria, Toscana, Abruzzo OECD, WHO Europe, European Commission DG-RESEARCH, DG-SANCO
2004	Senior Officer , Directorate General of Health Policy and Planning, Italian Ministry of Health, Rome, Italy
2000-2003	Associate Professor , Director of the Centre for Health Systems Research, Monash Institute of Health Services Research, Monash University, Melbourne, Australia
1997	Consultant , Harvard School of Public Health, Boston, USA
1992-2000	Senior Biostatistician , Head of the Unit of Statistics and Information Systems, Consorzio Mario Negri Sud, S.Maria Imbaro, Italy
1990	Laurea in Statistical and Economical Sciences , University of Bologna, Italy



ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Second cycle degree programme (LM) in Statistical Sciences

85278 – Methods and tools for Health Statistics

**85277 – Methods and tools for Official Statistics: Population
and Health Statistics**

a.a. 2022/2023

Stream 1. Regional, national and international health statistics

Topic 1.1.1

International data sources, projects and classifications (Health status and quality of life, Surveys, Reports and Health Databases) - WHO, European Commission, OECD)

Fabrizio Carinci
fabrizio.carinci@unibo.it

Monday, 13th February 2023

Taxonomy of available health data sources

Aims

- Research
- Routine administrative data
- National Surveys
- International Data Collections

Availability

- Raw
 - Individual data
- Pre-Processed
 - Online exploratory tools
 - Tables
 - Aggregate data
- Fully Processed
 - Papers
 - Reports

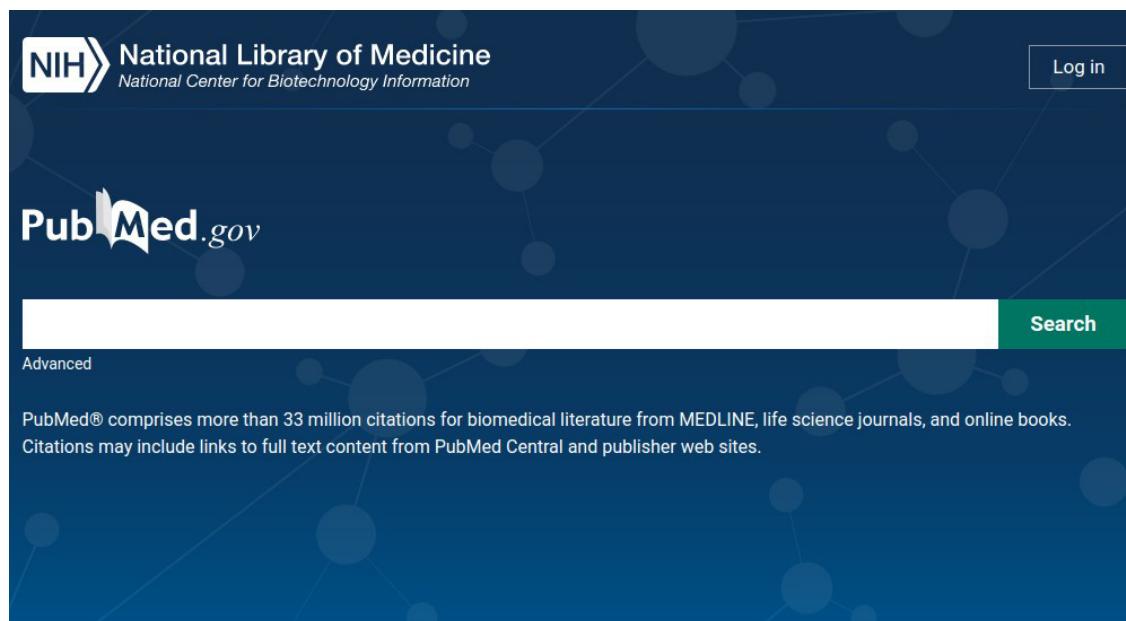
Health data sources

1. Global platforms

PubMed

<https://www.ncbi.nlm.nih.gov/pubmed>

- The main health database is Medline (PubMed), a search engine from which it is possible to extract all accredited (credible) sources of scientific evidence
- Scientific publications ordinarily deliver only fully processed data, but in many cases papers can link to supplementary files or provide essential details to access individual data



Learn

[About PubMed](#)
[FAQs & User Guide](#)
[Finding Full Text](#)



Find

[Advanced Search](#)
[Clinical Queries](#)
[Single Citation Matcher](#)



Download

[E-utilities API](#)
[FTP](#)
[Batch Citation Matcher](#)



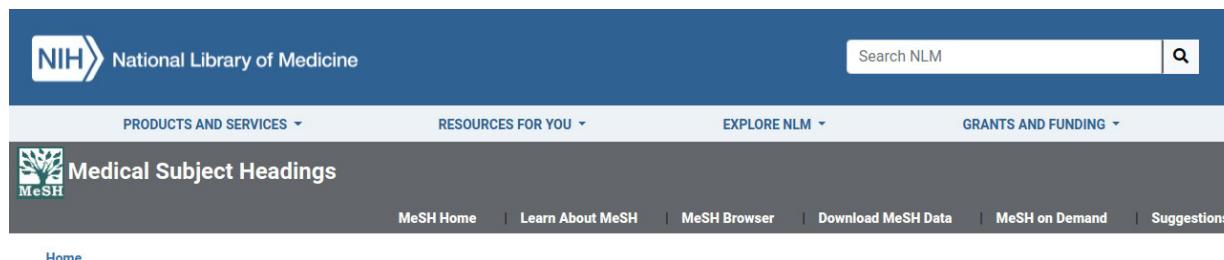
Explore

[MeSH Database](#)
[Journals](#)

Medical subject headings

<https://www.nlm.nih.gov/mesh/>

- Data standardization is essential to ensure interpretation: to make this possible at the international level, there is a need to agree on common classification criteria
- In the publication of scientific evidence, each paper, before being included in PubMed, is indexed according to a standardized taxonomy of keywords also known as “subject headings”
- MeSH is a vocabulary that helps the process of browsing the scientific literature by using an appropriate set of keywords through which papers are classified and consequently retrieved from a central archive



What's New

- Visit our [What's New](#) page to see all recent MeSH developments including the most recent ones listed below
- [2022 MeSH files are now in production](#)
 - The MeSH Browser now displays 2021 MeSH and 2022 MeSH vocabularies
 - Reports of MeSH changes are available from our [What's New](#) page
 - All 2022 MeSH files are now available via FTP download
 - [MeSH in Resource Description Format\(RDF\)](#) is now in production
 - The [downloadable file](#) contain a full representation of XML MeSH in RDF format
 - An [open MeSH API](#) is available for retrieving MeSH data
 - You can use our [SPARQL query editor](#) for querying MeSH data
 - [MeSH on Demand 2.0](#) has been re-engineered and improved in response to your

Learn About MeSH

- [Tutorials and Webinars](#)
- [MeSH Vocabulary](#)
 - [Introduction to MeSH](#)
 - [Browser Instructions](#)
 - [Finding Keywords for Publications](#)
 - [MeSH Qualifiers List](#)
 - [MeSH Qualifiers with Scope Notes](#)
 - [Publication Characteristics \(Publication Types\) with Scope Notes](#)
- [Search and Retrieval using MeSH](#)
 - [Cataloging with MeSH Terminology](#)
 - [Searching PubMed® Using MeSH Search Terms](#)
 - [PubMed® Online Training](#)

International Classification of Diseases (ICD)

- A fundamental classification that is used to record causes of death in a standardised way on a global scale is the International Classification of Diseases (ICD).
- The ICD is the international standard diagnostic classification made available by the WHO primarily to monitor **mortality trends**. Mortality is a fundamental measure of health status of a population that is frequently stratified by characteristic of interest (e.g. age, sex) and **cause of death**. Cause of death may include any disease, morbid conditions or injuries which contributed to death e.g. the circumstances of an accident that produced injuries leading to death.
- However, ICD is also increasingly used to monitor population health more broadly, through its adoption in disease diagnosis, which guides epidemiological analyses, monitoring of the incidence and prevalence of diseases, resource allocation, quality of care evaluation and production of clinical guidelines.

ICD-10

- ICD is used to classify diseases recorded on many administrative data sources e.g. death certificates and electronic health records. In addition to enabling the storage and retrieval of diagnostic information for clinical, epidemiological and quality purposes, ICD criteria form the basis for the compilation of national mortality and morbidity statistics by WHO Member States.
- **ICD-10** was endorsed by the 43rd WHO World Health Assembly in May 1990 and came into use in WHO Member States as from 1994. The classification is the latest in a series which has its origins in the 1850s. The first edition, known as the International List of Causes of Death, was adopted by the International Statistical Institute in 1893.
- WHO took over the responsibility for the ICD at its creation in 1948 when the Sixth Revision, which included causes of morbidity for the first time, was published. The World Health Assembly adopted in 1967 the WHO Nomenclature Regulations that stipulate use of ICD in its most current revision for mortality and morbidity statistics by all Member States.

Structure of ICD10

<http://apps.who.int/classifications/icd10/browse/2016/en#/I>

ICD-10 Version:2016

Search  [Advanced Search] **ICD-10** **Versions - Languages** **Info**

▼ **ICD-10 Version:2016** 

- I Certain infectious and parasitic diseases
- II Neoplasms
- III Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism
- IV Endocrine, nutritional and metabolic diseases
- V Mental and behavioural disorders
- VI Diseases of the nervous system
- VII Diseases of the eye and adnexa
- VIII Diseases of the ear and mastoid process
- IX Diseases of the circulatory system
- X Diseases of the respiratory system
- XI Diseases of the digestive system
- XII Diseases of the skin and subcutaneous tissue
- XIII Diseases of the musculoskeletal system and connective tissue
- XIV Diseases of the genitourinary system
- XV Pregnancy, childbirth and the puerperium
- XVI Certain conditions originating in the perinatal period
- XVII Congenital malformations, deformations and chromosomal abnormalities
- XVIII Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified
- XIX Injury, poisoning and certain other consequences of external causes
- XX External causes of morbidity and mortality
- XXI Factors influencing health status and contact with health services
- XXII Codes for special purposes

Chapter I
Certain infectious and parasitic diseases
(A00-B99)

Incl.: diseases generally recognized as communicable or transmissible

Use additional code (U82-U84), if desired, to identify resistance to antimicrobial drugs

Excl.: carrier or suspected carrier of infectious disease ([Z22.-](#))
certain localized infections - see body system-related chapters
infectious and parasitic diseases complicating pregnancy, childbirth and the puerperium [except obstetrical tetanus] ([O98.-](#))
infectious and parasitic diseases specific to the perinatal period [except tetanus neonatorum, congenital syphilis, perinatal gonococcal infection and perinatal human immunodeficiency virus [HIV] disease] ([P35-P39](#))
influenza and other acute respiratory infections ([J00-J22](#))

This chapter contains the following blocks:

- [A00-A09](#) Intestinal infectious diseases
- [A15-A19](#) Tuberculosis
- [A20-A28](#) Certain zoonotic bacterial diseases
- [A30-A49](#) Other bacterial diseases
- [A50-A64](#) Infections with a predominantly sexual mode of transmission
- [A65-A69](#) Other spirochaetal diseases
- [A70-A74](#) Other diseases caused by chlamydiae
- [A75-A79](#) Rickettsioses
- [A80-A89](#) Viral infections of the central nervous system
- [A92-A99](#) Arthropod-borne viral fevers and viral haemorrhagic fevers
- [B00-B09](#) Viral infections characterized by skin and mucous membrane lesions
- [B15-B19](#) Viral hepatitis
- [B20-B24](#) Human immunodeficiency virus [HIV] disease

ICD-9-CM

- The International Classification of Diseases, 9th Revision, Clinical Modification (**ICD-9-CM**), based on WHO ICD-9, is also important for its use in Hospital Discharge Abstracts in many countries.
- **ICD-9-CM** is the official system of assigning codes to diagnoses and procedures associated with hospital utilization in the United States (used for DRG payments). The ICD-9 was used to code and classify mortality data from death certificates until 1999, when use of ICD-10 for mortality coding started.
- The ICD-9-CM consists of:
 - a tabular list containing a numerical list of the disease code numbers in tabular form
 - an alphabetical index to the disease entries
 - a classification system for surgical, diagnostic, and therapeutic procedures (alphabetic index and tabular list).
- The National Center for Health Statistics (NCHS) and the Centers for Medicare and Medicaid Services are the U.S. governmental agencies responsible for overseeing all changes and modifications to the ICD-9-CM

WHO Global Health Observatory

<http://www.who.int/gho/en/>

- The World Health Organization (WHO) collects data from all the 193 Member countries.

The screenshot shows the homepage of the WHO Global Health Observatory. At the top, there is a black navigation bar with links for 'Global' and 'Regions'. On the right side of the bar are a search icon, a language selection icon, and a 'Select language' dropdown. Below the bar, the WHO logo is displayed next to the text 'World Health Organization'. The main header features the text 'THE GLOBAL HEALTH OBSERVATORY' in large, bold, orange capital letters, followed by the subtitle 'Explore a world of health data' in white. Below this, there are two buttons: 'Indicators' and 'Countries'. The background of the page has a dark purple gradient with abstract, glowing blue and yellow lines forming a network-like pattern. At the bottom of the page, there are three smaller images with captions: 'Universal Health Coverage' (a woman holding a document), 'Health Emergencies' (two people in medical scrubs), and 'Health and Well-Being' (a woman smiling with a bowl of fruit). A banner at the very bottom of the page reads 'Coronavirus disease (COVID-19) data'.

WHO Global Health Observatory

<http://www.who.int/gho/en/>

- The **Global Health Observatory (GHO)** provides access to data and analyses for monitoring the global health situation (reports are downloadable from 2005). The GHO issues analytical reports on the current situation and trends for priority health issues.
- A key output of the GHO is the annual publication *World Health Statistics*, which compiles statistics for key health indicators on an annual basis. In addition, the GHO provides analytical reports on cross-cutting topics such as the report on women and health and burden of disease. Lastly, the GHO provides the link to specific disease or programme reports with a strong analytical component.
- A specific section, the **Global Health Infobase**, delivers data on non-communicable diseases (NCDs) and their risk factors for all WHO Member States.

WHO European Health for All Database

<https://gateway.euro.who.int/en/datasets/european-health-for-all-database/>

- The WHO Regional Office for Europe (based in Copenhagen) coordinates information from 53 Member States, covering an area from Lisbon to Vladivostock

European Health Information Gateway | Health for All explorer

Gateway > Health for All explorer

Selected indicators ^

X Indicator not selected.

Y Indicator not selected.

Same size for all bubbles.

Highlight options ▾

Select indicators (1503/1503) ▾

Select countries (0/53) ▾

Show average value (1/11) ▾



Integrated data
Health for All databases in one place

Explore
in bubble charts, maps and line charts

Interact
see dynamic changes over time and geographic place

Search
find indicators by topic or by title

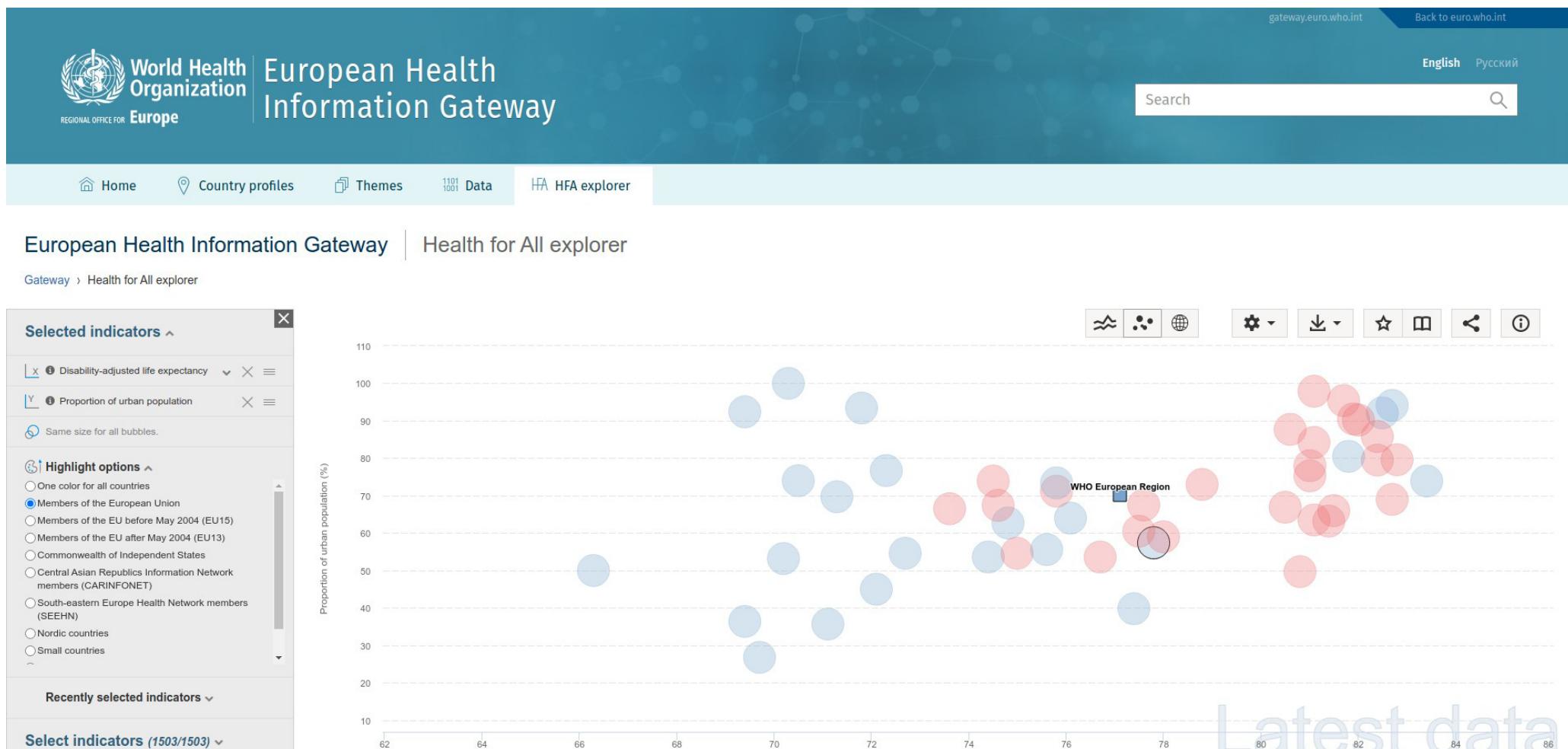
Export
data, metadata, graphs or share online

Want to know more?

TAKE A TOUR

WHO Health for All Tool

<https://gateway.euro.who.int/en/hfa-explorer/>



WHO Europe Health Information Gateway

<https://gateway.euro.who.int/en/>

The screenshot shows the homepage of the WHO Europe Health Information Gateway. At the top, there is the WHO logo and the text "World Health Organization" and "REGIONAL OFFICE FOR Europe". On the right, there are links for "English" and "Русский". Below the header is a navigation bar with links for "Home", "Health topics", "Countries", "Publications", "Data and evidence" (which is highlighted), "Media centre", and "About us". The main title "European Health Information Gateway" is displayed prominently in the center, with the subtitle "A wealth of information at your fingertips" below it. A search bar is located above a section titled "or browse by:". This section includes four categories: "COUNTRY PROFILES", "THEMES", "DATA", and "HFA EXPLORER". To the left, there is a sidebar titled "UPDATES" containing news items such as "Update of European Health for All databases" (09 December 2021), "Seasonal influenza vaccination policies and coverage has been updated" (26 November 2020), "Mortality & Health for All databases now updated with latest available data" (10 September 2020), and "Now available: Data of the HBSC 2018 report" (05 June 2020). To the right, there is a large data visualization titled "State of "Health 2020"" showing various health indicators for the WHO European Region. The indicators include Alcohol consumption (L per capita), Tobacco use (%), Overweight (%), Obesity (%), Premature mortality (Deaths/100 000), and Mortality from external causes (Deaths/100 000). The data is presented in a grid of scatter plots with red lines highlighting trends.

OECD.Stat

<http://stats.oecd.org/>

- The OECD maintains a large health database of aggregate data from 39 contributing Member States, including all more industrialised countries

The screenshot shows the OECD.Stat interface. On the left, a navigation menu includes categories like Globalisation, Health, and Health Care Utilisation, with 'Hospital discharges by diagnostic categories' highlighted. The main area displays a data grid titled 'Health Care Utilisation : Hospital discharges by diagnostic categories'. The grid has columns for Year (1990-1998) and Country (Australia, Austria, Belgium, Canada, Chile, Czech Republic), and rows for Variable (Infectious and parasitic diseases by diagnostic categories) and Measure (Number). A legend indicates a 'Break' symbol. To the right, there's an 'Information' panel with links to Source, Contact person/organisation, and email (health.contact@oecd.org), along with update details and source definitions. Arrows at the bottom point to 'Indicators' (left), 'Data' (center), and 'Sources' (right).

Variable	Measure	Country	Year								
			1990	1991	1992	1993	1994	1995	1996	1997	1998
Infectious and parasitic diseases by diagnostic categories	Number	Australia	55 566	56 137	62 477	60 806	61 446	65 9
		Austria	37 103	36 363	38 033	39 067	41 915	41 713	42 774	48 023	48 7
		Belgium
		Canada	50 142	52 334	50 467	49 779	(B) 50 609	48 539	45 326	44 906	47 0
		Chile
		Czech Republic									

Health data sources

2. European platforms

- The Statistical Office of the European Union (EUROSTAT) aims to provide the European Union with Official Statistics that enable comparisons between countries and regions. Data are published in Volumes and Bulletins
- Health statistics and databases refer to two main topics:
 - **Public health database.** It addresses citizens' needs for reliable information and high level knowledge at European level. Member States agreed to collect datasets ensuring high quality and comparability of evidence to support policy decisions in Europe.
 - **Health and safety at work.** Data are collected with the aim of creating more jobs and of better quality. A safe and healthy working environment is an essential element of the quality of work.

Public health tables have been divided into three main domains:

1. Health status indicators from surveys

tables on self-perceived health, life styles and restrictions. Data on health conditions also play a role in the calculation of the healthy life years expectancy. This collection includes also tables on employment of disabled persons based on the Labour Force survey.

2. Health care

- health care expenditure, i.e. data by type of provider (e.g. hospitals, general practitioners), function (products and services) and financing agent (e.g. social security, private insurance company, household)
- resources: including human (physicians, dentists, etc) as well as hospital statistics (hospital beds, surgical procedures in hospitals, hospital discharges by disease).
- health care indicators from surveys: tables on perceived unmet needs for medical or dental care, consultations of health care professionals, hospitalisation, cancer screening, etc.

3. Causes of death

- Eurostat disseminates statistics according to a shortlist of 65 causes ('Causes of death European shortlist', based on the ICD International Statistical Classification of Diseases and Related Health Problems, WHO). Data are available at national and regional level (NUTS 2) for total number, crude death rates (CDR) and standardized death rates (SDR), broken down by age groups and by sex.

**HEALTH****MAIN TABLES**[Overview](#)[▲ Data](#)**MAIN TABLES**[Database](#)[Publications](#)[Methodology](#)[Legislation](#)[Statistics illustrated](#)[Links](#)

- └ **Health (t_hlth)**
 - └ **Health status (t_hlth_state)**
 - ZIP Healthy life years and life expectancy at birth, by sex (tps00150) [M](#) [i](#)
 - ZIP Healthy life years and life expectancy at age 65 by sex (tepsr_sp320) [M](#) [i](#)
 - ZIP Share of people with good or very good perceived health by sex (sdg_03_20) [M](#) [i](#)
 - └ **Health care (t_hlth_care)**
 - ZIP Practising physicians (tps00044) [i](#)
 - ZIP Licensed physicians (tps00167) [i](#)
 - ZIP Practising dentists (tps00045) [i](#)
 - ZIP Physicians or doctors by NUTS 2 regions (tgs00062) [i](#)
 - ZIP Dentists by NUTS 2 regions (tgs00063) [i](#)
 - ZIP Hospital beds (tps00046) [i](#)
 - ZIP Curative care beds in hospitals (tps00168) [i](#)
 - ZIP Psychiatric care beds in hospitals (tps00047) [i](#)
 - ZIP Available beds in hospitals by NUTS 2 regions (tgs00064) [i](#)
 - ZIP Discharges from hospitals (tps00048) [i](#)
 - ZIP Self-reported unmet need for medical examination and care by sex (sdg_03_60) [M](#) [i](#)
 - └ **Causes of death (t_hlth_cdeath)**
 - ZIP Causes of death, by sex (tps00152) [i](#)
 - ZIP Death due to cancer, by sex (tps00116) [i](#)
 - ZIP Death due to other ischaemic heart diseases, by sex (tps00119) [i](#)
 - ZIP Death due to suicide, by sex (tps00122) [i](#)
 - ZIP Death due to accidents, by sex (tps00125) [i](#)
 - ZIP Death due to transport accidents, by sex (tps00165) [i](#)
 - ZIP Death due to pneumonia, by sex (tps00128) [i](#)
 - ZIP Death rate due to chronic diseases by sex (sdg_03_40) [M](#) [i](#)
 - ZIP Death due to chronic liver disease, by sex (tps00131) [i](#)
 - ZIP Death due to diseases of the nervous system, by sex (tps00134) [i](#)
 - ZIP Death due to diabetes mellitus, by sex (tps00137) [i](#)
 - ZIP Death due to alcoholic abuse, by sex (tps00140) [i](#)
 - ZIP Death due to AIDS (HIV-disease), by sex (tps00143) [i](#)
 - ZIP Death due to homicide, assault, by sex (tps00146) [i](#)
 - ZIP Death due to drugs dependence, by sex (tps00149) [i](#)
 - ZIP Suicide death rate by age group (tps00202) [i](#)
 - ZIP All causes of death by NUTS 2 regions (tgs00057) [M](#) [i](#)
 - ZIP Death due to cancer by NUTS 2 regions (tgs00058) [M](#) [i](#)
 - ZIP Death due to ischaemic heart diseases by NUTS 2 regions (tgs00059) [M](#) [i](#)
 - ZIP Death due to accidents by NUTS 2 regions (tgs00060) [M](#) [i](#)
 - ZIP Death due to transport accidents by NUTS 2 regions (tgs00061) [M](#) [i](#)

More detailed information available in each data section:

Statistics on **health status and health determinants** focus on various aspects of health status of population and its non-medical determinants, life styles and health behaviour.

Sources:

- **EU Statistics on Income and Living Conditions (EU-SILC)** - annual: self-perceived health, limitations, etc
- **European Health Interview Survey (EHIS)** - every 5 years: health status, care, determinants

The screenshot shows the Eurostat homepage with a navigation bar at the top. The main content area is titled "Your key to European statistics". Below the title, there are five tabs: News, Data, Publications, About Eurostat, and Help. The "Data" tab is selected. A breadcrumb navigation path is visible: European Commission > Eurostat > Health > Data > Database. On the left, there is a sidebar with sections for HEALTH (Overview, Data, Main tables, DATABASE, Publications, Methodology, Legislation, Statistics illustrated, Links) and DATABASE (Health (hlth), Health determinants (hlth_det), Health care (hlth_care), Disability (hlth_dsb), Causes of death (hlth_cdeath), Health and safety at work (hsw)). The right side displays a hierarchical tree view of the database structure under the "DATABASE" tab, with categories like Health (hlth), Health determinants (hlth_det), and Health care (hlth_care). Each category has several sub-items listed.

EUROSTAT

<https://ec.europa.eu/eurostat/web/health/data>

More detailed information available in each data section:

Statistics on **health care** focus on various aspects of health care systems

Sources:

- **Joint OECD-Eurostat-WHO Europe-WHO System of Health Accounts (SHA) Data Collection – annual:** health care use and expenditure
- **EU Statistics on Income and Living Conditions (EU-SILC)** – annual: unmet needs etc.
- **European Health Interview Survey (EHIS)** – every 5 years

The screenshot shows the Eurostat website's 'Database' section for the 'Health' category. The left sidebar has 'HEALTH' selected. Under 'DATA', there are links to 'Main tables', 'DATABASE', 'Publications', 'Methodology', 'Legislation', 'Statistics illustrated', and 'Links'. The main content area shows a hierarchical tree view of datasets under 'DATABASE'. The tree starts with 'Health (hlth)' and branches into 'Health status (hlth_state)', 'Health determinants (hlth_det)', 'Health care (hlth_care)', 'Health care expenditure (SHA 2011) (hlth_sha11)', 'Health care resources (hlth_res)', 'Health care staff (hlth_staff)', 'Health care facilities (hlth_facil)', 'Health care activities (hlth_act)', and several specific datasets like 'Hospital discharges and length of stay for inpatient and curative care (hlth_co_dischl)', 'Curative care bed occupancy rate (hlth_co_bedoc)', 'Non-residents among all hospital discharges, % (hlth_co_dischnr)', 'Hospital discharges - national data (hlth_hosd)', 'Hospital discharges - regional data (hlth_hosd_r)', 'Length of stay in hospital (hlth_hostay)', 'Operations, procedures and treatment (hlth_oper)', 'Consultations (hlth_consult)', 'Preventive services (hlth_prev)', 'Medicine use (hlth_med)', 'Home care and help (hlth_home)', 'Unmet needs for health care (hlth_unm)', 'Disability (hlth_dsb)', 'Causes of death (hlth_cdeath)', and 'Health and safety at work (hsw)'. Most datasets have a small 'M' icon next to them, indicating they are metadata files.

More detailed information available in each data section:

Disability statistics

Sources:

- **EU SILC – annual:** limitations due to health problems, social inclusion, living conditions, etc
- **EHIS - every 5 years:** level of functioning and activity limitations in the population and provides other information on health status, health determinants and health care use.
- **European Health and Social Integration Survey (EHSIS) 2012-13:** barriers to participation in different life areas
- **Labour Force Survey (LFS) 2002, 2011:** disabled persons on the labour market
- **European System of Integrated Social Protection Statistics (ESSPROS):** disability benefits

The screenshot shows the Eurostat homepage with a navigation bar at the top. The main content area has two tabs: 'HEALTH' and 'DATABASE'. The 'DATABASE' tab is selected, showing a hierarchical tree view of datasets under the 'Health' category. The tree includes 'Health (hlth)', 'Disability (hlth_dsb)', and several sub-folders like 'Health status (hlth_state)' and 'Health determinants (hlth_det)'. Each folder has a small icon and a link to its details.

More detailed information available in each data section:

Statistics on **causes of death** (COD) provide information on mortality patterns and form a major element of public health information.

- WHO's definition: COD is "the disease or injury which initiated the train of morbid events leading directly to death, or the circumstances of the accident or violence which produced the fatal injury".

Sources:

- Data collection from Member States - Annual.** Datasets using ICD10 available at national and regional level since 2011

The screenshot shows the Eurostat website interface. At the top, there is a navigation bar with links for 'Sign In | Register', 'Legal notice | RSS | Cookies | Links | Contact', and a language selector set to 'English'. Below the navigation bar is a search bar with the placeholder 'Type a keyword, a publication title, a dataset title...'. The main content area has a header 'eurostat Your key to European statistics'. Below the header, there are five main menu items: 'News', 'Data', 'Publications', 'About Eurostat', and 'Help'. Under the 'Data' menu, a sub-menu for 'European Commission > Eurostat > Health > Data > Database' is selected. The 'DATABASE' section is currently active, indicated by a blue bar. On the left, there is a sidebar with categories: 'HEALTH' (Overview, Data, Main tables, DATABASE, Publications, Methodology, Legislation, Statistics illustrated, Links) and 'DATABASE' (Health (hlth), General mortality (hlth_cd_gnmr), Infant mortality (hlth_cd_imor), Peri-neonatal mortality (hlth_cd_pnmor), Public health themes (hlth_cd_pbt)). The 'DATABASE' section displays a hierarchical tree view of datasets under 'Health (hlth)'. The tree includes nodes for 'Health status (hlth_state)', 'Health determinants (hlth_det)', 'Health care (hlth_care)', 'Disability (hlth_dsb)', 'Causes of death (hlth_cdeath)', 'General mortality (hlth_cd_gnmr)', 'Infant mortality (hlth_cd_imor)', 'Peri-neonatal mortality (hlth_cd_pnmor)', 'Public health themes (hlth_cd_pbt)', and 'Health and safety at work (hsw)'. Each node is accompanied by a small icon and a link.

- The collection **Health and safety** at work has been divided into two domains:
 - Accidents at work (ESAW European Statistics on Accidents at Work): accidents at the workplace or in the course of an occupational activity from declarations to the insurance or to another competent authority . ESAW statistics cover non-fatal accidents at work with more than 3 days of absence as well as fatal accidents at work. Data are available at national level for total number and incidence rates (per 100 000 employed workers), broken down by age groups, sex and economic activity of the employer.
 - Work-related accidents and health problems. To complement the administrative data, ad hoc modules on health and safety at work outcomes are carried out. These aim to cover: groups that are not comprehensively included in the administrative statistics (e.g. self-employed, the public sector), less severe accidents (less than 4 days of absence), and work-related diseases not recognized by the authorities.

EUROSTAT

<https://ec.europa.eu/eurostat/web/health/data>

More detailed information available in each data section:

Health and Safety at Work statistics accidents at work, work-related health problems and exposure to risk factors.

- In partnership with National Statistical Offices, Social Security Institutions and Ministries of Labour Affairs.

Sources:

- **European statistics on accidents at work – annual administrative data collection**
- **Labour Force Survey 1999, 2007 and 2013:** accidents at work, work-related health problems and exposure to risk factors
- Art.153 Treaty, Commission Regulation (EU) No 349/2011 on statistics on accidents at work

The screenshot shows the Eurostat website interface. At the top, there is a navigation bar with links for 'Sign In | Register', 'Legal notice', 'RSS', 'Cookies', 'Links', 'Contact', and a language selector set to 'English'. Below the navigation bar is a search bar with the placeholder 'Type a keyword, a publication title, a dataset title...'. The main content area features a banner with the Eurostat logo and the text 'Your key to European statistics'. Below the banner, there is a breadcrumb navigation path: 'European Commission > Eurostat > Health > Data > Database'. The page is divided into two main sections: 'HEALTH' on the left and 'DATABASE' on the right. The 'HEALTH' section contains links for 'Overview', 'Data' (with 'Main tables' and 'DATABASE' sub-links), 'Publications', 'Methodology', 'Legislation', 'Statistics illustrated', and 'Links'. The 'DATABASE' section displays a hierarchical tree view of datasets under 'Health (hlth)' and 'Health and safety at work (hsw)'. The 'Health (hlth)' category includes 'Health status (hlth_state)', 'Health determinants (hlth_det)', 'Health care (hlth_care)', 'Disability (hlth_dsb)', and 'Causes of death (hlth_cdeath)'. The 'Health and safety at work (hsw)' category is expanded, showing numerous sub-datasets such as 'Accidents at work (ESAW, 2008 onwards) (hsw_acc_work)', 'Main indicators (hsw_mi)', 'Details by NACE Rev. 2 activity (2008 onwards) (hsw_n2)', 'Causes and circumstances of accidents at work (ESAW Phase III) (hsw_ph3)', 'Accidents at work (ESAW) - until 2007 (hsw_acc7_work)', and many others related to traffic accidents, enterprise size, and employment status.

European Health Interview Survey (EHIS)

- The European Health Interview Survey (EHIS) is a major source of information for the calculation of health indicators, consisting of four modules on health status, health care use, health determinants and socio-economic background variables. EHIS targets the population aged at least 15 and living in private households.
- The four modules cover the following topics:
 - Background variables on demography and socio economic status such as sex, age, household type, etc.
 - Health status such as self-perceived health, chronic conditions, limitation in daily activities, disease specific morbidity, physical and sensory functional limitations, etc.
 - Health care use such as hospitalisation, consultations, unmet needs, use of medicines, preventive actions, etc.
 - Health determinants e.g. height and weight, consumption of fruits, smoking, alcohol consumption, etc.

European Core Health Indicators

<https://webgate.ec.europa.eu/dyna/echi/>

- The European Commission also maintains own specialised public health portal, where the results of funded projects are permanently stored and followed up in their implementation, e.g. the ECHI project.

European Core Health Indicators (ECHI)

Click on the button "Choose your indicator" to navigate through the ECHI data tool

The screenshot shows the ECHI Data Tool interface. At the top right is a button labeled "Choose your indicator(s)". On the left, under "Select Indicators in chapters", there are two main categories: "Demographic and Socio-economic factors indicators" and "Health Status indicators". Under "Health Status indicators", several sub-indicators are listed, each with a checkbox and an "ECHI" button. In the center, there is a large empty area with a vertical scrollbar. On the right, there are two columns: "Select Countries" and "Select Years". The "Select Countries" column lists countries with checkboxes: Belgium, Bulgaria, Czech Republic, Germany, Estonia, Greece, Spain, France, and Cyprus. Most checkboxes are checked. The "Select Years" column has two options: "2008" and "all", with "2008" checked. There is also a small upward arrow icon at the top right of the interface.

ECHI Data Tool

Choose your indicator(s)

Select Indicators in chapters

Reset

Demographic and Socio-economic factors indicators

Health Status indicators

- Asthma: self-reported prevalence
- Chronic obstructive pulmonary disease (COPD): self-reported prevalence
- Dementia / Alzheimer : Estimated number of people with dementia (Age group 30-95+)
- Physical and sensory functional limitations
- Depression: self-reported prevalence
- Diabetes: self-reported prevalence
 - Proportion of persons reporting diagnosed diabetes in the past 12 months (ECHI)
 - Proportion of men reporting diagnosed diabetes in the past 12 months (ECHI)
 - Proportion of women reporting diagnosed diabetes in the past 12 months (ECHI)
 - Proportion of persons reporting diagnosed diabetes in the past 12 months whose highest completed (ECHI)

Select Countries

Select Years

2008

all

OECD Health at a Glance: Europe

<https://www.oecd.org/health/health-at-a-glance-europe/>

- Relevant recent developments **commissioned** by the EU to the OECD. Reports published in a *cycle of two years* in collaboration between the European Commission, the Organization for Economic Cooperation and Development (OECD), the European Observatory on Health Systems and Policies and EU Member States. **State of Health in the EU** Health at a Glance Report published every *even* year. State of Health in the EU “**Country Health Profiles**” published every *odd* year. Indicators included in the flagship OECD publication “Health at a Glance” used as a major source

The image shows the homepage of the **Health at a Glance: Europe 2020** website. The top navigation bar includes links for **OECD.org**, **Data**, **Publications**, **More sites**, **News**, and **Job vacancies**. A search bar with "Google Custom" and "A to Z" options is also present. The main menu features **OECD Home**, **About**, **Countries**, **Topics** (with a dropdown for "Coronavirus (COVID-19)"), and a large green button for "Coronavirus (COVID-19)". Below the menu, a breadcrumb trail shows [OECD Home](#) > [Health](#) > [Health at a Glance: Europe 2020](#).

The main content area features a large image of a woman wearing a blue surgical mask, overlaid with several stylized white COVID-19 virus particles. The text "State of Health in the EU Cycle" is displayed prominently. A paragraph discusses the challenges posed by the coronavirus pandemic, mentioning the need for effective testing, tracing, and isolation policies. A "READ MORE" button with a right-pointing arrow is located at the bottom right of this section.

To the right of the main content, there is a vertical sidebar with a green background. It features the **European Commission** logo, a circular icon with a map of Europe, the text "State of Health in the EU Italy Country Health Profile 2021", and three smaller circular icons representing different health metrics. At the bottom of this sidebar, the **OECD** and **European Observatory** logos are visible.

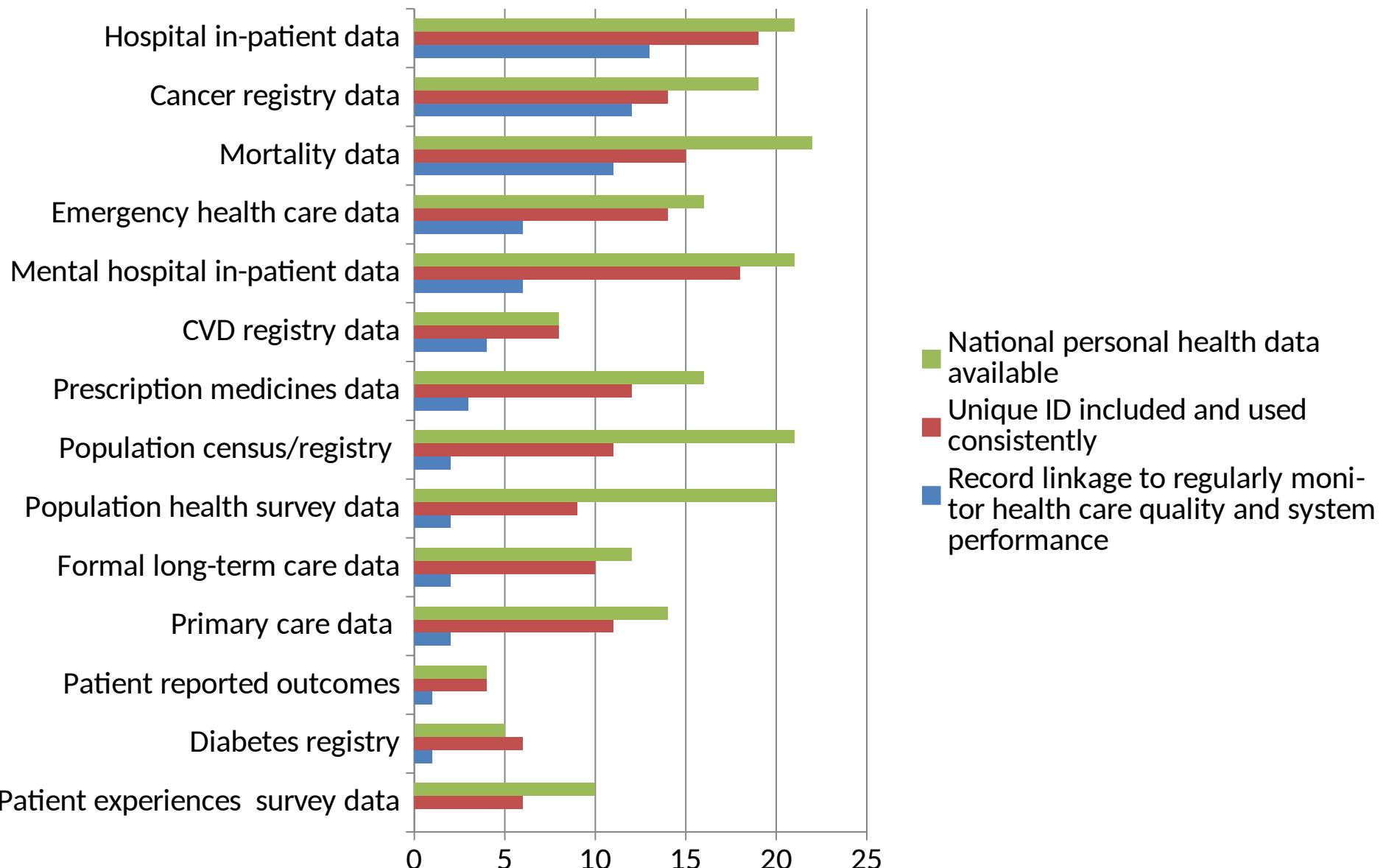
A thumbnail image of the **Health at a Glance: Europe 2020** report cover is shown on the right. The cover has a blue background and features the title "Health at a Glance: Europe 2020 STATE OF HEALTH IN THE EU CYCLE" along with the **OECD** and **European Observatory** logos. Below the title, there is a small image of a woman wearing a mask and a series of interconnected icons representing various health and economic indicators.

Health data sources

3. National experiences

OECD 2015 - Health Information Infrastructure

State of play of data linkage



Population Data Science

International Journal of Population Data Science (2018) 3:4

International Journal of Population Data Science

Journal Website: www.ijpds.org



A Position Statement on Population Data Science: The Science of Data about People

Kimberlyn M McGrail^{1*}, Kerina Jones², Ashley Akbari², Tellen D Bennett³, Andy Boyd⁴, Fabrizio Carinci⁵, Xinjie Cui⁶, Spiros Denaxas⁷, Nadine Dougal⁸, David Ford⁹, Russell Kirby¹⁰, Hye-Chung Kum¹¹, Rachael Moorin¹², Ros Moran¹³, Christine M O'Keefe¹⁴, David Preen¹⁵, Hude Quan⁹, Claudia Sanmartin¹⁶, Michael Schull¹⁷, Mark Smith¹⁸, Christine Williams¹⁹, Tyler Williamson⁹, Grant MA Wyper²⁰, and Milton Kotelchuk²¹

Abstract

Information is increasingly digital, creating opportunities to respond to pressing issues about human populations using linked datasets that are large, complex, and diverse. The potential social and individual benefits that can come from data-intensive science are large, but raise challenges of balancing individual privacy and the public good, building appropriate socio-technical systems to support data-intensive science, and determining whether defining a new field of inquiry might help move those collective interests and activities forward. A combination of expert engagement, literature review, and iterative conversations led to our conclusion that defining the field of Population Data Science (challenge 3) will help address the other two challenges as well. We define Population Data Science succinctly as *the science of data about people* and note that it is related to but distinct from the fields of data science and informatics. A broader definition names four characteristics of: data use for positive impact on citizens and society; bringing together and analyzing data from multiple sources; finding population-level insights; and developing safe, privacy-sensitive and ethical infrastructure to support research. One implication of these characteristics is that few people possess all of the requisite knowledge and skills of Population Data Science, so this is by nature a multi-disciplinary field. Other implications include the need to advance various aspects of science, such as data linkage technology, various forms of analytics, and methods of public engagement. These implications are the beginnings of a research agenda for Population Data Science, which if approached as a collective field, can catalyze significant advances in our understanding of trends in society, health, and human behavior.

Submitted: 06/09/2017
Accepted: 30/11/2017
Published: 22/02/2018

¹The University of British Columbia
²Swansea University
³University of Colorado School of Medicine
⁴University of Bristol
⁵University of Bologna
⁶PolicyWise for Children & Families
⁷University College London
⁸Edinburgh Napier University
⁹University of Calgary
¹⁰University of South Florida
¹¹Yale University
¹²Qatar University
¹³Health Research Board, Ireland
¹⁴CSIRO, Australia
¹⁵University of Western Australia
¹⁶Statistics Canada
¹⁷International Council for Clinical Evaluation Sciences (ICES)
¹⁸University of Manitoba, Manitoba Centre for Health Policy
¹⁹Australian Bureau of Statistics
²⁰Public Health and Intelligence, NHS National Services Scotland
²¹Harvard Medical School

Introduction

Developments in information and communications technologies have altered the research capabilities of almost every academic field. While advances are now new, the pace of change has increased rapidly over the last few decades. The real differences for research come from the exponential increases in computer storage and the digitization of information: <1 % of the world's information was estimated to be in digital form in 1986, compared to 94% in 2007, as shown on Figure 1 (1). Readily available digital information creates new opportunities to answer questions, on an ever-increasing population scale, about human health and well-being, the delivery of public services, and the functioning of societies.

Digital storage and collection technologies also translate to amassing more information, with one repeated assertion being

that 90% of the world's information has been collected in the last two years alone (2). Every interaction, service contact, device use, social media post, and clinical encounter is construed as a data resource from which we can extract information or meaning. More traditional administrative data, in a range of sectors, are also becoming increasingly digitized and available, with the capacity to link these data at the unit-record level now commonplace in many countries (3). Linked datasets that are large, complex, diverse, and increasingly available in near real-time open new challenges and new possibilities for the pursuit of scientific knowledge.

Data fuels the knowledge economy, and there is an increasing tendency of both private industry and the public sector to view data as an asset as well as a 'frontier for innovation' (4). The size and complexity of datasets, particularly when they derive from multiple sources, makes assembling data for

This new field of activity will provide further input to Official Statistics, as it will develop particularly on the secondary use of routine data that are normally protected by strict governmental restrictions.

This requires a multidisciplinary effort with a range of professionals.

Relevant work opportunities for all biostatisticians, if not ready other professions will take over in the health sector (e.g. data scientists, engineers, etc) !!

*Corresponding Author:
Email Address: kim.mcgrail@ubc.ca (K McGrail)

Western Australian Data Linkage System

http://www.publish.csiro.au/?act=view_file&file_id=AH080766.pdf



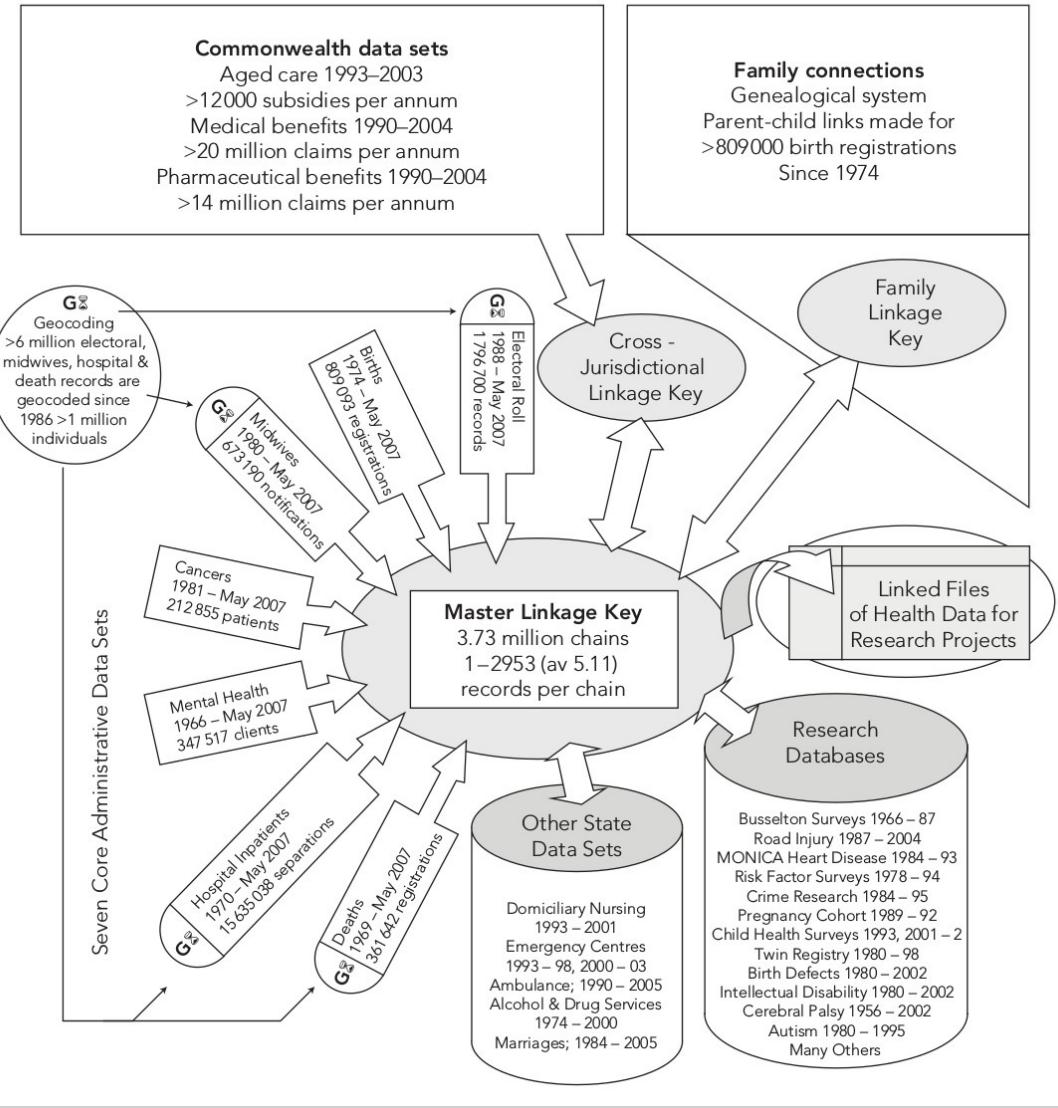
Health Information

A decade of data linkage in Western Australia: strategic design, applications and benefits of the WA data linkage system

C D'Arcy, J Holman, A John Bass, Diana L Rosman, Merran B Smith, James B Sen, Emma J Glasson, Emma L Brook, Brooke Trutwein, Ian L Rouse, Charles R Wa Nicholas H de Klerk and Fiona J Stanley

- One of the longest projects in the field of health data linkage
- Others: Oxford, Manitoba, SAIL database in Wales (UK)

I Scope of the Western Australian Data Linkage System as at May 2007





[https://web.www.healthdatagateway.org/search?
search=&datasetpublisher=SAIL&datasetSort=latest&tab=Datasets](https://web.www.healthdatagateway.org/search?search=&datasetpublisher=SAIL&datasetSort=latest&tab=Datasets)

The screenshot shows a search interface with a sidebar on the left containing links to 'SAIL Datasets', 'Application Process', and 'Glossary'. The main area displays a list of datasets.

Dataset Name
Annual District Birth Extract (ADBE)
Annual District Death Extract (ADDE)
Bowel Screening Wales (BSW)
Breast Test Wales (BTW)
Cervical Screening Wales (CSW)
Congenital Anomaly Register and Information Service (CARIS)
Emergency department Data Set (EDDS)
National Community Child Health Database (NCCHD)
Outpatient Dataset (OPD)
Patient Episode Database for Wales (PEDW)
Primary Care GP dataset
UK Health Dimensions
Welsh Cancer Intelligence and Surveillance Unit (WCISU)
Welsh Demographic Service (WDS)

The screenshot shows a user profile for 'Toby Lerone' with details like job title (Researcher) and email (SAIL.Databank@Swansea.ac.uk). It also shows a timeline of recent activity and service status information.

Service	Status
Access SAIL Gateway	Enabled
Access SAIL Gateway (old system)	Enabled
Access Approved Files	Enabled
Service Status	Normal
NWIS Switching Service	Normal
HIRU Switching Service	Normal
SAIL Gateway	Normal
FTP Services	Normal
SAIL Database	Normal
Data Mining	Normal

[Annual District Birth Extract \(ADBE\)](#)

[Annual District Death Extract \(ADDE\)](#)

[Bowel Screening Wales \(BSW\)](#)

[Breast Test Wales \(BTW\)](#)

[Cervical Screening Wales \(CSW\)](#)

[Congenital Anomaly Register and Information Service \(CARIS\)](#)

[Emergency department Data Set \(EDDS\)](#)

[National Community Child Health Database \(NCCHD\)](#)

[Outpatient Dataset \(OPD\)](#)

[Patient Episode Database for Wales \(PEDW\)](#)

[Primary Care GP dataset](#)

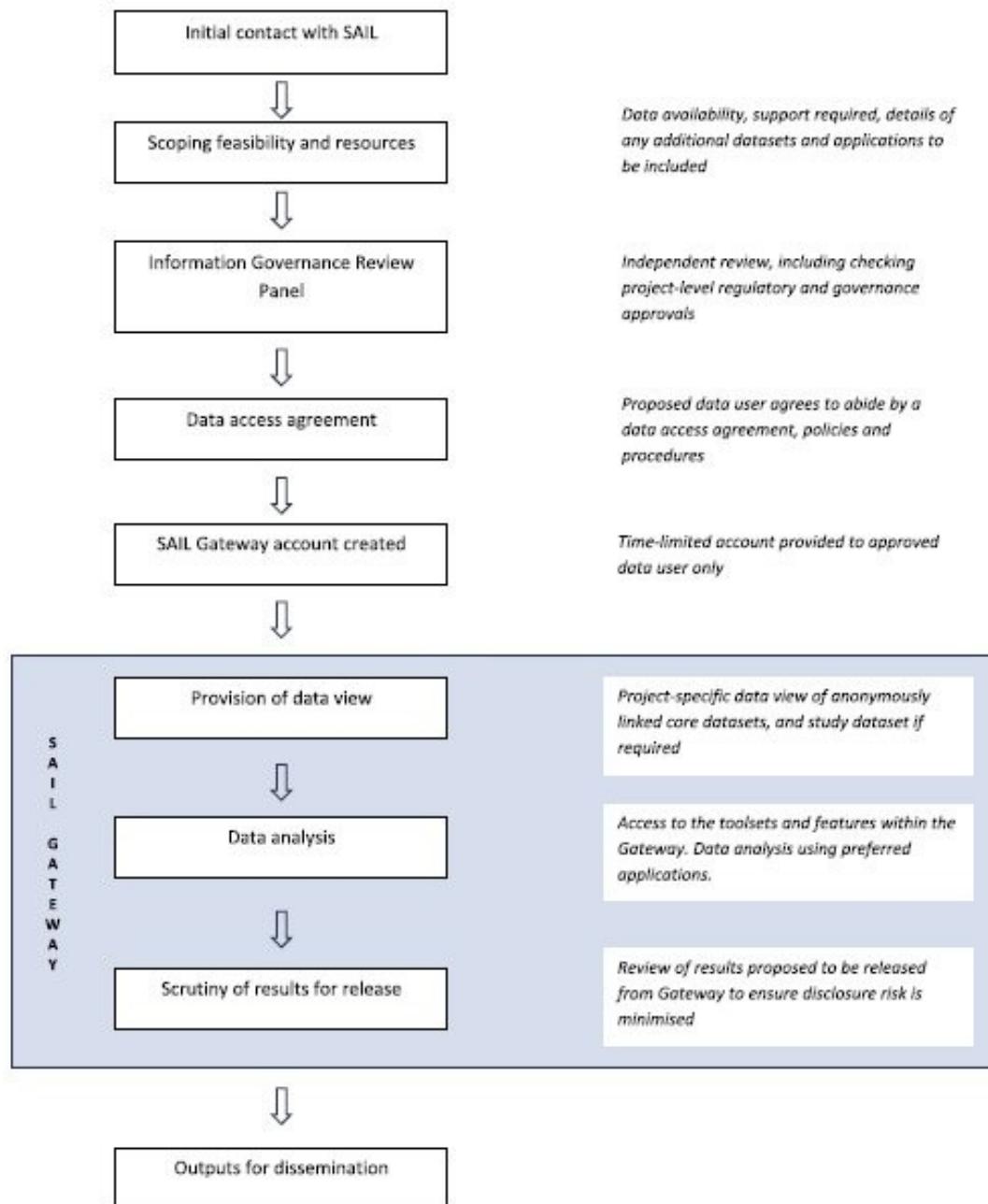
[UK Health Dimensions](#)

[Welsh Cancer Intelligence and Surveillance Unit \(WCISU\)](#)

[Welsh Demographic Service \(WDS\)](#)



Sail Data Access and Use flow chart



A screenshot of a Windows desktop environment. On the left, there's a vertical toolbar with icons for various software: Recycle Bin, Design Studio, Eclipse, InfoCentral, Prtgns, RStudio, SPSS Statistics 15, WinSQL, and Data docs. The main window shows several overlapping applications:

- A web browser window titled "OS Q&A" showing a list of questions and answers. It displays 73 questions and 130 answers. The URL is <http://osqa.chiswan.ac.uk/q>.
- An IBM InfoSphere Warehouse 10.1 interface showing a database structure with tables like "selected_project", "0000 - Useful SQL Code", and "C003 - SAIL Data Quality Reports". A specific query is visible in the SQL editor:

```
-- list all values in the column  
select 'PROV_UNIT_CD' as var, PROV_UNIT_CD, count(*) from SAILPEDWW.DIAG group by PROV_UNIT_CD order by 2;  
  
select distinct prov_unit_cd from sailrefrv.prov_unit_cd where prov_unit_cd in (select distinct PROV_UNIT_CD from S.  
  
--select 'SPELL_NUM_E' as var, SPELL_NUM_E, count(*) from SAILPEDWW.DIAG group by SPELL_NUM_E order by 2;  
select 'EPI_NUM' as var, EPI_NUM, count(*) as var, DIAG_NUM, count(*) as var, DIAG_CD_123 as var, DIAG_CD_123;
```

- A Clinical Terminology Browser window showing a tree structure of medical codes under the "Body Mass Index" category. One node is selected: "Body Mass Index normal K/M2".
- A "Conversation" window in the bottom-left corner showing a message exchange between users Joanne Demmer and Joanne Demmer.

The taskbar at the bottom shows icons for Internet Explorer, File Explorer, Task View, and the Start button.

US CDC National Center for Health Statistics

<https://www.cdc.gov/nchs/>

CDC Centers for Disease Control and Prevention
CDC 24/7: Saving Lives, Protecting People™

[A-Z Index](#)

Search

Search NCHS ▾



[Advanced Search](#)

National Center for Health Statistics



Coronavirus Disease 2019 (COVID-19)

Explore the latest COVID-19 data from NCHS and sign-up to receive e-mail updates.

[Learn more](#)



National
Health
Interview
Survey



National Survey
of Family Growth



CDC WONDER

Access the newest provisional
mortality data

US Center for Disease Control

National Health and Nutrition Examination Survey

https://www.cdc.gov/nchs/nhanes/about_nhanes.htm



National Center for Health Statistics

CDC > NCHS



National Health and Nutrition Examination Survey

About NHANES +

What's New +

Questionnaires, Datasets, and Related Documentation +

Survey Participants +

Biospecimen Program +

New Content and Proposal Guidelines

Survey Results and Products +

Tutorials +

Listserv



National Health and Nutrition Examination Survey

NHANES 2017-2018 Dietary Data

- [NHANES 2017-2018 Dietary Variable List](#)
- [Questionnaire Instruments](#)
- [Exam Procedure Manuals](#)
- [Measuring Guides for the Dietary Recall Interview](#)
- [SAS Universal Viewer](#)

Data File Name	Doc File	Data File	Date Published
Dietary Interview - Individual Foods, First Day	DR1IFF_J Doc	DR1IFF_J Data [XPT - 72.2 MB]	June 2020
Dietary Interview - Individual Foods, Second Day	DR2IFF_J Doc	DR2IFF_J Data [XPT - 59.9 MB]	June 2020
Dietary Interview - Total Nutrient Intakes, First Day	DR1TOT_J Doc	DR1TOT_J Data [XPT - 11.2 MB]	June 2020

CDC Research Data Center

<https://www.cdc.gov/rdc/>



SEARCH



CDC A-Z INDEX ▾

Research Data Center



NCHS Research Data Center (RDC)

The National Center for Health Statistics (NCHS) developed the Research Data Centers (RDC) to allow researchers access to restricted data. Today, in addition to providing access to NCHS data, the RDC also hosts restricted data from a variety of groups within the Department of Health and Human Services (HHS).



The RDC is responsible for protecting the confidentiality of survey respondents, study subjects, or institutions from which data were collected. In order to do this, we request all researchers submit a research proposal outlining the need for this more sensitive data. The proposal provides a framework for us to identify potential disclosure risk. Once approved, we work with you to create a data file specific to your research question. We cannot send you the dataset, but we offer several modes of access.

- Research Data Centers are increasingly needed to safeguard microdata and distribute individual records only to accredited researchers who require to formally apply for access
- This limitation has created barriers that in some cases are also of economic nature, given that data maintenance and processing may be expensive, charges are not minimal



Data Access Request Service (DARS): process

The DARS process consists of the following stages: Enquiry, Application, Approval, Access, Audit and Deletion of data.

Page contents

[Top of page](#)

[Enquiry](#)

[Application](#)

[Access](#)

[Audit](#)

[Deletion of data](#)

[Our service levels](#)

[Further information](#)

Enquiry

Not all enquiries progress to the application stage - it may be that your requirement can be satisfied through existing published data.

If you need any support with your enquiry you can contact us on 0300 303 5678 or email enquiries@nhsdigital.nhs.uk to discuss your data access requirements.

By discussing your data requirements now we can offer informed advice on:

- what you can do to improve the quality of your application
- whether your application is likely to be approved

Application

The application stage sets out the nature of the requested data and the purpose for which it is being requested.

NHS Digital Data Access Research Service - Charges

<https://digital.nhs.uk/services/data-access-request-service-dars/data-access-request-service-dars-charges#table-of-charges>

[Page contents](#)

[Top of page](#)

[How our charges are calculated](#)

[Table of charges](#)

How our charges are calculated

NHS Digital is publicly funded and we operate on a cost-recovery basis. We do not charge for data but we do apply charges to cover the cost of processing and delivering our service. We ensure charges are applied fairly and consistently, broadly determined by the amount of effort and approvals required.

Charges are calculated on the following components:

- the type of application
- the volume of data requested including any data linkage
- the frequency that you require data to be disseminated

Should the volume and complexity of requests require additional development work, these may be subject to additional charges.

Actual costs will be agreed with the customer during the application process.

We also offer data services to support clinical trials and research. Find out about [NHS DigiTrials](#) and the [Trusted Research Environment service for England](#).

Table of charges

Service components ^		Unit cost ^
Application type	New application	£1,030
Amendment / renewal / extension	£820	
Annual review fee ¹	£500	
Data volumes	Per year / per dataset ²	£320
Dissemination	Per dissemination ³	£930
Service required	Bespoke data linkage per dataset per dissemination ⁴	£2060
Initial tracing pack	£420	
Data Access Environment (DAE) (per user per year)	£3200	
Tabulations: per table	£800	
Permission to sub-liscence (per annum)	£10,000	

Key messages

- The scenario of health data sources at the international, national and regional level is in continuous evolution. A frequent update of the available sources is required to get a clear idea of the state of the art.
- Health data may be available either in aggregate/tабular form or as raw data. Statisticians would be naturally interested in complete individual data, but this format is increasingly difficult to acquire given the increasingly restrictive privacy regulations.
- In the near future, a formal application from accredited institutions would be normally required to gain access to individual health data

Materials

- Course notes
- *Web links included in this presentation*



ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Second cycle degree programme (LM) in Statistical Sciences

85278 – Methods and tools for Health Statistics

**85277 – Methods and tools for Official Statistics: Population
and Health Statistics**

a.a. 2022/2023

Stream 1. Regional, national and international health statistics

Topic 1.1.2

*Standardized health care data sources in the Italian National Health System
Covid-19 Dashboards*

Fabrizio Carinci

fabrizio.carinci@unibo.it

Monday, 13th February 2023

Health data sources

4. Italy

ISTAT "Noi Italia": Overview

<http://noi-italia.istat.it>

The screenshot shows the homepage of the Noi Italia 2021 website. At the top right, there are links for 'Download' (in red), 'ITA' (in blue), and 'EN' (in white). Below this is the Istat logo with the text 'Istituto Nazionale di Statistica'. A navigation bar at the top has six categories: 'POPULATION AND SOCIETY', 'EDUCATION AND LABOUR MARKET', 'HEALTH AND WELFARE', 'INDUSTRY AND SERVICES', 'ENVIRONMENT AND AGRICULTURE', and 'ECONOMY AND PUBLIC FINANCE'. The 'HEALTH AND WELFARE' category is highlighted in red. Underneath, a sub-menu for 'HEALTH' is shown. At the bottom left, there are links for 'ITALY', 'REGIONS', and 'EUROPE', with 'ITALY' underlined in red. On the bottom right, there are four circular icons with labels: 'download', 'dashboard', 'link', and 'glossary'.

Healthcare and social security are the pillars of welfare. The aim of the national healthcare systems is to promote and improve the citizens' health through empowerment, prevention, diagnosis, treatment and rehabilitation initiatives. Health indicators measure a central item in the state budget and above all a primary element of the social assistance. Over more than a decade, in Italy and in the European Union, the healthcare system has been subject to reforms aimed at rationalising and containing expenditure.

Briefly

- In 2018, in Italy, the public expenditure on health was lower than that of other European countries. Germany ranked first for per capita expenditure.
- In 2019, Italian households contributed for about 26% to the total expenditure on health, placing Italy among the top ten of EU countries.
- Mortality from neoplasms further decreased (24.7 per 10 thousand inhabitants in 2018) and gender gap reduced. At a national level, the highest rate for men was recorded in Campania (35.4 deaths per 10 thousand inhabitants).
- In the Northeast, the highest proportion of alcohol consumers at risk (18.1%) was recorded; in the South, the highest proportion of obese persons (12.1%) and in the Centre of smokers (20.8%) were recorded.
- In 2018 in Italy, there were 1,048 hospitals; 3.1 beds per thousand inhabitants, compared to an EU average value of 5.0. In particular, in the North-West and Northeast the number of beds per thousand inhabitants was equal to 3.4, in the South and Islands it was equal to 2.8 and in Calabria it was equal to 2.5, the lowest value ever registered.
- Infant mortality in Italy was among the lowest in Europe, but in the South and Islands, it was higher than in the Centre and North, and the difference has not narrowed over the last ten years.

ITALY | AN OVERVIEW

ISTAT "Noi Italia": comparisons

<http://noi-italia.istat.it>

The screenshot shows the homepage of the Noi Italia 2021 website. At the top right, there are links for 'Download' (in red), 'ITA' (Italian), and 'EN' (English). Below this is the Istat logo with the text 'Istituto Nazionale di Statistica'. The main navigation menu has six categories: 'POPULATION AND SOCIETY', 'EDUCATION AND LABOUR MARKET', 'HEALTH AND WELFARE', 'INDUSTRY AND SERVICES', 'ENVIRONMENT AND AGRICULTURE', and 'ECONOMY AND PUBLIC FINANCE'. Under 'HEALTH AND WELFARE', there is a link to 'HEALTH'. Below the menu, there are three tabs: 'ITALY' (underlined in red), 'REGIONS', and 'EUROPE'. On the right side, there are four circular icons with labels: 'download', 'dashboard', 'link', and 'glossary'.

REGIONS | ITALY AND ITS REGIONS

In 2019, the phenomena of high-risk alcohol consumption and obesity highlighted regional disparities: in the Centre and North, the share of alcohol consumers at risk was higher (16.7%), while in the South and in the Northeast, that of obese persons was higher (12.1% and 11.4% respectively). With regard to smokers, the highest share was recorded in the Centre, in particular in Lazio (22.7%) and Umbria (21.7%).

In 2018, the expenditure on health per inhabitant in the Northeast (1,915 euros) and North-west (1,930 euros) were similar and above the national average (1,875 euros), the Centre showed a value just below the national average (1,862 euros), while in the South and Islands (1,816 euros) the per capita expenditure was lower than the national average.

Total health expenditure (public and private) in 2018 amounted to 8.6% of GDP, financed by out-of-pocket households' expenditure for 1.2 percentage points. Households' contribution to the total health expenditure decreased between 2004 and 2014, but started to increase from 2015. Instead, in the same period, the overall health expenditure to GDP ratio increased by 0.5 percentage points. The increase was entirely financed through a growth in public spending (although the contribution of private spending to the overall expenditure has grown in recent years).

The incidence of the households' health expenditure as a percentage of the regional GDP was higher in the South and Islands (2.5%) and in the Northeast (2.2%). Instead, in Calabria, Friuli-Venezia Giulia, Valle d'Aosta / Vallée d'Aoste, Puglia, Basilicata, Molise and Sardegna the share was higher (over 2.5% of the regional GDP). Considering the distribution of health expenditure between the two components, public and private, the contribution of families to overall health expenditure is lower in the Southern Italy (20.1%) than in the Centre-North, where it stood at 27.2%, with a peak of 29.3% in the Northeast.

ISTAT demographic data

<http://demo.istat.it>

Contacts Legal notice IT EN

demo

demography in figures



More recent official administrative data on resident population in the Italian municipalities are available in this site. Personalized queries of data (by year, territory, citizenship, etc.) are available and allow to built the tables and download the data. Data are collected from the Population Register Offices and updated from time to time with the last available period (year, month, etc.). Data on main demographic phenomena (such as birth rates, mortality rates, population projections, ageing index, mean age, etc.) are here available.

Latest updates



Changes of Residence

Changes of Residence, years 2002 - 2021 (italian only)



Monthly Demographic Balance

Monthly demographic balance and resident population by sex, January-November 2022



Supercentenarians

Semi-supercentenarians population, years 2009-2022



Search by variable, place and period

Main structural characteristics of the population

Population dynamics

Demographic indicators

Birthrate and fertility

ISTAT Database on Health Service

http://dati.istat.it/Index.aspx?DataSetCode=DCIS_RSERVSAN#

The screenshot shows the ISTAT database interface for 'Doctor consultations and diagnostic tests'. The left sidebar includes 'Data by theme' (selected), 'Popular queries', 'Find in Themes', 'Reset', and a dropdown for 'Health statistics' containing links to the 2010 Agricultural Census, Industry Services Census 2011, Population Housing Census 2011, and various health statistics categories like Life styles and risk factors, Health conditions, Causes of death, Road accidents, Women Reproductive health, and Use of health services. The main content area displays a table with data for 2013, comparing 4 weeks and 12 months. The table has columns for 'Reference period for doctor consultations or diagnostic tests' (4 weeks and 12 months), 'Type of doctor consultations or diagnostic tests' (medical consultations, general practitioner/pediatrician, medical specialist), and 'Value' (e.g., 61.2, 33.3, 27.8). The right sidebar provides information about the survey, including its purpose, source, and data source(s). Arrows at the bottom point from the text 'Indicators' and 'Data' to the respective sections of the interface.

Reference period for doctor consultations or diagnostic tests	Year		2013	
	4 weeks	12 months	▲▼	▲▼
consultations (per 100 people with the same characteristics)	medical consultations	61.2	..	
	consultations of general practitioner or pediatrician	33.3	..	
	consultations of medical specialist	27.8	..	
persons who have used one or more consultations (per 100 people with the same characteristics)	medical consultations	32.4	..	
	consultations of general practitioner or pediatrician	21.9	..	
	consultations of medical specialist	16.8	50.6	
consultations (age-standardised rate)	medical consultations	61.4	..	
	consultations of general practitioner or pediatrician	33.3	..	
	consultations of medical specialist	28	..	
persons who have used one or more consultations (age-standardised rate)	medical consultations	32.4	..	

Indicators

Data

Sources

ISTAT Multipurpose Surveys

ISTAT started the system of Multipurpose Surveys in 1993 with the aim of collecting information about individuals and families which now provides a solid basis for social information in Italy.

It includes seven different surveys covering the following main areas:

- daily life aspects (annual),
- tourism (quarterly)
- 5 thematic surveys (every five years):
 - health conditions and use of health systems
 - citizens and free time
 - citizen safety
 - families and social individuals
 - use of free time

ISTAT multipurpose surveys: “Health conditions and use of health systems”

- “**Health conditions and use of health systems**” is a survey that allows integrating information from administrative sources with demographic and socio-economic conditions. Moreover, it allows evaluating the perceived health status as well as other data for use in health planning⁷
- It is divided in four thematic areas:
 - Health status perception, disease in acute form and trauma, chronic disease and disability
 - Main risk factors (smoking, obesity, physical inactivity)
 - Use of health service (drug consumption, examinations, hospitalization, social and health care)
 - Maternity (pregnancy, childbirth and breastfeeding) of women who had a baby during the last five years

Perceived health status

- The ISTAT Multipurpose Survey on “Health conditions and use of health systems” also includes measurement of **perceived health status** as a fundamental concept beyond morbidity and mortality.
- The questionnaire includes a multi item questionnaire including 12 items also known as “SF-12” (Short Form Health Survey).
- The structure of the DF-12 includes a first item about the general health status, followed by items referred to different dimensions of health status contributing to two separate indexes: the PCS index (Physical Component Summary) and the MCS index (Mental Component Summary), both intended as measures of the level of perceived health.

ISTAT multipurpose surveys: “Aspects of daily life”

<http://www.istat.it/en/archive/129959>

- The survey on **aspects of daily life** is an annual data collection that concerns personal behaviour, lifestyles and satisfaction for health services provided.
- Various publications are produced from this survey, including data regarding:
 - eating habits
 - drinking
 - smoking
 - health status and drug consumption
 - use of hospital services
 - accidents at home

ISTAT multipurpose surveys: “Aspects of daily life”

<http://www.istat.it/en/archive/129959>



Istituto Nazionale
di Statistica

VERSIONE IN ITALIANO



POPULATION
& HOUSEHOLDS

INSTITUTIONS
& SOCIETY

EDUCATION
& LABOUR

ECONOMY

ENVIRONMENT
& TERRITORY

SEARCH IN THIS
WEBSITE

A-Z Statistics

Glossary

[[ITALIANO](#)]

MICRODATA

ASPECTS OF DAILY LIFE: PUBLIC USE MICROSTAT FILES



2019

Last update: 22 mar 2021 20:19:23

[Survey Methodology \(it\)](#)

[Variables list \(it\)](#)

[Download](#)

2018

Last update: 30 Mar 2020 19:30:13

[Survey Methodology \(it\)](#)

[Variables list \(it\)](#)

[Download](#)

2017

Last update: 02 Oct 2019 14:39:48

[Survey Methodology \(it\)](#)

[Variables list \(it\)](#)

[Download](#)

REFERENCE PERIOD: YEAR 2017, 2016, 2015, 2014, 2013

DATE OF ISSUE: 02 OCTOBER 2019

Italian Health Administrative Databases

(Nuovo Sistema Informativo Sanitario – NSIS, Ministero della Salute)

At present, the NSIS includes the following national databases made available by all regions to the Ministry of Health in a standardised electronic format:

- Hospital discharges (SDO, 1993)
- Maternal delivery (CEDAP 2001)
- Ambulatory care (2003)
- Pharmaceutical Prescriptions – pharmacies (2003)
- Pharmaceutical Prescriptions – direct (2007)
- Emergency Services (“sistema 118”, 2008)
- Emergency Care (“Pronto soccorso”, 2008)
- Residential Care (2008)
- Home care (2008)
- Sentinel events / malpractice claims (2009)
- Addiction (2010)
- Mental Health (2010)
- Hospice (2012)

Unique Identification Number

- A core element of the Italian NSIS is the existence of a reliable unique identification number (UID) covering all served population. The UID corresponds to the tax file number assigned to each Italian citizen. Visits, diagnostic tests or pharmaceutical prescriptions are recorded in the relevant database through a **National Health Card (Tessera Sanitaria, TS)** assigned to each individual.
- Among the National System of Information Databases, the primary database used for the calculation of quality indicators is the National Database of Hospital Discharges (see above). Each subject included in the hospital database holds a UID (pseudonymised from the original TS database) and is carefully classified according to the place of residency (council, province, and region) for reimbursement purposes. The error rate for the residency is quantified in the order of 40 per 100 000 cases (Rapporto Annuale sui Ricoveri Ospedalieri 2011, 2012).

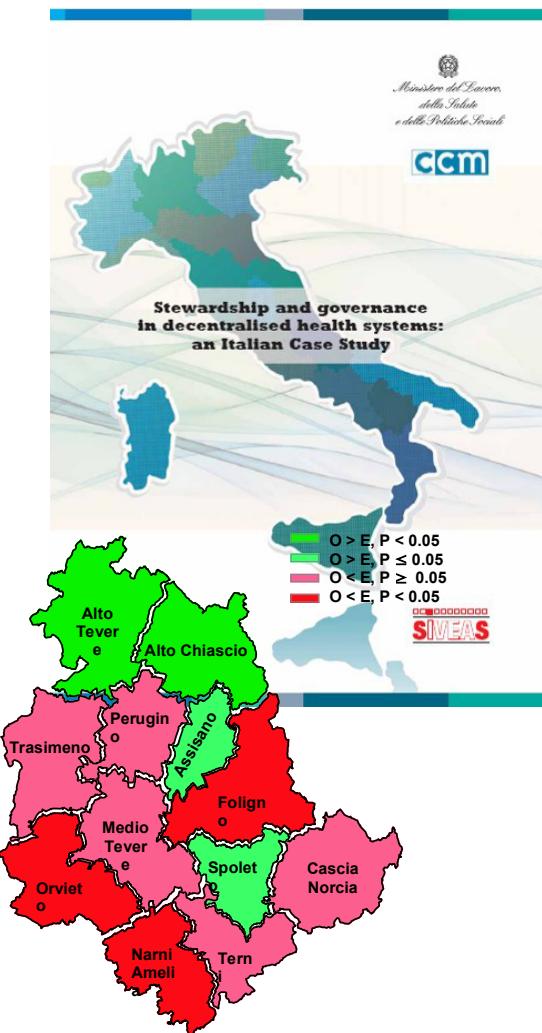
Italian Hospital Discharge Database

- The National Database of Hospital Discharges (Scheda di Dimissione Ospedaliera, SDO database) provides comprehensive and accurate data around acute care for the whole country and all Italian hospitals. It is maintained by the Ministry of Health as an official data collection from hospital discharge abstracts submitted by law by all Italian regions. The national data collection has been active since 1994.
- The SDO database includes casemix classification based on ICD-9-CM 2002 and DRG v.19 (2006-2008), ICD-9-CM 2007 and DRG v.24 (2009-today). It includes **one Principal Diagnosis** and **one Main Procedure** (including Date of Intervention) and **up to five Secondary Diagnoses and five Secondary Procedures**. In 2013, the SDO database included a total of N=6 634 977 inpatient discharges and N=1 459 hospitals. Diagnoses codes for accidents ("E codes") have been introduced in 2010 and began to stabilize after one year. The platform has been regularly used by the Ministry of Health, recently in collaboration with AGENAS, to deliver quality indicators to the OECD.

Health data sources

5. Italian regional experiences

DVSS Project – Regione Umbria - 2008



HEALTH RECORDS

- Hospital Discharge Abstracts (SDO) (2001-2005)
- Mortality Register (RENCAM) (2001-2005)
- Outpatient Specialist Visits and Pathology Tests (2001-2005)
- Pharmaceutical Prescriptions (2001-2005)
- Birth Register [file A - mother] [file B1 - mother] [file B2 - neonate] (2002-2005)
- Abortion Register (2001-2004)
- Nursing Home Stay (RUG) (2001-2004)
- Cervix Screening (2003-2005)
- Mammographic Screening (2002-2005)

STRUCTURES

- Regions
- Local Health Authorities (AO, ASL)
- Health Districts
- Hospitals
- Psychiatric Centres (SPDC)
- Physicians
- Pharmacies
- General Practitioners

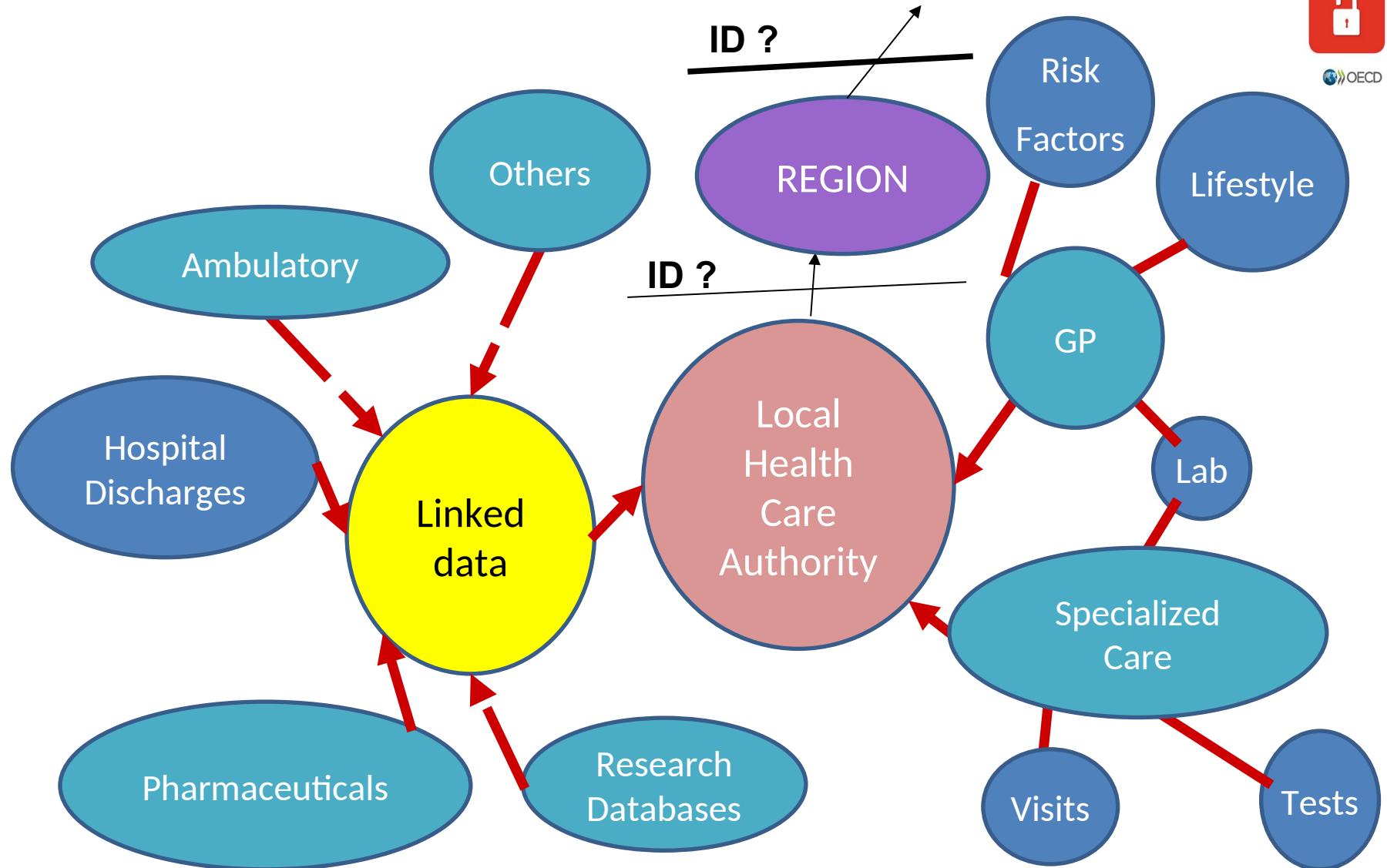
POPULATION

- Client Master Index (2005)
- Census Population (ISTAT)
- Municipalities

CLASSIFICATIONS

- Type of Physician
- Pharmaceutical Prescriptions (ATC)
- Class Pharmaceutical Treatment
- Compound
- Specialty
- Defined Daily Dose (DDD)
- ICD Revised Classification 1997 (ICD9CM-97)
- ICD Revised Classification 1997 Extended ((ICD9CM-97-ext))
- DRG Version 10
- DRG Version 19

General characteristics of decentralised Regional/National Data Linkage



Health data sources

5. Open data and public information

Global Health Data Exchange

<https://ghdx.healthdata.org/>

The screenshot shows the GHDx homepage with a green header bar containing links for IHME, GHDx, and GBD Compare, along with a search bar and login link. The main navigation menu includes Home, Countries, Series and Systems, Organizations, Keywords, IHME Data, About the GHDx, and Help. A banner at the top states: "After December 16, 2022, IHME paused its COVID-19 modeling for the foreseeable future. Past estimates and COVID-related resources will remain publicly available via healthdata.org/covid". The main content area features a "Global Health Data Exchange" section with a welcome message, links to GBD 2019 data and All IHME data, and information about data usage terms. To the right, there are sections for "Recent IHME Datasets" (listing Gross Domestic Product Per Capita 1960-2050, Global Expected Health Spending 2020-2050, etc.) and "More Ways to Explore the GHDx" (links for By Data Type, By Keyword, By Organization, and By Survey Family, Series or Systems). At the bottom, there is a "Resources" section and a footer with the Institute for Health Metrics and Evaluation logo and contact information.

IHME | GHDx | GBD Compare

Search Login

 GHDx

Global Health Data Exchange
Discover the World's Health Data

Home Countries Series and Systems Organizations Keywords IHME Data About the GHDx Help

After December 16, 2022, IHME paused its COVID-19 modeling for the foreseeable future. Past estimates and COVID-related resources will remain publicly available via healthdata.org/covid.

Global Health Data Exchange

Welcome to the GHDx, the world's most comprehensive catalog of surveys, censuses, vital statistics, and other health-related data. It's the place to start your health data search. Learn more about the catalog in [GHDx Help](#).

- [GBD 2019 data](#)
- [All IHME data](#)

Data made available for download by IHME can be used, shared, modified, or built upon by non-commercial users in accordance with the [IHME FREE-OF-CHARGE NON-COMMERCIAL USER AGREEMENT](#). For more information (and inquiries about commercial use), visit [IHME Terms and Conditions](#).

Search Data

Advanced search >>> ?

Countries

Afghanistan

Recent IHME Datasets

- [Gross Domestic Product Per Capita 1960-2050 - FGH 2021](#)
- [Global Expected Health Spending 2020-2050](#)
- [Global Health Spending 1995-2019](#)
- [Development Assistance for Health Database 1990-2021](#)
- [Development Assistance for COVID-19 Vaccine Delivery 2020-2021](#)
- [Development Assistance for Health on COVID-19 2020-2021](#)

[View all](#)  [Subscribe](#)

More Ways to Explore the GHDx

- [By Data Type](#)
- [By Keyword](#)
- [By Organization](#)
- [By Survey Family, Series or Systems](#)

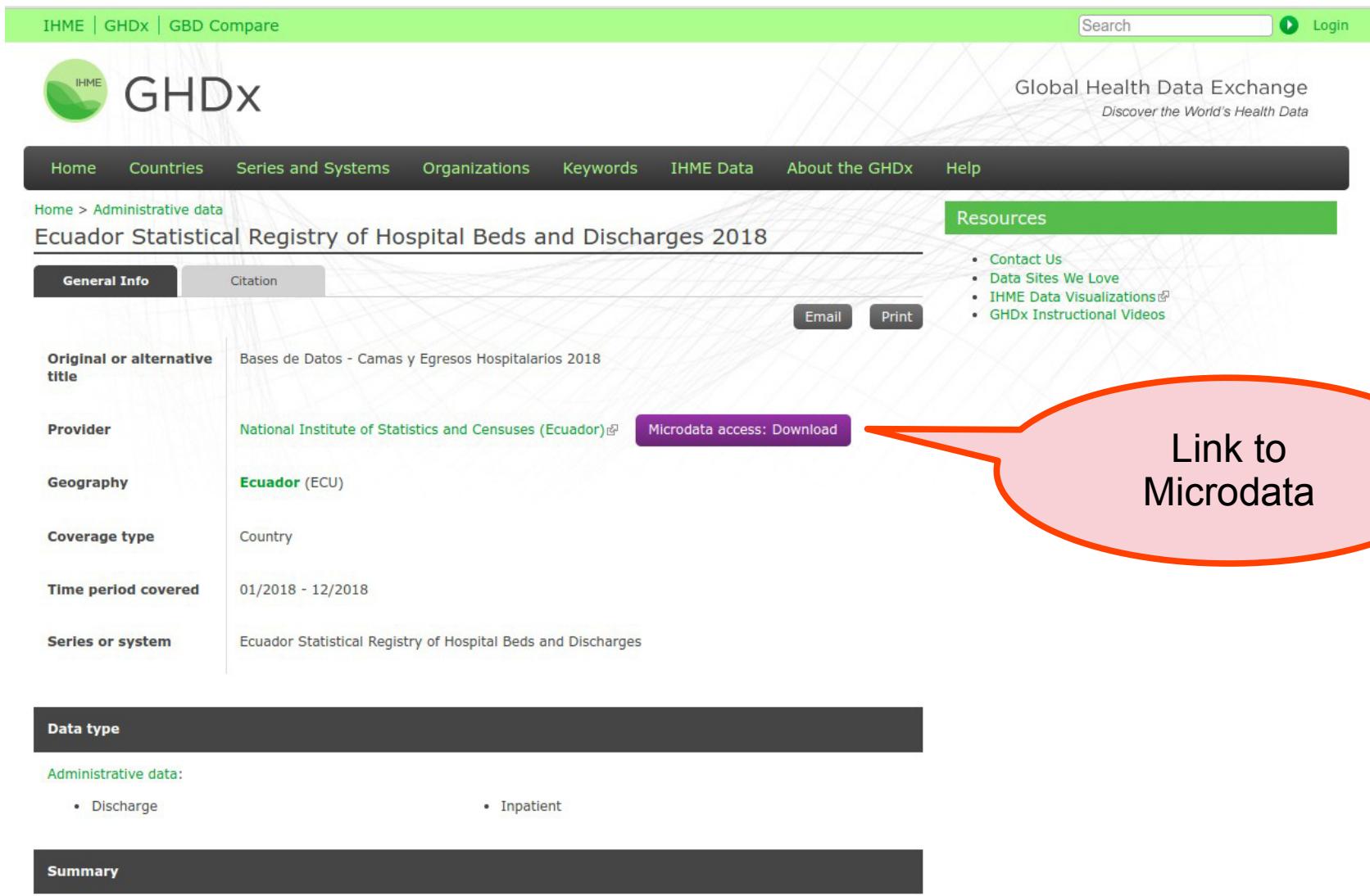
Resources

- [Contact Us](#)
- [Data Sites We Love](#)
- [IHME Data Visualizations](#)
- [GHDx Instructional Videos](#)

 Institute for Health Metrics and Evaluation
Population Health Building/Hans Rosling Center, 3980 15th Ave. NE, Seattle, WA 98195, USA
UW Campus Box #351615 | Tel: +1.206.897.2800 | Fax: +1.206.897.2899
© 2023 University of Washington

Global Health Data Exchange

<https://ghdx.healthdata.org/>



The screenshot shows the GHDx website interface. At the top, there is a navigation bar with links for IHME, GHDx, and GBD Compare, along with a search bar and a login link. The main header features the GHDx logo and the text "Global Health Data Exchange" with the subtitle "Discover the World's Health Data". Below the header, a secondary navigation bar includes links for Home, Countries, Series and Systems, Organizations, Keywords, IHME Data, About the GHDx, and Help. The main content area displays the details of the "Ecuador Statistical Registry of Hospital Beds and Discharges 2018" dataset. This includes sections for General Info, Citation, Email, and Print. The dataset details include: Original or alternative title (Bases de Datos - Camas y Egresos Hospitalarios 2018), Provider (National Institute of Statistics and Censuses (Ecuador)), Geography (Ecuador (ECU)), Coverage type (Country), Time period covered (01/2018 - 12/2018), Series or system (Ecuador Statistical Registry of Hospital Beds and Discharges), and Data type (Administrative data). Under Administrative data, options for Discharge and Inpatient are listed. To the right of the dataset details, a "Resources" sidebar lists Contact Us, Data Sites We Love, IHME Data Visualizations, and GHDx Instructional Videos. A red callout bubble points to the "Microdata access: Download" button, which is located next to the provider information.

Global Health Data Exchange
Discover the World's Health Data

Home Countries Series and Systems Organizations Keywords IHME Data About the GHDx Help

Home > Administrative data

Ecuador Statistical Registry of Hospital Beds and Discharges 2018

General Info Citation Email Print

Original or alternative title: Bases de Datos - Camas y Egresos Hospitalarios 2018

Provider: National Institute of Statistics and Censuses (Ecuador) [Download](#)

Geography: Ecuador (ECU)

Coverage type: Country

Time period covered: 01/2018 - 12/2018

Series or system: Ecuador Statistical Registry of Hospital Beds and Discharges

Data type

Administrative data:

- Discharge
- Inpatient

Summary

This dataset contains inpatient hospital discharge information as reported to the National Institute of Statistics and Censuses (INEC) in Ecuador. INEC collects these data on a monthly basis, from all health facilities in Ecuador, in both the public and private sectors.

Link to Microdata

Microdata

Microdata are **unit-level** data obtained from sample surveys, censuses, and administrative systems. They provide information about characteristics of individual people or entities such as households, business enterprises, facilities, farms or even geographical areas such as villages or towns. They allow in-depth understanding of socio-economic issues by studying relationships and interactions among phenomena.

Microdata are thus key to designing projects and formulating policies, targeting interventions and monitoring and measuring the impact and results of projects, interventions and policies.

(World Bank)

Data granularity is the level of detail to which the **unit-level** data refers, e.g. measurements based on daily admissions for Covid-19, or single hospital episode for a specific person

Ecuador National Discharge Database

<https://www.ecuadorencifras.gob.ec/camas-y-egresos-hospitalarios/>



Búsqueda



Estadísticas por tema

Estadísticas por fuente

Geografía Estadística

Banco de Datos

Consultas Especializadas

Comunicamos ▾

Contacto

Camas y Egresos Hospitalarios

Salud

Vacunación – COVID 19 **NUEVO**

Actividad Física y Sedentarismo

Actividades y Recursos de Salud

Camas y Egresos Hospitalarios

Salud, Salud Reproductiva y Nutrición

Uso del Tiempo

Tecnologías de la Información y Comunicación-TIC

Calidad de los Servicios Públicos

Presupuestos familiares

Encuesta de Estratificación del Nivel Socioeconómico

Encuesta Nacional de Ingresos y Gastos de los Hogares Urbanos y Rurales

Encuesta Nacional de Alquileres-ENALQUIL



El Registro Estadístico de Camas y Egresos Hospitalarios, recaba información, sobre la morbilidad hospitalaria, la utilización de camas hospitalarias de dotación normal y camas disponibles de los establecimientos de salud que prestan internación hospitalaria, de la Red Pública Integral de Salud (RPIS) y Red Complementaria (RC).

En 2021 se registraron 1.038.235 egresos hospitalarios, 23.196 camas disponibles en 630 establecimientos de salud a nivel nacional..

El INEC es el encargado del procesamiento.

La periodicidad de la publicación es anual.

Ecuador National Discharge Database

<https://www.ecuadorencifras.gob.ec/camas-y-egresos-hospitalarios/>



Tabulados y series históricas
Contiene los resultados del registro en forma de tablas y cuadros estadísticos.

Excel

CSV



Base de datos - periodo vigente
Acceso a la base de datos y documentación adicional que permite la interpretación y comprensión de las bases.

SPSS

CSV



Ficha de indicadores
Contiene las fichas metodológicas de las operaciones estadísticas.



Sintaxis
Contiene códigos de programación útiles para la réplica de tabulados y principales indicadores.



Diccionario de variables
Descripción de las variables que conforman las bases de datos.

Open data



Metodología

Documentación que describe los métodos, procedimientos e instrumentos para producir e interpretar la operación estadística.



Metodologías de los Registros Estadísticos de Camas y Egresos Hospitalarios:
Aspectos metodológicos y conceptuales de la operación estadística.



Historia de los Estadísticos de Camas y Egresos Hospitalarios:
Contiene los cambios por los que ha transcurrido la operación estadística.



Formularios:
Instrumento de recolección de la información de las operaciones estadísticas.



Documentación adicional

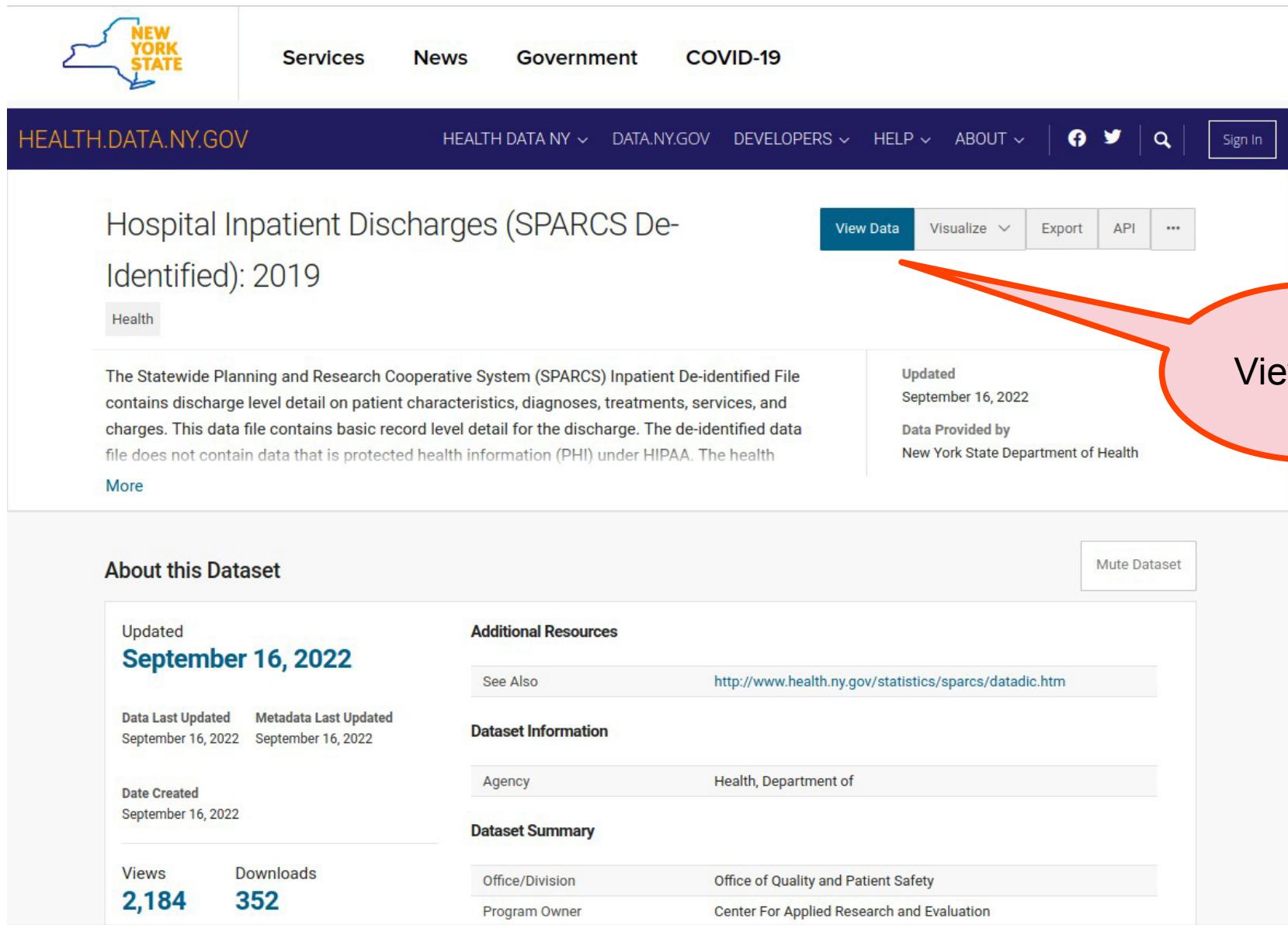
Documentación relacionada con la temática.



Información de años anteriores
Información de años anteriores al vigente, para su interpretación.

New York Inpatient Discharges (SPARCS)

<https://health.data.ny.gov/Health/Hospital-Inpatient-Discharges-SPARCS-De-Identified/4ny4-j5zv>



The screenshot shows the New York State Health Data website. At the top, there's a navigation bar with links for Services, News, Government, and COVID-19. Below that is a secondary navigation bar with links for HEALTH.DATA.NY.GOV, HEALTH DATA NY, DATA.NY.GOV, DEVELOPERS, HELP, ABOUT, and Sign In. On the left, there's a link to the Health section. The main content area displays the "Hospital Inpatient Discharges (SPARCS De-Identified): 2019" dataset. It includes a summary text about the SPARCS system, a "More" link, and a "View Data" button. To the right, there's information about the last update (September 16, 2022) and the data provider (New York State Department of Health). A large red callout bubble points to the "View Data" button. Below this, there's a section titled "About this Dataset" with tabs for "Additional Resources", "Dataset Information", and "Dataset Summary". It lists various metadata such as update dates, agency, and program owner.

Hospital Inpatient Discharges (SPARCS De-Identified): 2019

The Statewide Planning and Research Cooperative System (SPARCS) Inpatient De-identified File contains discharge level detail on patient characteristics, diagnoses, treatments, services, and charges. This data file contains basic record level detail for the discharge. The de-identified data file does not contain data that is protected health information (PHI) under HIPAA. The health

[More](#)

View Data [Visualize](#) [Export](#) [API](#) [...](#)

Updated
September 16, 2022

Data Provided by
New York State Department of Health

About this Dataset

Additional Resources

See Also <http://www.health.ny.gov/statistics/spars/datadic.htm>

Dataset Information

Agency: Health, Department of

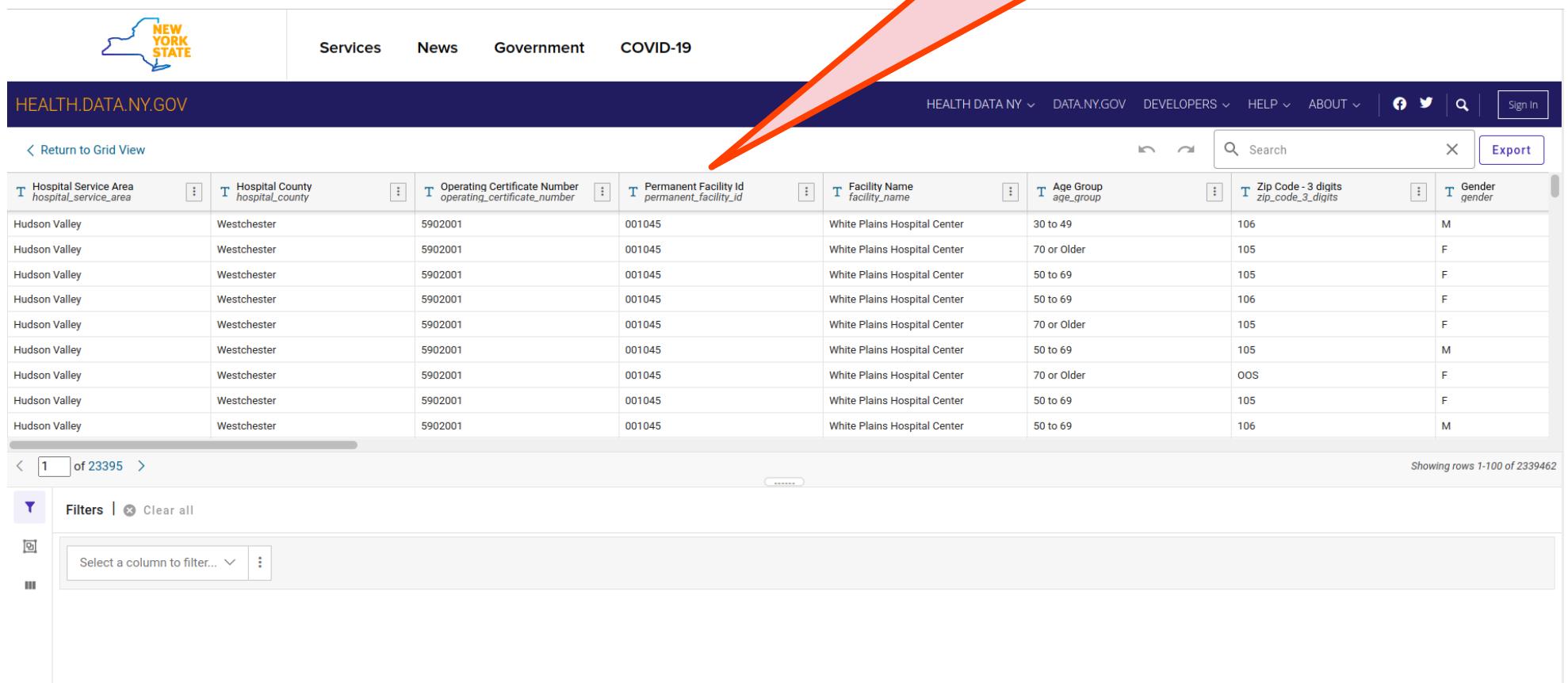
Dataset Summary

Office/Division: Office of Quality and Patient Safety
Program Owner: Center For Applied Research and Evaluation

Views: 2,184 Downloads: 352

New York Inpatient Discharges (SPARCS)

Use data dictionary for
standardized field
description



The screenshot shows a data grid from the [HEALTH.DATA.NY.GOV](https://health.data.ny.gov) website. The grid displays inpatient discharge data with the following columns:

Hospital Service Area	Hospital County	Operating Certificate Number	Permanent Facility Id	Facility Name	Age Group	Zip Code - 3 digits	Gender
Hudson Valley	Westchester	5902001	001045	White Plains Hospital Center	30 to 49	106	M
Hudson Valley	Westchester	5902001	001045	White Plains Hospital Center	70 or Older	105	F
Hudson Valley	Westchester	5902001	001045	White Plains Hospital Center	50 to 69	105	F
Hudson Valley	Westchester	5902001	001045	White Plains Hospital Center	50 to 69	106	F
Hudson Valley	Westchester	5902001	001045	White Plains Hospital Center	70 or Older	105	F
Hudson Valley	Westchester	5902001	001045	White Plains Hospital Center	50 to 69	105	M
Hudson Valley	Westchester	5902001	001045	White Plains Hospital Center	70 or Older	OOS	F
Hudson Valley	Westchester	5902001	001045	White Plains Hospital Center	50 to 69	105	F
Hudson Valley	Westchester	5902001	001045	White Plains Hospital Center	50 to 69	106	M

At the bottom left, there are filters and a search bar. The page footer indicates "Showing rows 1-100 of 2339462".

New York Inpatient Discharges (SPARCS)

<https://health.data.ny.gov/browse?q=sparcs&sortBy=relevance>



Services

News

Government

COVID-19

HEALTH.DATA.NY.GOV

HEALTH DATA NY

DATA.NY.GOV

DEVELOPERS

HELP

ABOUT



sparcs

Categories

Health

View Types

Calendars

Charts

Data Lens pages

Datasets

External Datasets

Files and Documents

Filtered Views

Forms

Maps

Tags

2008

74 Results

Sort by

Hospital Inpatient Discharges (SPARCS De-Identified): 2012

Dataset

Health

The Statewide Planning and Research Cooperative System (SPARCS) Inpatient De-Identified dataset contains discharge level detail on patient characteristics, diagnoses, treatments, services, and charges. [More](#)

Tags [discharge](#), [hospital](#), [id removed](#), [inpatient](#), [quality](#), and 1 more

[API Docs](#)

Updated September 13, 2019
Views 112,516

Hospital Inpatient Discharges (SPARCS De-Identified): 2011

Dataset

Health

The Statewide Planning and Research Cooperative System (SPARCS) Inpatient De-identified dataset contains discharge level detail on patient characteristics, diagnoses, treatments, services, charges, and [More](#)

Tags [discharge](#), [hospital](#), [inpatient](#), [quality-safety-costs](#), [sparcs](#)

[API Docs](#)

Updated September 13, 2019
Views 32,973

Hospital Inpatient Discharges (SPARCS De-Identified): 2013

Dataset

Health

The Statewide Planning and Research Cooperative System (SPARCS) Inpatient De-identified File contains discharge level detail on patient characteristics, diagnoses, treatments, services, and charges. [More](#)

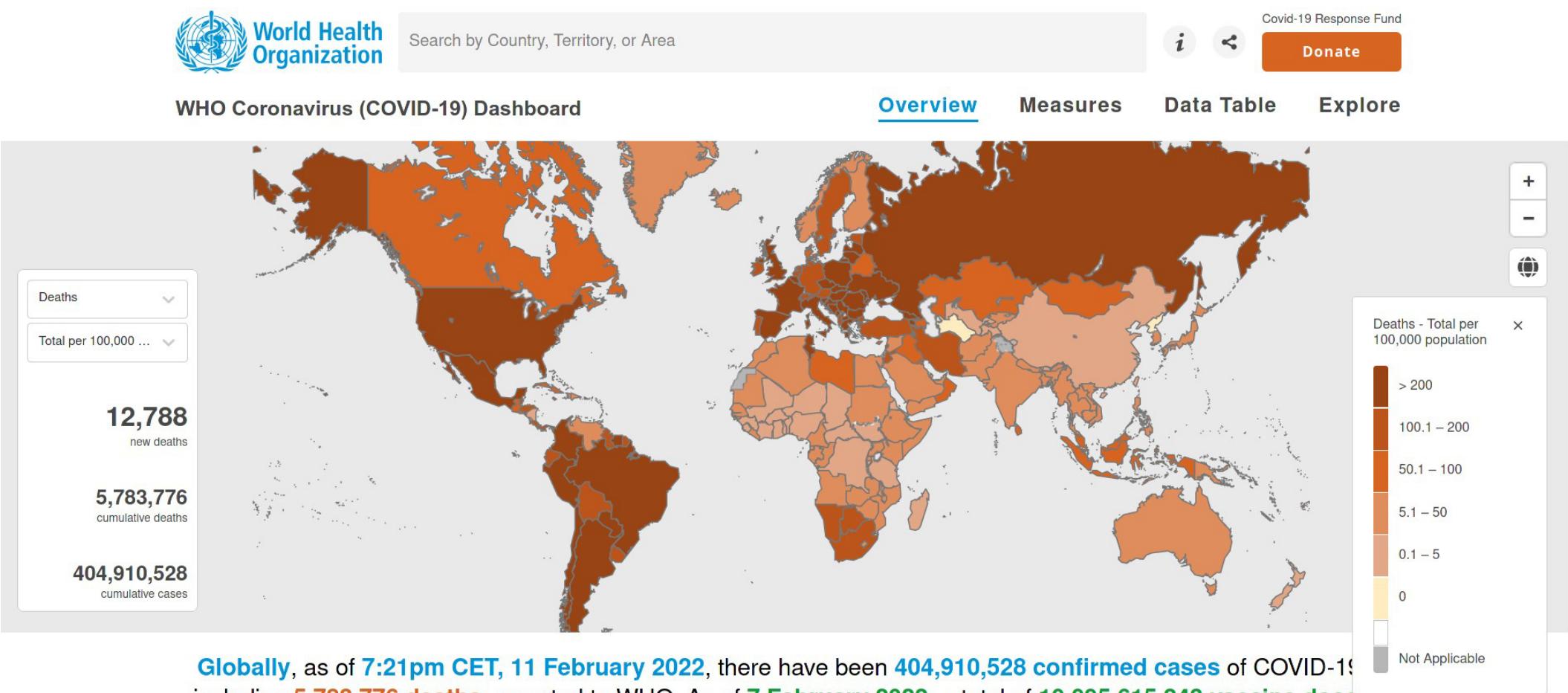
Tags [charge transparency](#), [costs](#), [discharge](#), [hospital](#), [id removed](#), and 4 more

[API Docs](#)

Updated September 13, 2019
Views 16,028

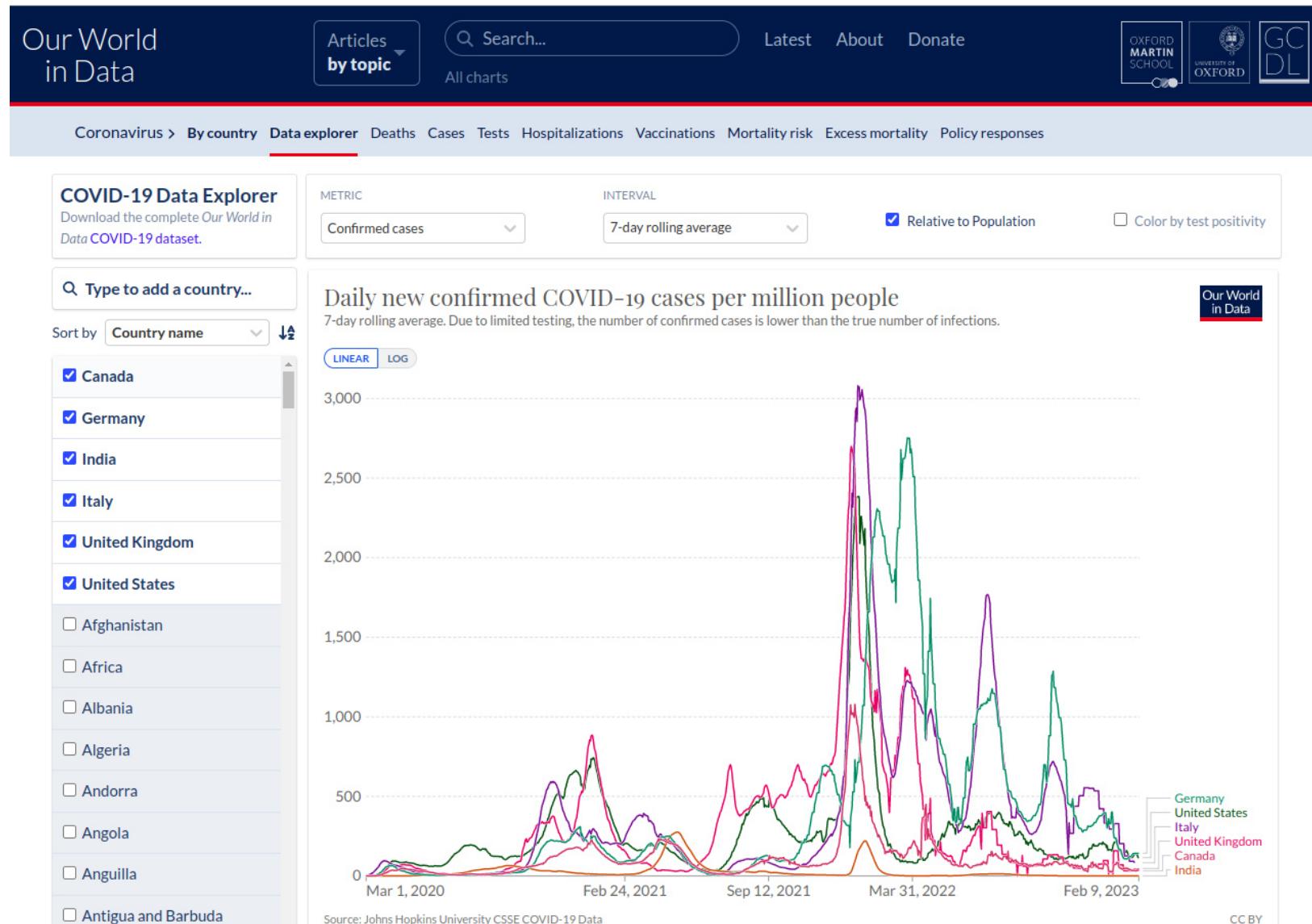
Covid-19 Health Statistics: WHO International

<https://covid19.who.int/>



Our World in Data Covid-19

<https://ourworldindata.org/explorers/coronavirus-data-explorer>



Covid-19 Health Statistics: AGENAS, Italy

<https://www.agenas.gov.it/covid19/>

The screenshot shows the AGENAS Covid-19 Portal homepage. At the top, there is a blue header bar with the Agenzia Nazionale per i Servizi Sanitari Regionali logo and a BETA indicator. On the right, there is an English language selection (ENG) and a dropdown menu. Below the header, the title "Covid-19 Portal" is displayed next to a small icon. To the right of the title are social media links for Twitter, Facebook, and YouTube. The main navigation menu includes "Homepage", "Emergency Department" (with a "Novità" button), "Vaccines", "Forecasts", "Resilience", "Graphs and Tables", "Measures", and "Good Practices". A timestamp at the top left indicates "Data updated to 11 February 2022 time 20:45". Below the menu, three monitoring sections are listed: "Current situation", "National Monitoring", and "Regional Monitoring". The main content area features a "AGENAS ANALYTICS" section with a blue ribbon icon. It displays two circular gauge charts: one for "Italy - Intensive Care" and another for "Italy - Non Intensive Care ("Non Critical Area")". Both charts have scales from 0 to 100, with the needle pointing towards the red zone.

Data updated to 11 February 2022 time 20:45

Current situation National Monitoring Regional Monitoring

AGENAS ANALYTICS

Updated on 11 February 2022

Select Region

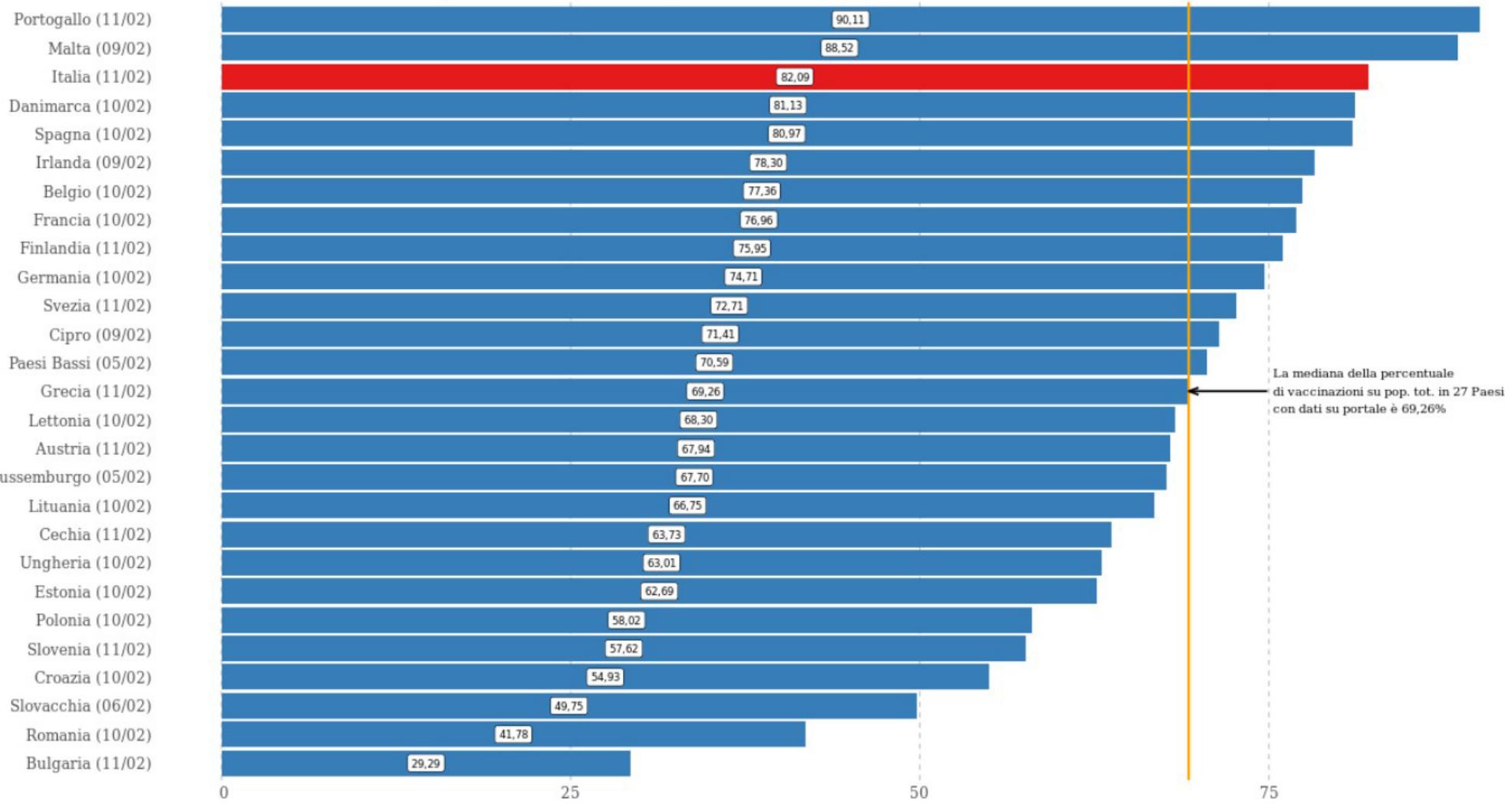
Italy - Intensive Care

Italy - Non Intensive Care ("Non Critical Area")

Covid-19 Portal AGENAS - Vaccines

<https://www.agenas.gov.it/covid19/web/index.php?r=english%2Findex>

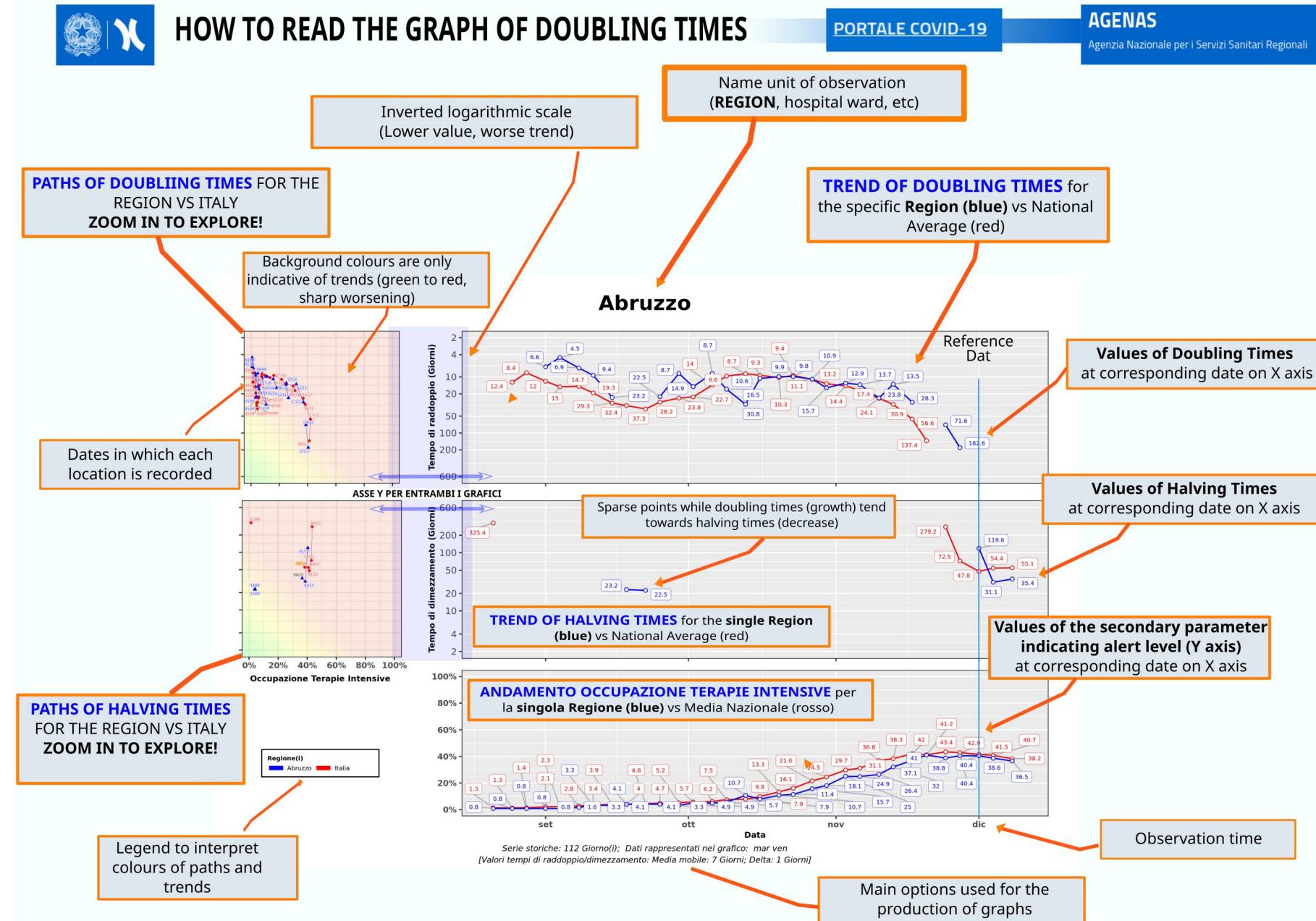
Percentuale con ciclo di vaccinazione completa per Coronavirus (Covid-19) su popolazione totale per Paesi UE
Elaborazione AGENAS su dati disponibili alle ore 09:41 del giorno 12/02/2022



Fonti: <https://github.com/italia/covid19-opendata-vaccini/blob/master/dati>, <https://ourworldindata.org/covid-vaccinations>, <https://ec.europa.eu/eurostat/databrowser/view/tps00001/default>

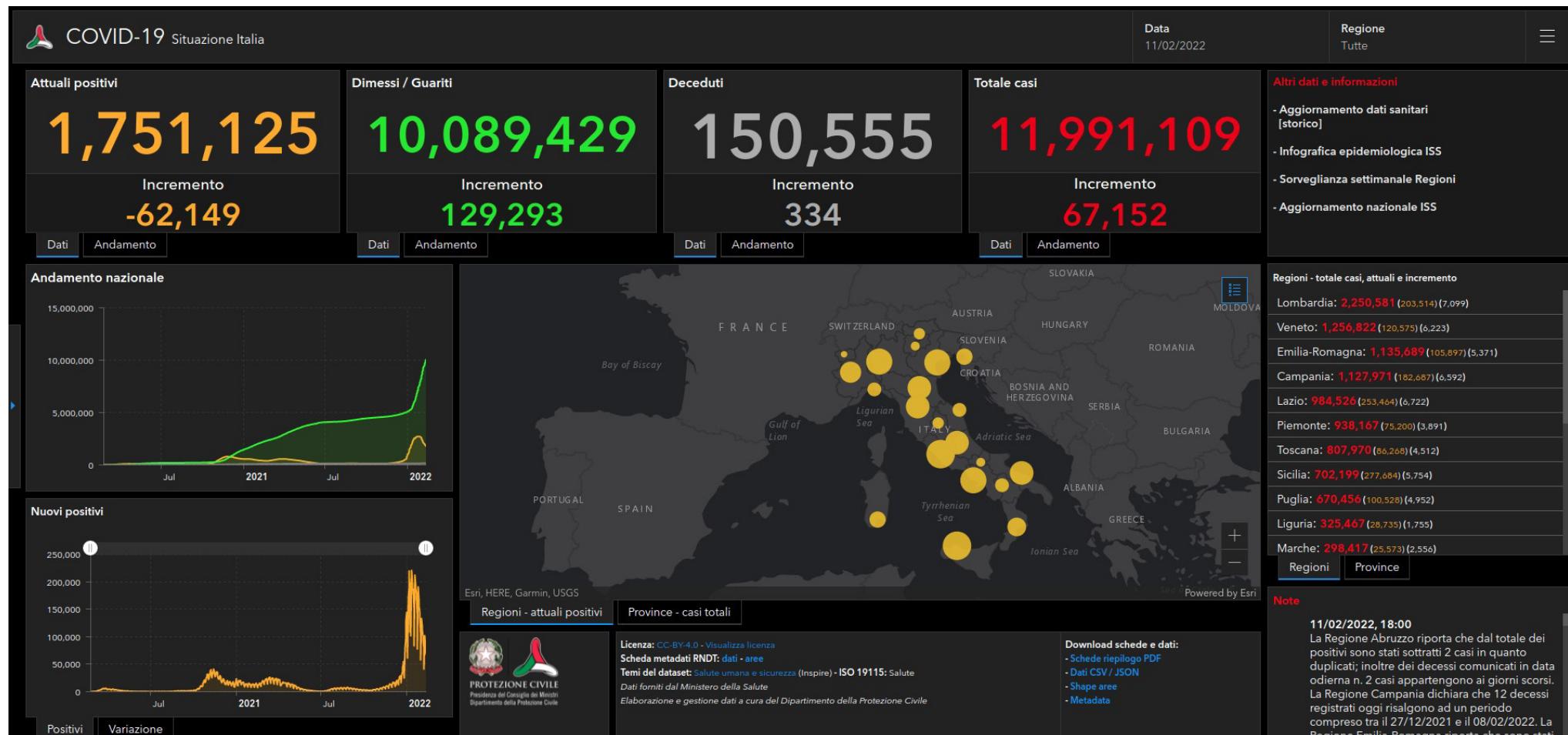
Doubling times AGENAS

<https://www.agenas.gov.it/covid19/>



Covid-19 ArcGIS: Protezione Civile

<https://www.salute.gov.it/covid-19/dati/desktop>



Covid-19 Open Data: Protezione Civile

<https://github.com/pcm-dpc/COVID-19>

Screenshot of the GitHub repository page for `pcm-dpc / COVID-19`.

The repository is public and has 171 issues, 2 pull requests, and 1914 commits.

Key details from the repository page:

- Code**: master branch, 2 branches, 0 tags.
- Commit History**: Last commit by `pierluigicara` on 2022-02-12 at 1 hour ago, with 1,914 commits.
- Files**: .github/ISSUE_TEMPLATE, Update errore.md, areae, assets/img, dati-andamento-nazionale, dati-contratti-dpc-forniture, dati-json, dati-province, dati-regioni, dati-statistici-riferimento, metadata, metriche, note, schede-iss, schede-riepilogative.
- Contributors**: 22 contributors (icons shown).
- Topics**: gov, pcm, dpc, covid-19.
- Activity**: 3.9k stars, 211 watching, 2.3k forks.
- Pages**: Readme, View license, Code of conduct.
- Releases**: No releases published.
- Packages**: No packages published.
- Contributors**: 22 contributors (icons shown).

Covid-19 Open Data: browsing latest data

<https://github.com/pcm-dpc/COVID-19/blob/master/dati-regioni/dpc-covid19-ita-regioni-latest.csv>

pcm-dpc / COVID-19 Public

Code Issues 171 Pull requests Actions Security Insights

master COVID-19 / dati-regioni / dpc-covid19-ita-regioni-latest.csv Go to file ...

dpc-005 2022-02-11 Latest commit 90a6a2b 19 hours ago History

13 contributors +1

22 lines (22 sloc) | 7.62 KB Raw Blame ⌂ ⌐ ⌒

Search this file...

1	data	stato	codice_regione	denominazione_regione	lat	long	ricoverati_con_sintomi	terapia_intensiva	totale_ospedalizzati	isolamento_domiciliare	totale_posi
2	2022-02-11T17:00:00	ITA	13	Abruzzo	42.35122196	13.39843823	508	28	536	107395	107931
3	2022-02-11T17:00:00	ITA	17	Basilicata	40.63947052	15.80514834	98	7	105	20014	20119
4	2022-02-11T17:00:00	ITA	18	Calabria	38.90597598	16.59440194	354	22	376	44821	45197
5	2022-02-11T17:00:00	ITA	15	Campania	40.83956555	14.25084984	1248	72	1320	181367	182687
6	2022-02-11T17:00:00	ITA	08	Emilia-Romagna	44.49436681	11.3417208	2147	132	2279	103618	105897
7	2022-02-11T17:00:00	ITA	06	Friuli Venezia Giulia	45.6494354	13.76813649	390	41	431	37338	37769
8	2022-02-11T17:00:00	ITA	12	Lazio	41.89277044	12.48366722	1989	184	2173	251291	253464
9	2022-02-11T17:00:00	ITA	07	Liguria	44.41149315	8.9326992	633	29	662	28073	28735
10	2022-02-11T17:00:00	ITA	03	Lombardia	45.46679409	9.190347404	2161	174	2335	201179	203514
11	2022-02-11T17:00:00	ITA	11	Marche	43.61675973	13.5188753	301	47	348	25225	25573
12	2022-02-11T17:00:00	ITA	14	Molise	41.55774754	14.65916051	43	4	47	7253	7300
13	2022-02-11T17:00:00	ITA	21	P.A. Bolzano	46.49933453	11.35662422	114	8	122	12110	12232
14	2022-02-11T17:00:00	ITA	22	P.A. Trento	46.06893511	11.12123097	135	12	147	9099	9246
...

Covid-19 Open Data: Protezione Civile

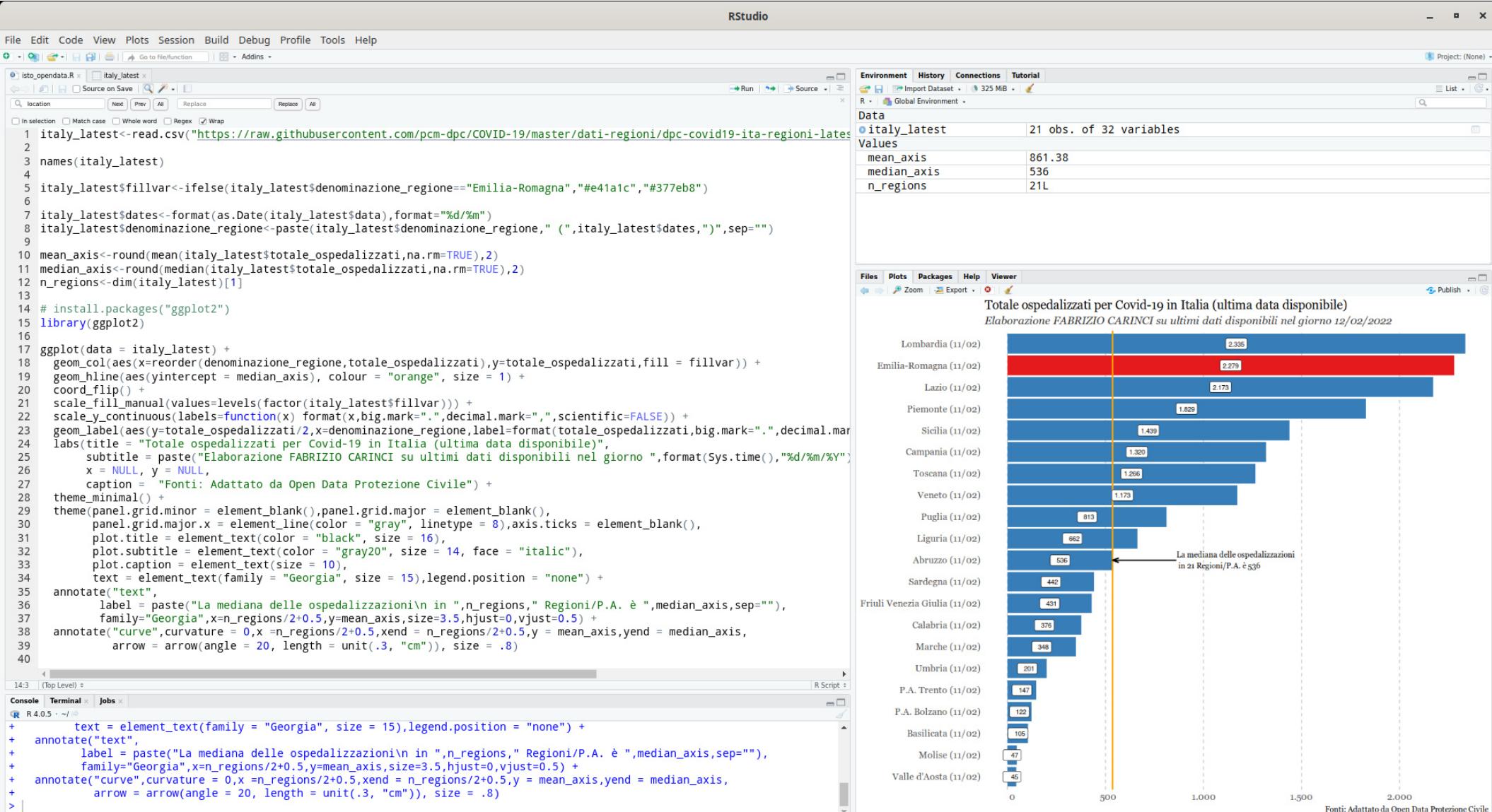
<https://raw.githubusercontent.com/pcm-dpc/COVID-19/master/dati-regioni/dpc-covid19-ita-regioni-latest.csv>

data	stato	codice_regione	denominazione_regione	lat	long	ricoverati_con_sintomi	terapia_intensiva	totale_ospedalizzati	isolamento_domiciliare	totale_positivi	variazione_totale_positivi	nuovi_positivi	dimessi_guariti	deceduti	casi_da_sospetto_diagnostico	casi_da_screening	totale_casi	tamponi_casi_testati	note_ingressi_terapia_intensiva	note_test	note_casi	totale_positivi_test_molecolare	totale_positivi_test_antigenico_rapido	tamponi_test_molecolare	tamponi_test_antigenico_rapido	codice_nuts_1	codice_nuts_2		
2022-02-11T17:00:00	ITA	13	Abruzzo	42.35122196	13.39843823	508	28	536	107395	107931	-1704	1872	129873	2869	,,	240673	4783668	1150167	Dal totale dei positivi sono stati sottratti 2 casi in quanto duplicati. Nota informativa: Dal totale dei decessi comunicati in data odierna 2 casi appartengono ai giorni scorsi.,,,148133,92540,1994698,2788970,ITF,ITF1										
2022-02-11T17:00:00	ITA	17	Basilicata	40.63947052	15.80514834	98	7	105	20014	20119	-58	666	53039	708	,,	73866	748757	296388	"Il numero totale dei decessi ne comprende n. 25 a carico di pazienti non residenti, deceduti in strutture ospedaliere della Regione Basilicata. ",,,55878,17988,600881,147876,ITF,ITF5										
2022-02-11T17:00:00	ITA	18	Calabria	38.59057598	16.59440194	354	22	376	44821	45197	662	2055	141949	1987	,,	189133	2122078	1644236	,,"L'asp di Cosenza comunica : "" Nel setting fuori regione si registra 1 nuovo caso ricoverato\nSi segnala, inoltre, che 11 casi a domicilio e 1 caso guarito, precedentemente comunicati tra i residenti in regione, a seguito di verifiche oggi sono compresi tra i non residenti".".,156370,32763,1446976,675102,ITF,ITF6										
2022-02-11T17:00:00	ITA	15	Campania	40.83956555	14.25084984	1248	72	1320	181367	182687	-5923	6592	935789	9495	,,	1127971	13098133	4492742	,,10,,"a seguito delle verifiche quotidiane si evince che 12 decessi registrati oggi, risalgono ad un periodo compreso tra il 27/12/2021 e il 08/02/2022".,830564,297407,7833196,5264937,ITF,ITF3										
2022-02-11T17:00:00	ITA	08	Emilia-Romagna	44.49436681	11.3417208	2147	132	2279	103618	105897	-10092	5371	1014287	15505	,,	1135689	14126514	2606842	"Sono stati eliminati 9 casi, comunicati nei giorni precedenti, in quanto giudicati non casi COVID-19.",,,829785,305904,8283341,5843173,ITH,ITH5										
2022-02-11T17:00:00	ITA	06	Friuli Venezia Giulia	45.6494354	13.76813649	390	41	431	37338	37769	-2295	1436	251426	4622	,,	293817	5552871	1071246	"Il totale dei casi positivi è stato ridotto di 5 a seguito di 5 test positivi rimossi dopo revisione dei casi (2 casi relativi alla provincia di PN, 1 caso relativo alla provincia di UD, 1 caso relativo alla provincia di GO, 1 caso relativo alla provincia di TS). ",,,177719,116098,2986210,2566661,ITH,ITH4										
2022-02-11T17:00:00	ITA	12	Lazio	41.89277044	12.48366722	1989	184	2173	251291	253464	-3879	6722	720954	10108	,,	984526	168837408	5313397	,,9,,"790307,194219,7880688,8956720,ITI,ITI4										
2022-02-11T17:00:00	ITA	07	Liguria	44.41149315	8.9326992	633	29	662	28073	28735	-1832	1755	291738	4994	,,	325467	4587240	1209183	"Si precisa che nel flusso informativo degli ospedalizzati in Area Medica e Terapia Intensiva sono conteggiati tutti i pazienti SARS-CoV2 positivi ricoverati sia per patologia Covid-19 correlata sia per altre cause. Inoltre, si fa presente che i pazienti attualmente ospedalizzati per patologia non Covid-19 correlata ammontano a circa il 30% del totale degli ospedalizzati positivi per SARS-CoV2.",,,*di cui 8425 reinfezioni a partire da 3/09/2021 [circ. min sal. n.37911 del 20/08/2021].,208256,117211,2231489,2355751,ITC,ITC3										
2022-02-11T17:00:00	ITA	03	Lombardia	45.466679409	9.190347404	2161	174	2335	201179	203514	-14405	7099	2009140	37927	,,	2250581	31895899	7716705	,,9,,"1300976,949605,14746068,17149831,ITC,ITC4										
2022-02-11T17:00:00	ITA	11	Marche	43.61675973	13.5188753	301	47	348	25225	25573	-150	2556	269352	3492	,,	298417	2622084	1660688	,,2,,196975,101442,1874516,747568,ITI,ITI3										
2022-02-11T17:00:00	ITA	14	Molise	41.55774754	14.65916051	43	4	47	7253	7300	-54	353	26409	551	,,	34260	448236	401801	,,,21666	12594	357742	90494,ITF,ITF2							
2022-02-11T17:00:00	ITA	21	P.A. Bolzan	46.49933453	11.35662422	114	8	122	12110	12232	-1108	1031	163870	1380	,,	177482	3542133	613991	"1031 nuovi positivi di cui 6 test antigenici confermati da test molecolare e 920 test antigenici. A questi sono stati sottratti 1 stormi di pazienti positivi al test antigenico ma negativi a test molecolare eseguito entro 3 giorni, per un totale netto di 919 positivi a test antigenico comunicati",,0,,"1031 nuovi positivi di cui 6 test antigenici confermati da test molecolare e 920 test antigenici comunicati",,88685,88797,834843,2707290,ITH,ITH1										
2022-02-11T17:00:00	ITA	22	P.A. Trento	46.06893511	11.12123097	135	12	147	9099	9246	-923	570	122098	1494	,,	132838	2311845	525370	,,,42140	90698	817405	1494440,ITH,ITH2							
2022-02-11T17:00:00	ITA	01	Piemonte	45.0732745	7.686867483	1323	97	1829	73371	75200	-4554	3891	850150	12817	,,	938167	14582005	3654330	,12,,	467334	470833	4606153	9975852,ITC,ITC1						
2022-02-11T17:00:00	ITA	16	Puglia	41.12559576	16.86736689	745	68	813	99715	100528	-1016	4952	562520	7408	,,	670456	8235298	2044328	,,,424672	245784	3862459	4372839,ITF,ITF4							
2022-02-11T17:00:00	ITA	20	Sardegna	39.21531192	9.110616306	413	29	442	32197	32639	-1020	2575	114845	1927	,,	149411	3581781	1418658	,4,,"L'incremento dei nuovi casi tiene conto anche dei casi diagnosticati con test antigenico, ai sensi di quanto disposto con Ordinanza n. 2 del 3 febbraio 2022 del Presidente della Regione Sardegna. Pertanto si specifica che dei 2575 casi dichiarati oggi, 1807 sono stati diagnosticati da tamponi antigenico Si segnala il decesso dei paz.: - 1 uomo 82 aa Residente nella Provincia del Sud Sardegna - 1 uomo 86 aa Residente nella Provincia di Oristano - 2 decessi in Provincia di Nuoro",,136289,13122,1721415,1860366,ITG,ITG2										
2022-02-11T17:00:00	ITA	19	Sicilia	38.11569725	13.3623567	1323	116	1439	276245	277684	2811	5754	415533	8982	,,	702199	10257263	46998050	DECEDUTI: N. 4 IL 11/02/22 - N. 8 IL 10/02/22 - N. 19 IL 09/02/22 - N. 2 IL 08/02/22 - N. 1 IL 02/02/22	,,9,,"Sul numero complessivo dei casi confermati comunicati in data odierna, n. 342 sono relativi a giorni precedenti al 08/02/22 (di cui n. 31 il 07/02/22, n. 3 il 06/02/22, n. 19 il 05/02/22, n. 88 il 04/02/22)",,471086,231113,4253860,6003403,ITG,ITG1									
2022-02-11T17:00:00	ITA	09	Toscana	43.76923077	11.25588885	1175	91	1266	85002	86268	-6368	4512	713115	8587	,,	807970	11709561	4260326	,8,,	535229	272741	6222253	5487308,ITI,ITI2						
2022-02-11T17:00:00	ITA	10	Umbria	43.10675841	12.38824698	194	7	201	16102	16303	-530	1106	152748	1682	,,	170733	3560209	641438	-,	Si fa presente che 10 dei ricoveri non UTI appartengono ai codici disciplina di Ostetricia & Ginecologia e Pediatria - Si fa presente che 17 dei ricoveri non UTI appartengono ad altri codici disciplina - Si fa presente che 3 deceduti comunicati non appartengono alle 24h precedenti, ma sono frutto di allineamento dei sistemi informativi.									
2022-02-11T17:00:00	ITA	02	Valle d'Aosta	45.73750286	7.320149366	42	3	45	2219	2264	-194	61	27853	514	,,	30631	453089	126092	,,,13980	16651	133251	319838,ITC,ITC2							
2022-02-11T17:00:00	ITA	05	Veneto	45.43490485	12.33845213	1079	94	1173	119402	120575	-11557	6223	1122741	13506	,,	1256822	24591960	4499125	Nei valori riportati per le terapie intensive si è verificato un disallineamento temporale del flusso informativo pertanto per convenzione è stato riportato 28 dimessi da TI invece del n. 8 effettivi che include anche i negativizzati.,,11,,732023,524799,9134784,15457176,ITH,ITH3										

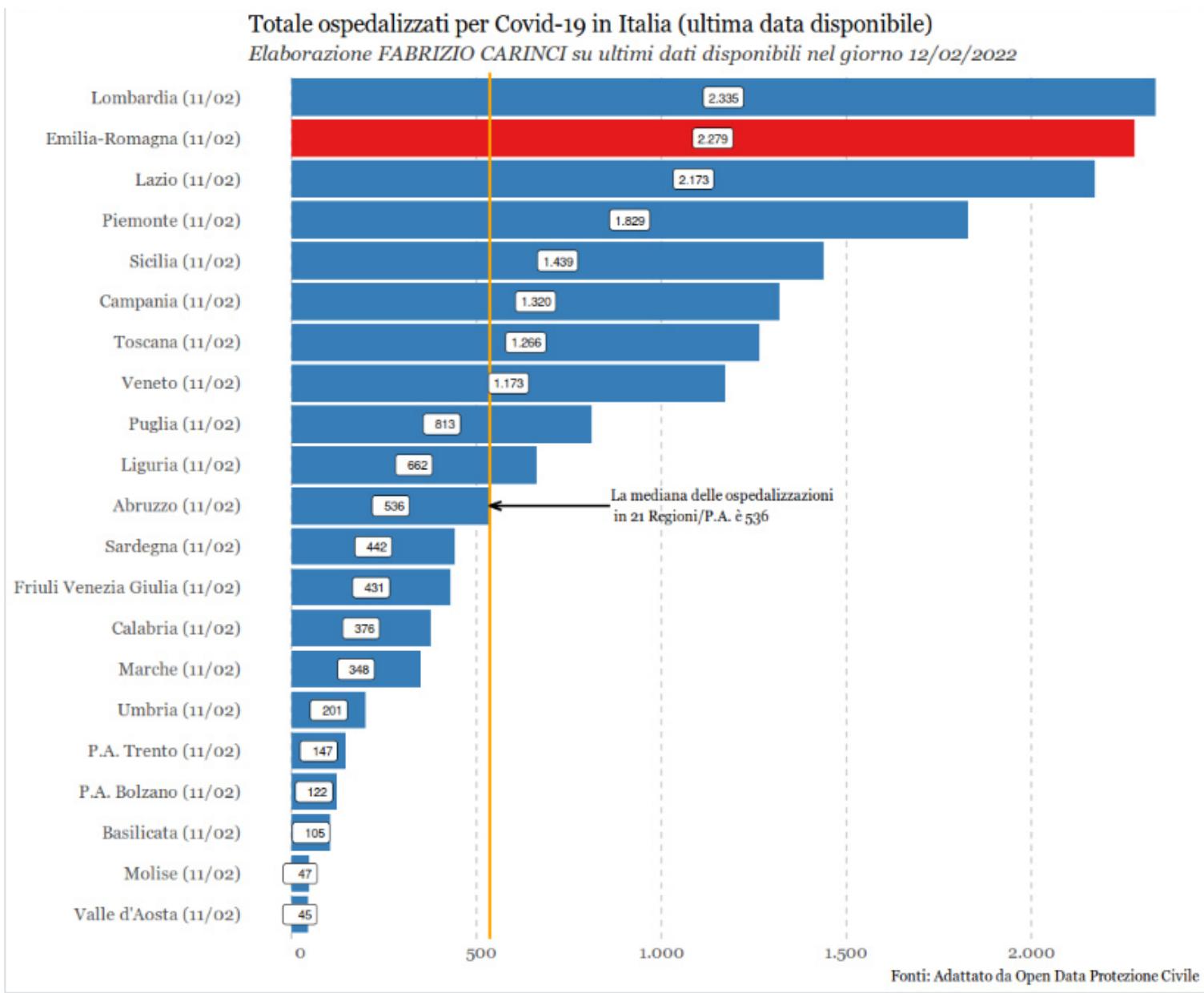
Processing Covid-19 Open Data in R

```
italy_latest<-read.csv("https://raw.githubusercontent.com/pcm-dpc/COVID-19/master/dati-regioni/dpc-covid19-ita-regioni-latest.csv",sep=",",stringsAsFactors = FALSE,na.strings="")  
  
names(italy_latest)  
  
italy_latest$fillvar<-ifelse(italy_latest$denominazione_regione=="Emilia-Romagna","#e41a1c","#377eb8")  
  
italy_latest$dates<-format(as.Date(italy_latest$data),format="%d/%m")  
italy_latest$denominazione_regione<-paste(italy_latest$denominazione_regione," (",italy_latest$dates,") ",sep="")  
  
mean_axis<-round(mean(italy_latest$totale_ospedalizzati,na.rm=TRUE),2)  
median_axis<-round(median(italy_latest$totale_ospedalizzati,na.rm=TRUE),2)  
n_regions<-dim(italy_latest)[1]  
  
install.packages("ggplot2")  
library(ggplot2)  
  
ggplot(data = italy_latest) +  
  geom_col(aes(x=reorder(denominazione_regione,totale_ospedalizzati),y=totale_ospedalizzati,fill = fillvar)) +  
  geom_hline(aes(yintercept = median_axis), colour = "orange", size = 1) +  
  coord_flip() +  
  scale_fill_manual(values=levels(factor(italy_latest$fillvar))) +  
  scale_y_continuous(labels=function(x) format(x,big.mark=".",decimal.mark=",",scientific=FALSE)) +  
  geom_label(aes(y=totale_ospedalizzati/2,x=denominazione_regione,label=format(totale_ospedalizzati,big.mark=".",decimal.mark=",",)),size=3) +  
  labs(title = "Totale ospedalizzati per Covid-19 in Italia (ultima data disponibile)",  
       subtitle = paste("Elaborazione FABRIZIO CARINCI su ultimi dati disponibili nel giorno ",format(Sys.time(),"%d/%m/%Y"),sep=""),  
       x = NULL, y = NULL,  
       caption = "Fonti: Adattato da Open Data Protezione Civile") +  
  theme_minimal() +  
  theme(panel.grid.minor = element_blank(),panel.grid.major = element_blank(),  
        panel.grid.major.x = element_line(color = "gray", linetype = 8),axis.ticks = element_blank(),  
        plot.title = element_text(color = "black", size = 16),  
        plot.subtitle = element_text(color = "gray20", size = 14, face = "italic"),  
        plot.caption = element_text(size = 10),  
        text = element_text(family = "Georgia", size = 15),legend.position = "none") +  
  annotate("text",  
          label = paste("La mediana delle ospedalizzazioni\nin ",n_regions," Regioni/P.A. è ",median_axis,sep=""),  
          family="Georgia",x=n_regions/2+0.5,y=mean_axis,size=3.5,hjust=0,vjust=0.5) +  
  annotate("curve",curvature = 0,x = n_regions/2+0.5,xend = n_regions/2+0.5,y = mean_axis,yend = median_axis,  
          arrow = arrow(angle = 20, length = unit(.3, "cm")), size = .8)
```

Processing Covid-19 Open Data in R Studio

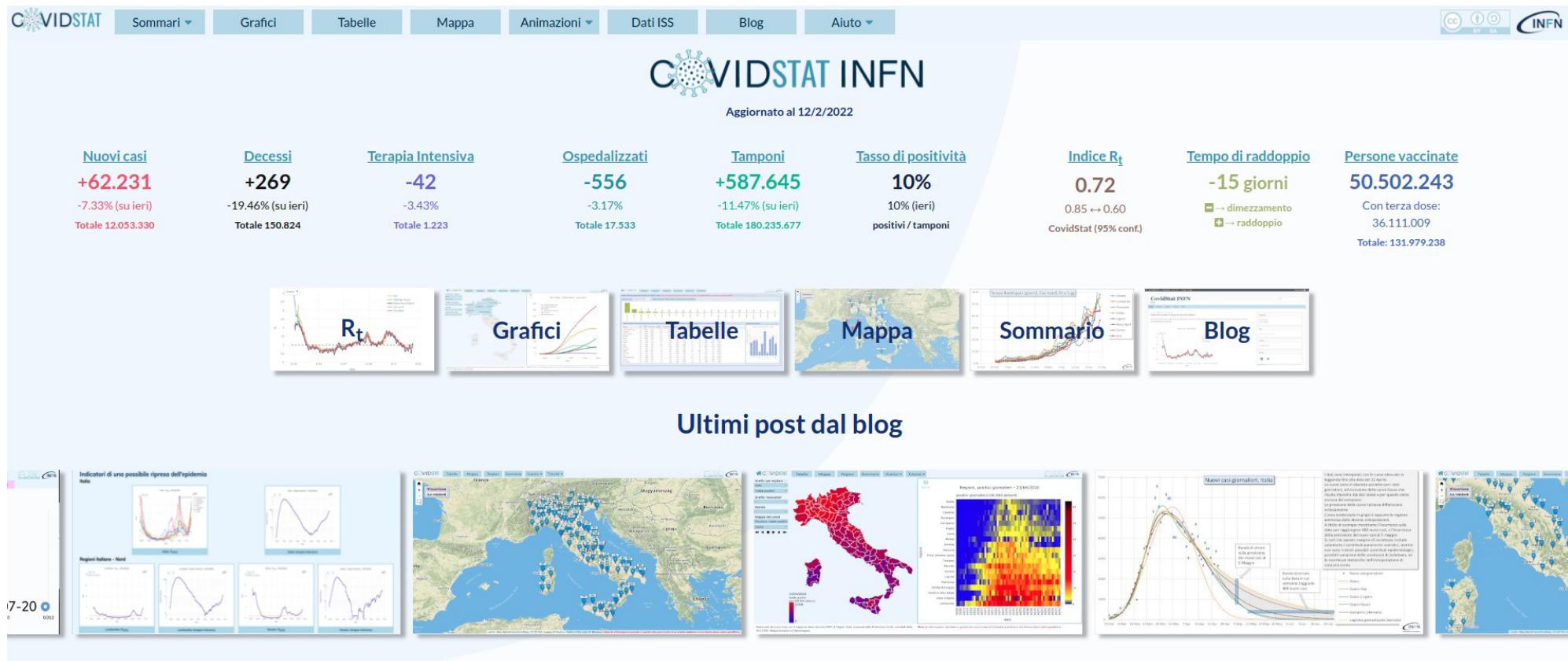


Daily update of Covid-19 Hospital Admissions



CovidStat INFN

<https://covid19.infn.it/>



Features of Covid-19 Actionable Dashboards

<https://www.jmir.org/2021/2/e25682>

JMIR Publications
Advancing Digital Health & Open Science

Articles ▾ Search articles Search

Resource Center ▾ Login Register

Journal of Medical Internet Research Journal Information ▾ Browse Journal ▾ Submit Article

Published on 24.2.2021 in Vol 23, No 2 (2021): February
Preprints (earlier versions) of this paper are available at <https://preprints.jmir.org/preprint/25682>, first published November 13, 2020.



Features Constituting Actionable COVID-19 Dashboards: Descriptive Assessment and Expert Appraisal of 158 Public Web-Based COVID-19 Dashboards

Damir Ivanković ¹ ; Erica Barbazza ¹ ; Véronique Bos ¹ ; Óscar Brito Fernandes ^{1,2} ; Kendall Jamieson Gilmore ³ ; Tessa Jansen ¹ ; Pinar Kara ^{4,5} ; Nicolas Larraín ^{6,7} ; Shan Lu ⁸ ; Bernardo Meza-Torres ^{9,10} ; Joko Mulyanto ^{1,11} ; Mircha Poldrugovac ¹ ; Alexandru Rotar ¹ ; Sophie Wang ^{6,7} ; Claire Willmington ³ ; Yuanhang Yang ^{4,5} ; Zhamin Yelgezekova ¹² ; Sara Allin ¹³ ; Niek Klazinga ¹ ; Dionne Kringos ¹ 

Article	Authors	Cited by (3)	Tweetations (33)	Metrics
---------	---------	--------------	------------------	---------

Abstract

Background: Since the outbreak of COVID-19, the development of dashboards as dynamic, visual tools for communicating COVID-19 data has surged worldwide. Dashboards can inform decision-making and support behavior change. To do so, they must be actionable. The features that constitute an actionable dashboard in the context of the COVID-19 pandemic have not been rigorously assessed.

Objective: The aim of this study is to explore the characteristics of public web-based COVID-19 dashboards by assessing their purpose and users ("why"), content and data ("what"), and analyses and displays ("how" they communicate COVID-19 data), and ultimately to appraise the common features of highly actionable dashboards.

Citation

Please cite as:

Ivanković D, Barbazza E, Bos V, Brito Fernandes Ó, Jamieson Gilmore K, Jansen T, Kara P, Larraín N, Lu S, Meza-Torres B, Mulyanto J, Poldrugovac M, Rotar A, Wang S, Willmington C, Yang Y, Yelgezekova Z, Allin S, Klazinga N, Kringos D
Features Constituting Actionable COVID-19 Dashboards: Descriptive Assessment and Expert Appraisal of 158 Public Web-Based COVID-19 Dashboards
J Med Internet Res 2021;23(2):e25682
doi: 10.2196/25682
PMID: 33577467
PMCID: 7906125

Copy Citation to Clipboard

Export Metadata

END for: Endnote
BibTeX for: BibDesk, LaTeX
RIS for: RefMan, Procite, Endnote, RefWorks
[Add this article to your Mendeley library](#)

This paper is in the following e-collection/theme issue:

JMIR Theme Issue: COVID-19 Special Issue (1394)
Outbreak and Pandemic Preparedness and

Checklist for “actionable” dashboards

1. Know the audience and their information needs
- 2 Manage the type, volume and flow of information
- 3 Make data sources and methods clear
- 4 Link time trends to policy (decisions)
- 5 Provide data ‘close to home’
6. Breakdown the population to relevant sub-groups
- 7 Use story-telling and visual cues

Expanding the vision

One Health is an integrated, unifying approach to balance and optimize the health of people, animals and the environment. It is particularly important to prevent, predict, detect, and respond to global health threats such as the COVID-19 pandemic.

One Health involves the public health, veterinary, public health and environmental sectors. The One Health approach is particularly relevant for food and water safety, nutrition, the control of zoonoses (diseases that can spread between animals and humans, such as flu, rabies and Rift Valley fever), pollution management, and fighting antimicrobial resistance (the emergence of microbes that are resistant to antibiotic therapy).

(World Health Organization)

Planetary health sets the ambitious task of understanding the dynamic and systemic relationships between global environmental changes, their effects on natural systems, and how changes to natural systems affect human health and wellbeing at multiple scales: global (eg, climate), regional (eg, transboundary fire emissions), and local (eg, persistent organic pollutants).

Planetary health, by emphasising interconnections between human health and environmental changes and enabling holistic thinking about overlapping challenges and integrated solutions for present and future generations, offers an opportunity to identify co-benefits across targets, encourage effective cross-sector action and partnerships, and ensure policy coherence.

(United Nations)

Growing topics

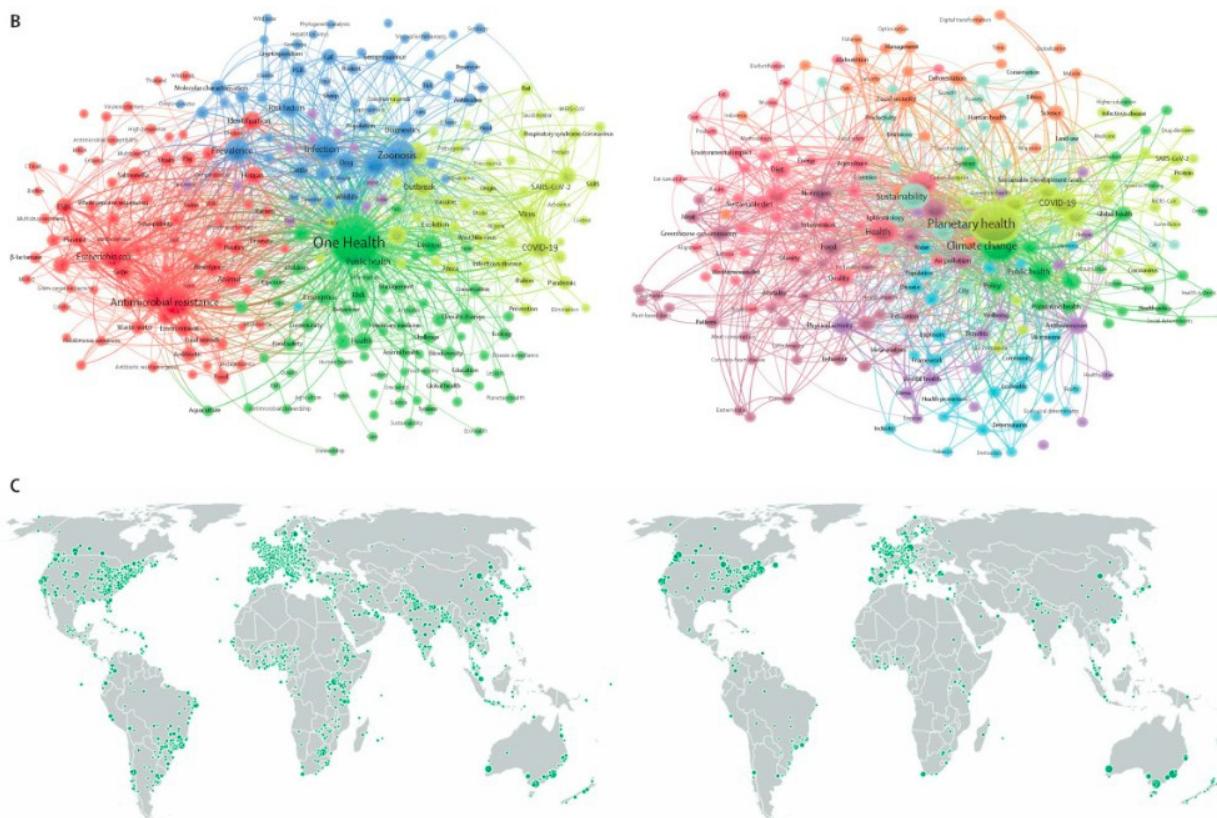
[https://www.thelancet.com/journals/lanplh/article/PIIS2542-5196\(23\)00002-5/fulltext](https://www.thelancet.com/journals/lanplh/article/PIIS2542-5196(23)00002-5/fulltext)

One Health and planetary health research: leveraging differences to grow together



The COVID-19 pandemic and the anthropogenic impact on Earth's life-support systems and planetary boundaries have reinvigorated the One Health and planetary health con
he
bot
thei
aro

To better understand the evolution of One Health and planetary health, we conducted a bibliometric analysis in the Web of Science since the emergence of COVID-19



Lancet Planetary Health 2023

Strong increase in One Health and planetary health research in 2020 and 2021, both in absolute numbers of publications (an increase of 137% for One Health and 170% for planetary health compared with 2018 and 2019 combined) and relative to the total number of publications indexed in Web of Science (figure A). All topics related to infectious diseases were the most represented in One Health publications (eg, COVID-19, antimicrobial resistance, and zoonoses; figure B). Planetary health publications also addressed COVID-19, but climate change was the dominant topic. Non-communicable diseases and issues related to food systems or physical activity and inactivity were part of planetary health, but not One Health, research.

Veterinary Information System

https://www.vetinfo.it/j6_statistiche/index.html#/

National Reference Center for Veterinary Epidemiology
Istituto Zooprofilattico Sperimentale Teramo



Who we are Services Con

Public Health Food Envi

Sistema Informativo Veterinario - Statistiche

Consistenza allevamenti suini Variazioni allevamenti e capi nel tempo Densità allevamenti e capi suini per provincia Densità allevamenti e capi suini per regione Consistenza allevamenti e capi per orientamento produttivo Consistenza allevamenti e capi per tipo di allevamento Consistenza allevi classe di consisti

CONSISTENZA ALLEVAMENTI E CAPI SUINI

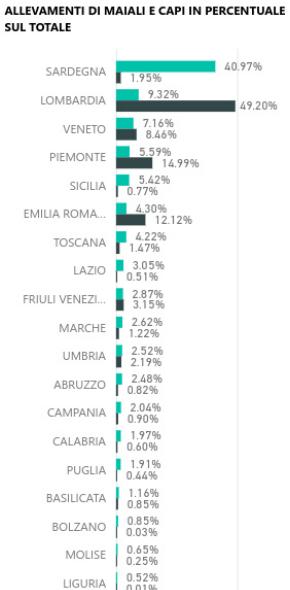
TIPO ATTIVITA
ALLEVAMENTO

DATA RIFERIMENTO
31/12/2022

ORIENTAMENTO PRODUTTIVO

- Select all
- COLLEZIONE FAUNISTICA - DIVERSA DA GIARDINO ZOOLO...
- COLLEZIONE FAUNISTICA - GIARDINO ZOOLOGICO
- DA RIPRODUZIONE (CICLO NON INDICATO)
- DA RIPRODUZIONE A CICLO APERTO
- DA RIPRODUZIONE A CICLO CHIUSO
- FAMILIARE
- NON DPA
- NON INDICATO
- PRODUZIONE DA INGRASSO
- STRUTTURA FAUNISTICA VENATORIA PER CINGHIALI

REGIONE	DATA RIFERIMENTO	NUMERO ALLEVAMENTI	DI CUI CON SOLO CINGHIALI	DI CUI CON MAIALI E CINGHIALI	NUMERO CAPI	DI CUI CINGHIALI	DI CUI MAIALI	DI CUI GRASSI	DI CUI MAGRONI	DI CUI MAGRONCELLI	DI CUI LATTONZOLI	DI CUI SCROFE	DI CUI SCROFETTE	DI CUI VERRI
ABRUZZO	31/12/2022	730	20	8	69,119	61	69,058	15,320	9,274	12,482	23,448	7,416	885	183
BASILICATA	31/12/2022	340	5	3	72,230	51	72,179	12,592	14,438	18,870	19,771	4,309	1,981	195
BOLZANO	31/12/2022	250	0	1	2,458	0	2,458	17	504	299	80	344	1,121	91
CALABRIA	31/12/2022	579	3	5	50,779	28	50,751	9,172	17,868	10,254	8,635	4,272	211	308
CAMPANIA	31/12/2022	600	7	6	76,162	18	76,144	16,205	25,602	9,834	18,311	4,902	1,039	204
EMILIA ROMAGNA	31/12/2022	1,263	46	6	1,024,215	337	1,023,878	325,343	247,652	164,409	215,766	49,228	19,197	530
FRIULI VENEZIA GIULIA	31/12/2022	844	7	5	266,144	65	266,079	75,812	31,939	61,738	71,686	17,211	7,573	120
LAZIO	31/12/2022	896	32	13	42,973	719	42,254	18,640	7,132	5,803	6,771	2,913	506	411
LIGURIA	31/12/2022	153	11	8	450	66	384	93	93	21	14	111	16	36
LOMBARDIA	31/12/2022	2,739	17	2	4,156,583	149	4,156,434	1,212,612	1,053,684	823,110	793,915	223,620	47,542	1,951
MARCHE	31/12/2022	771	23	12	102,846	235	102,611	28,037	19,973	10,237	37,199	4,498	2,513	154
MOLISE	31/12/2022	192	0	0	20,853	0	20,853	7,021	6,272	5,637	869	283	27	50
PIEMONTE	31/12/2022	1,644	54	0	1,266,630	267	1,266,363	445,410	277,853	172,339	303,211	55,264	11,761	522
PUGLIA	31/12/2022	561	4	7	37,229	104	37,125	7,071	6,145	4,573	7,327	1,387	299	282
SARDEGNA	31/12/2022	12,043	68	36	164,776	972	163,804	3,073	42,434	13,679	26,981	61,598	3,959	12,080
SICILIA	31/12/2022	1,593	9	14	65,173	147	65,026	5,181	27,979	8,563	13,978	7,052	929	1,092
TOSCANA	31/12/2022	1,242	197	24	124,240	3,117	121,123	35,513	33,152	11,841	29,929	7,764	2,398	526
TRENTO	31/12/2022	70	0	1	5,774	0	5,774	3,544	857	588	461	320	0	4



GIS Predictions (West Nile Virus)

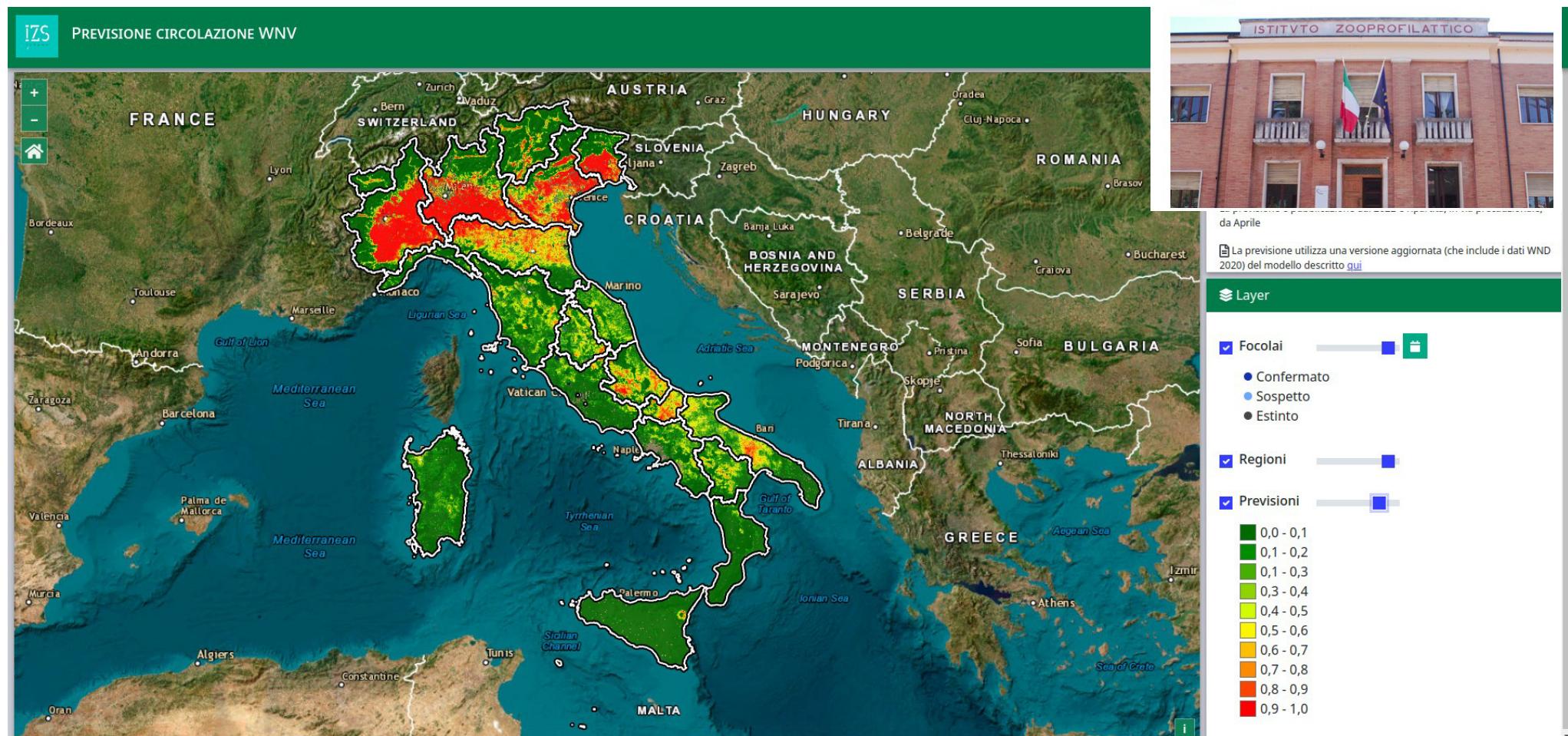
https://mapserver.izs.it/gis_wn_predictions/#

National Reference Center for Veterinary Epidemiology
Istituto Zooprofilattico Sperimentale Teramo



Who we are Services Con

Public Health Food Envi



GIS Contaminated Sites (Terra dei fuochi)

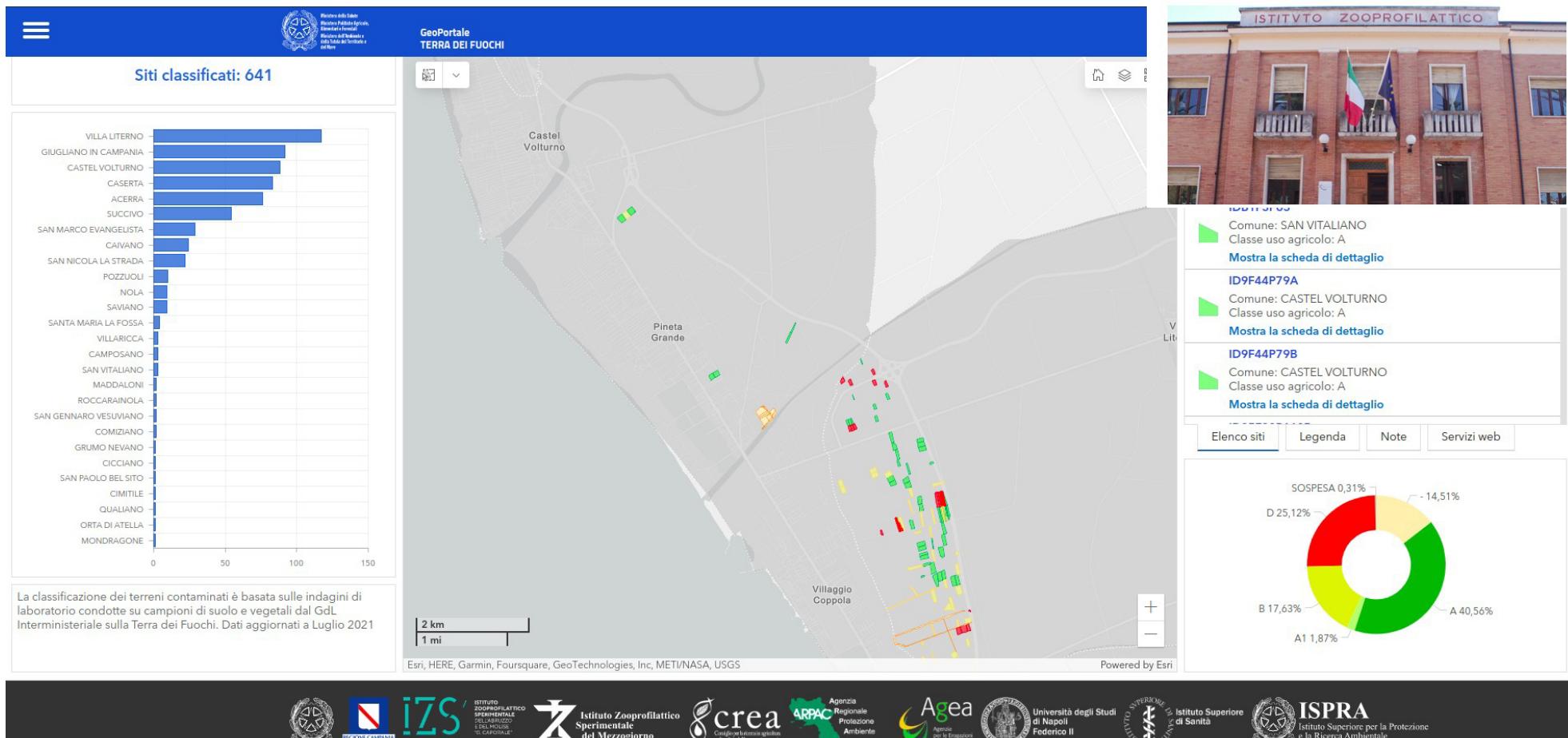
https://terradeifuochi.izs.it/terra_dei_fuochi/geoportale.html

Opportunities for data, training, statistical positions =>



Who we are Services Con

Public Health Food Envi



Key messages

- For their minimal privacy implications, micro-aggregates (e.g. tables by age, sex) have become normally available in the health sector. These type of data can be directly downloaded from the web.
- For research projects making use of routine data, data linkage will become essential so that the potential of health data sources could be fully exploited.
- Dashboards of health indicators have become widely available with Covid-19. The evolution of this systems is quite rapid, and will surely continue in the future. A qualitative assessment of their contents, considering different aspects even related to the statistical methods, will be increasingly needed.
- An expanded vision of health will also open new opportunities to gather data and use statistics for the direct benefit of our global environment.

Materials

- Course notes
- *Web links included in this presentation*



ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Second cycle degree programme (LM) in Statistical Sciences

85278 – Methods and tools for Health Statistics

**85277 – Methods and tools for Official Statistics: Population
and Health Statistics**

a.a. 2022/2023

Stream 1. Regional, national and international health statistics

Topic 1.2.1

Life Expectancy: Measurement issues and recent trends

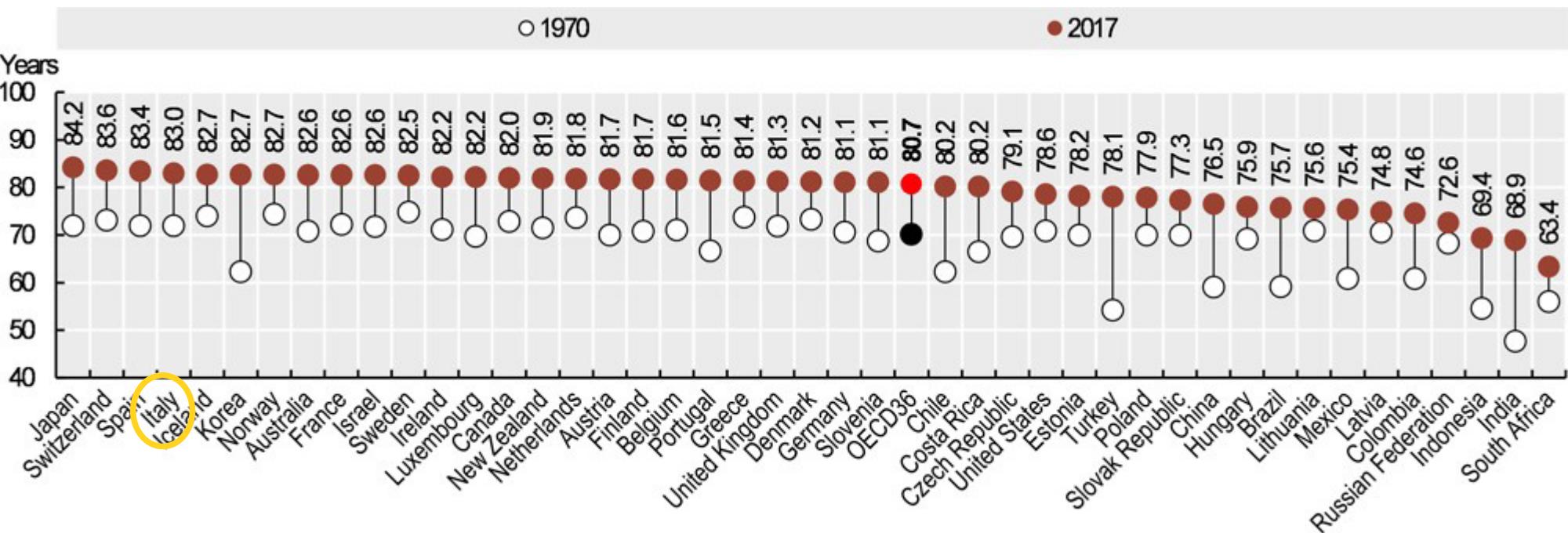
Fabrizio Carinci

fabrizio.carinci@unibo.it

Tuesday, 14th February 2023

Life expectancy at birth, 1970 and 2017 (or nearest years)

Fonte: OECD Health at a Glance 2019



Practical guide

EHEMU Technical report 2006_3 June 2007

**Health Expectancy Calculation by the Sullivan Method:
A Practical Guide**

3rd Edition

The logo features a stylized 'E' composed of vertical bars of varying heights. To the right of the 'E', the text 'European health expectancy monitoring unit' is written in a small, sans-serif font. Below the 'E', the word 'EHEMU' is written in a large, bold, blue serif font. Underneath 'EHEMU', the text 'Observatoire européen des espérances de santé' is written in a smaller, blue serif font.

Calculation method Page 9

[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]
Age group	Mid-year population	No. deaths	Central Death rate	Conditional probability of death	Numbers surviving to age x	Person years lived at age x	Total number of years lived from age x	Total Life Expectancy
x	P _x	D _x	m _x	q _x	l _x	L _x	T _x	e _x
0	54795.5	202	0.003686	0.003606	100000.00	99711.50	8141517.37	81.4
1	54818	21	0.000383	0.000383	99639.37	99620.29	8041805.87	80.7
2	55665.5	11	0.000198	0.000198	99601.21	99591.37	7942185.57	79.7
3	55969.5	8	0.000143	0.000143	99581.53	99574.41	7842594.20	78.8
4	55805.5	12	0.000215	0.000215	99567.30	99556.59	7743019.79	77.8
5	56401.5	8	0.000142	0.000142	99545.89	99538.83	7643463.19	76.8
....
74	49530.5	1059	0.021381	0.021155	80489.15	79637.79	1015838.01	12.6
75	49546.5	1161	0.023433	0.023161	78786.43	77874.04	936200.22	11.9
76	49715	1532	0.030816	0.030348	76961.65	75793.83	858326.18	11.2
77	49928	1887	0.037794	0.037093	74626.01	73241.94	782532.35	10.5
78	48534.5	2042	0.042073	0.041206	71857.87	70377.37	709290.40	9.9
79	45249	2121	0.046874	0.045801	68896.88	67319.12	638913.03	9.3

Life table method in Survival Analysis

- The life table method is also known as the actuarial method.
- The approach is to divide the period of observation into a series of time intervals, and estimate the conditional (interval-specific) survival rate for each interval.
- The cumulative survivor function, $S(t)$, at the end of a specified interval is then given by the product of the interval-specific survival rates for all intervals up to and including the specified interval.
- A “censored” observation is one whose status at the start of the interval is known, but is not at the end (eg lost to follow-up for different reasons).

Life table method calculations

- In the absence of censoring, the interval-specific survival rate is:

$$p = (l - d) / l$$

where d is the number of events (deaths) observed during the interval and l is the number of patients alive at the start of the interval.

- In the presence of censoring, it is assumed that censoring occurs at random throughout the interval so that each individual with a censored survival time is at risk for, on average, **half** of the interval. This assumption is known as the *actuarial assumption*.
- The **actual number of patients at risk during the interval** is given by

$$l' = l - 0.5w$$

l = number of patients alive at the start of the interval

w = number of censored observations during the interval

Life table method: example (1)

time	l	d	w	l'	p	$S(t)$
0 – 1)	35	8	0	35.0	0.771	0.771
1 – 2)	27	2	2	26.0	0.923	0.712
2 – 3)	23	5	4	21.0	0.762	0.543
3 – 4)	14	2	1	13.5	0.852	0.462
4 – 5)	11	0	1	10.5	1.000	0.462
5 – 6)	10	0	0	10.0	1.000	0.462
6 – 7)	10	0	3	8.5	1.000	0.462
7 – 8)	7	0	1	6.5	1.000	0.462
8 – 9)	6	2	3	4.5	0.556	0.257
9 – 10)	1	0	1	0.5	1.000	0.257

l is the number alive at the start of the interval

d is the number of events (deaths) during the interval

w is the number of censored observations during the interval

l' is the effective number at risk for the interval

p is the interval-specific survival probability

$S(t)$ is the estimated **cumulative survival function** at the end of the interval

Life table method: example (2)

- The interval-specific survival rate is equal to $p = (l' - d)/l'$.
- For the first interval, $l = l' = 35$ and $p = (35 - 8)/35 = 0.771$.
- The estimated 1-year survival rate is therefore $\hat{S}(1) = 0.771$.
- For the second interval, $l' = 27 - 0.5 \times 2 = 26$ and $p = (26 - 2)/26 = 0.923$. The estimated 2-year survival rate is then
$$\hat{S}(2) = 0.771 \times 0.923 = 0.712.$$
- The **cumulative survival rate** is estimated as the product of conditional survival rates, where the estimate of each conditional survival rate is based upon only those individuals under follow-up.

Type of Life Tables

- **Cohort or Generation Life Table:** the table is constructed using a sequence that is directly derived from the natural longitudinal life experience of the cohort until expiration (no subject alive anymore).
This may be “true” but not adequately representing the current conditions. *“A study providing an observed life expectancy would now require a follow-up of about 115 years, and this would apply to the past 115 years, not the future” (Singer 2005)*
- **Period Life Table:** Data for a single cross-section of time, representing current mortality patterns by age (single year: “unabridged”, classes of age: “abridged”). *“Their values of life expectancy are absolutely dependent on the forecast that the mortality rates in the table will remain constant. That the forecast is patently wrong is obvious from all the experience for more than 170 years of national life table preparation” (Singer 2005)*

Life Expectancy from Life Table (1)

Cross-sectional

[1] Age group	[2] Mid-year population	[3] No. deaths	[4] Central Death rate	[5] Conditional probability of death
x	P_x	D_x	m_x	q_x
0	54795.5	202	0.003686	0.003606
1	54818	21	0.000383	0.000383
2	55665.5	11	0.000198	0.000198
3	55969.5	8	0.000143	0.000143
4	55805.5	12	0.000215	0.000215
5	56401.5	8	0.000142	0.000142
.....	
74	49530.5	1059	0.021381	0.021155
75	49546.5	1161	0.023433	0.023161
76	49715	1532	0.030816	0.030348
77	49928	1887	0.037794	0.037093
78	48534.5	2042	0.042073	0.041206
79	45249	2121	0.046874	0.045801

$$q_x = \frac{m_x}{1+(1-a_x)m_x} \quad a_x = 0.5$$

Probability of death in each age interval conditional on having survived to that age
(basically we add half of the deaths to the denominator)

$$(1059/49530.5) / (1+0.5*(1059/49530.5))=0.021154614$$

Life Expectancy from Life Table (2)

Cross-sectional

Longitudinal

$$L_x = l_{x+1} + 0.5(l_x - l_{x+1}) = \frac{l_x + l_{x+1}}{2}$$



number surviving to age l_x

number arriving to age l_{x+1}

[1]	[2]	[3]	[4]	[5]	[6]	[7]
Age group	Mid-year population	No. deaths	Central Death rate	Conditional probability of death	Numbers surviving to age x	Person years lived at age x
x	P_x	D_x	m_x	q_x	ℓ_x	L_x
0	54795.5	202	0.003686	0.003606	100000.00	99711.50
1	54818	21	0.000383	0.000383	99639.37	99620.29
2	55665.5	11	0.000198	0.000198	99601.21	99591.37
3	55969.5	8	0.000143	0.000143	99581.53	99574.41
4	55805.5	12	0.000215	0.000215	99567.30	99556.59
5	56401.5	8	0.000142	0.000142	99545.89	99538.83
.....
74	49530.5	1059	0.021381	0.021155	80489.15	79637.79
75	49546.5	1161	0.023433	0.023161	78786.43	77874.04
76	49715	1532	0.030816	0.030348	76961.65	75793.83
77	49928	1887	0.037794	0.037093	74626.01	73241.94
78	48534.5	2042	0.042073	0.041206	71857.87	70377.37
79	45249	2121	0.046874	0.045801	68896.88	67319.12

$$L_{74} = 78786.43 + 0.5 * (80489.15 - 78786.43) = 0.5 * (80489.15 + 78786.43) = 79637.79$$

Life Expectancy from Life Table (3)

EUROSTAT before age of 1: $L_x = 0.2 * l_x + 0.8 * l_{x+1}$

[1]	[2]	[3]	[4]	[5]	[6]	[7]
Age group	Mid-year population	No. deaths	Central Death rate	Conditional probability of death	Numbers surviving to age x	Person years lived at age x
x	P_x	D_x	m_x	q_x	ℓ_x	L_x
0	54795.5	202	0.003686	0.003606	100000.00	99711.50
1	54818	21	0.000383	0.000383	99639.37	99620.29
2	55665.5	11	0.000198	0.000198	99601.21	99591.37
3	55969.5	8	0.000143	0.000143	99581.53	99574.41
4	55805.5	12	0.000215	0.000215	99567.30	99556.59
5	56401.5	8	0.000142	0.000142	99545.89	99538.83
.....
74	49530.5	1059	0.021381	0.021155	80489.15	79637.79
75	49546.5	1161	0.023433	0.023161	78786.43	77874.04
76	49715	1532	0.030816	0.030348	76961.65	75793.83
77	49928	1887	0.037794	0.037093	74626.01	73241.94
78	48534.5	2042	0.042073	0.041206	71857.87	70377.37
79	45249	2121	0.046874	0.045801	68896.88	67319.12

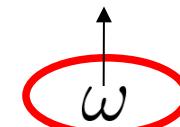
Makes estimation more precise based on number survived until year 1, given the higher mortality rates during the first months of life

$$L_0 = 0.2 * 100000 + 0.8 * 99639.37 = 99711.496$$

Life Expectancy from Life Table (4)

$$T_x = \sum_{i=x}^{\omega} L_i$$

[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]
Age group	Mid-year population	No. deaths	Central Death rate	Conditional probability of death	Numbers surviving to age x	Person years lived at age x	Total number of years lived from age x
x	P _x	D _x	m _x	q _x	ℓ _x	L _x	T _x
0	54795.5	202	0.003686	0.003606	100000.00	99711.50	8141517.37
1	54818	21	0.000383	0.000383	99639.37	99620.29	8041805.87
2	55665.5	11	0.000198	0.000198	99601.21	99591.37	7942185.57
3	55969.5	8	0.000143	0.000143	99581.53	99574.41	7842594.20
4	55805.5	12	0.000215	0.000215	99567.30	99556.59	7743019.79
5	56401.5	8	0.000142	0.000142	99545.89	99538.83	7643463.19
.....
74	49530.5	1059	0.021381	0.021155	80489.15	79637.79	1015838.01
75	49546.5	1161	0.023433	0.023161	78786.43	77874.04	936200.22
76	49715	1532	0.030816	0.030348	76961.65	75793.83	858326.18
77	49928	1887	0.037794	0.037093	74626.01	73241.94	782532.35
78	48534.5	2042	0.042073	0.041206	71857.87	70377.37	709290.40
79	45249	2121	0.046874	0.045801	68896.88	67319.12	638913.03


 ω

Life Expectancy from Life Table (5)

$$e_x = \frac{1}{l_x} T_x$$

[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]
Age group	Mid-year population	No. deaths	Central Death rate	Conditional probability of death	Numbers surviving to age x	Person years lived at age x	Total number of years lived from age x	Total Life Expectancy
x	P _x	D _x	m _x	q _x	l _x	L _x	T _x	e _x
0	54795.5	202	0.003686	0.003606	100000.00	99711.50	8141517.37	81.4
1	54818	21	0.000383	0.000383	99639.37	99620.29	8041805.87	80.7
2	55665.5	11	0.000198	0.000198	99601.21	99591.37	7942185.57	79.7
3	55969.5	8	0.000143	0.000143	99581.53	99574.41	7842594.20	78.8
4	55805.5	12	0.000215	0.000215	99567.30	99556.59	7743019.79	77.8
5	56401.5	8	0.000142	0.000142	99545.89	99538.83	7643463.19	76.8
.....
74	49530.5	1059	0.021381	0.021155	80489.15	79637.79	1015838.01	12.6
75	49546.5	1161	0.023433	0.023161	78786.43	77874.04	936200.22	11.9
76	49715	1532	0.030816	0.030348	76961.65	75793.83	858326.18	11.2
77	49928	1887	0.037794	0.037093	74626.01	73241.94	782532.35	10.5
78	48534.5	2042	0.042073	0.041206	71857.87	70377.37	709290.40	9.9
79	45249	2121	0.046874	0.045801	68896.88	67319.12	638913.03	9.3

This is an average for the entire population surviving at each time

So you have that value calculated for each age!

$$e_2 = (1/99601.21) * 7942185.57 = 79.739850249$$

Disability Free Life Expectancy from Life Table

$$\text{DFLE}_x = \frac{1}{l_x} \sum_{i=x}^{\omega} L_i(DF)$$

Table 1.5 Calculation of Disability-Free Life Expectancy (DFLE) by the Sullivan method using a single-year life table (method 1)

[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]	[10]	[11]	[12]	[13]	[14]
Age	Mid-year population	No. deaths	Central Death rate	Conditional probability of death	Numbers surviving to age x	Person years lived at age x	Total number of years lived from age x	Total Life Expectancy	Proportion with disability	Person years lived without disability at age x	Total years lived without disability from age x	Disability-free life expectancy	Prop. of life spent disability free
X	P_x	D_x	m_x	q_x	ℓ_x	L_x	T_x	e_x	π_x	$[1-\pi_x]L_x$	$\Sigma[1-\pi_x]L_x$	DFLE_x	%DFLE/ e_x
0	54795.5	202	0.003686	0.003606	100000.00	99711.50	8141517.37	81.4	0	99711.50	6657215.85	66.6	81.8
1	54818	21	0.000383	0.000383	99639.37	99620.29	8041805.87	80.7	0.048	94838.51	6557604.35	65.8	81.5
2	55665.5	11	0.000198	0.000198	99601.21	99591.37	7942185.57	79.7	0.048	94810.99	6462763.83	64.9	81.4
3	55969.5	8	0.000143	0.000143	99581.53	99574.41	7842594.20	78.8	0.048	94794.84	6367954.85	63.9	81.2
4	55805.5	12	0.000215	0.000215	99567.30	99556.59	7743019.79	77.8	0.048	94777.88	6273160.00	63.0	81.0
5	56401.5	8	0.000142	0.000142	99545.89	99538.83	7643463.19	76.8	0.030	96552.67	6178382.13	62.1	80.8
.....
74	49530.5	1059	0.021381	0.021155	80489.15	79637.79	1015838.01	12.6	0.345	52162.75	562021.21	7.0	55.3
75	49546.5	1161	0.023433	0.023161	78786.43	77874.04	936200.22	11.9	0.431	44310.33	509858.45	6.5	54.5
76	49715	1532	0.030816	0.030348	76961.65	75793.83	858326.18	11.2	0.431	43126.69	465548.12	6.0	54.2
--	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----

$$(1/99639.37) * 6557604.35 = 65.813386315$$

Recent trends in life expectancy across high income countries: retrospective observational study

Jessica Y Ho,¹ Arun S Hendi²

Women, 2014-15

Men, 2014-15

ABSTRACT

OBJECTIVE

To assess whether declines in life expectancy occurred across high income countries during 2014-16, to identify the causes of death contributing to these declines, and to examine the extent to which these declines were driven by shared or differing factors across countries.

DESIGN

Demographic analysis using aggregated data.

SETTING

Vital statistics systems of 18 member countries of the Organisation for Economic Co-operation and Development.

PARTICIPANTS

18 countries with high quality all cause and cause specific mortality data available in 2014-16.

MAIN OUTCOME MEASURES

Life expectancy at birth, 0-65 years, and 65 or more years and cause of death contributions to changes in life expectancy at birth.

RESULTS

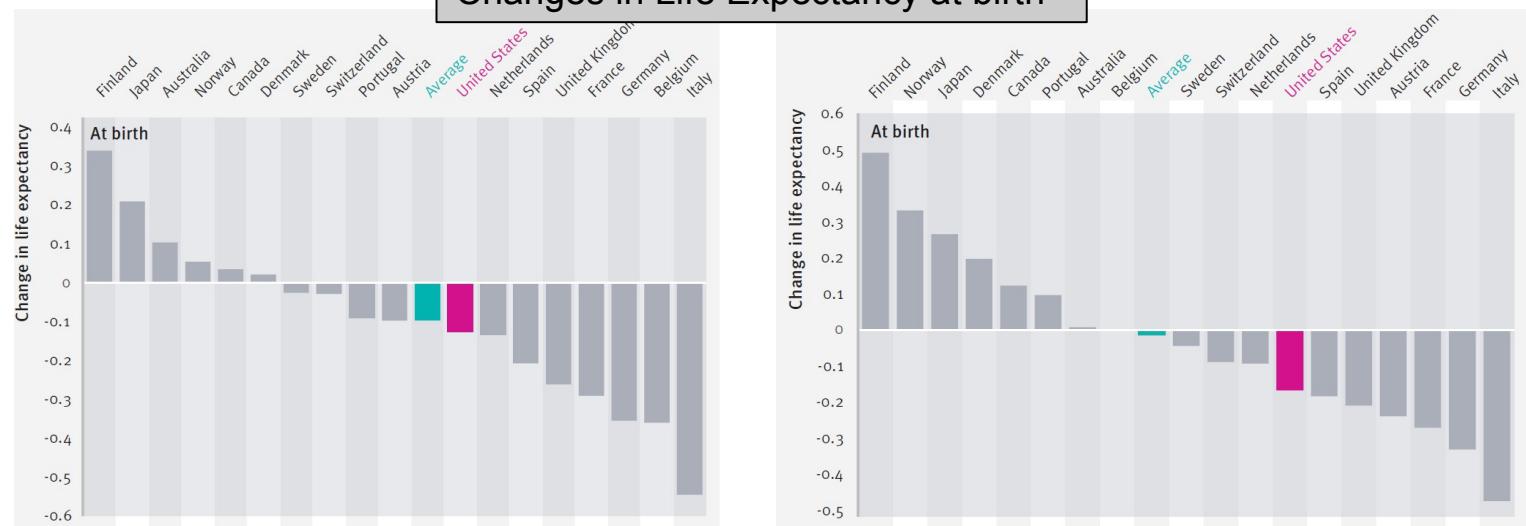
The majority of high income countries in the study experienced declines in life expectancy during 2014-15; of the 18 countries, 12 experienced declines in life expectancy among women and 11 experienced declines in life expectancy among men. The average decline was 0.21 years for women and 0.18 years for men. In most countries experiencing declines in life expectancy, these declines were predominantly driven by trends in older age (≥ 65 years) mortality and in deaths related to respiratory disease,

cardiovascular disease, nervous system disease, and mental disorders. In the United States, declines in life expectancy were more concentrated at younger ages (0-65 years), and drug overdose and other external causes of death played important roles in driving these declines.

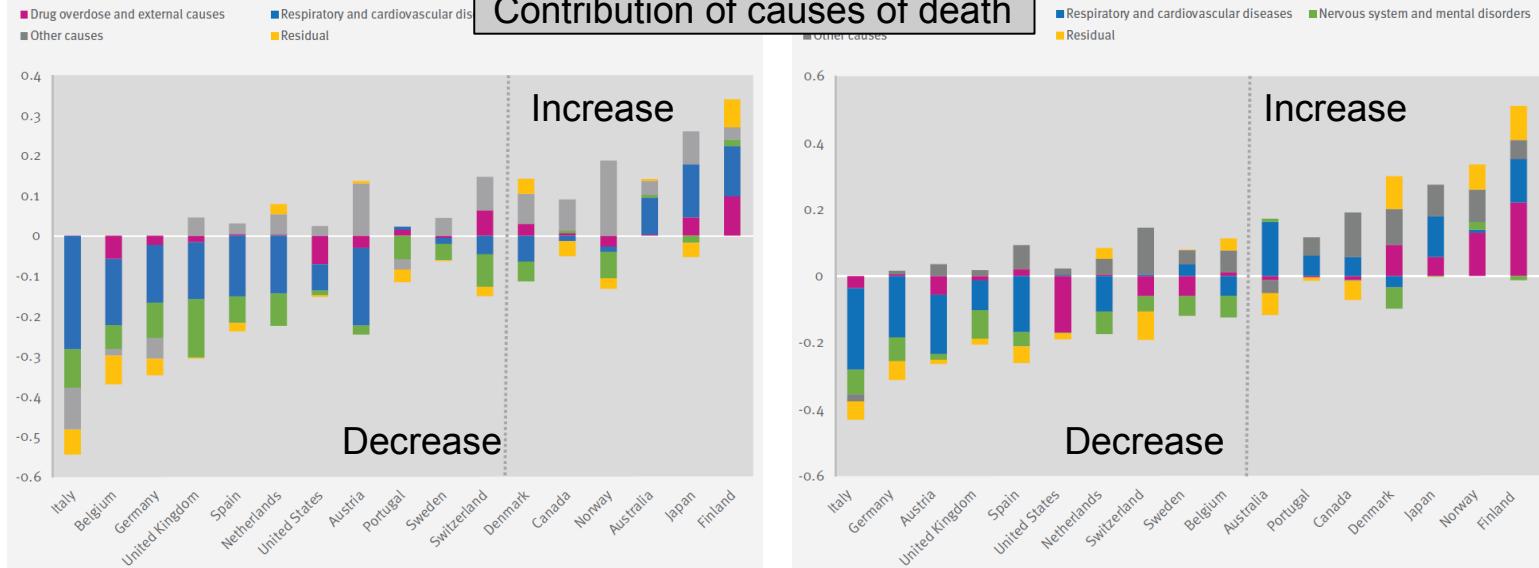
CONCLUSIONS

Most of the countries that experienced declines in life expectancy during 2014-15 experienced robust gains in life expectancy during 2015-16 that more than compensated for the declines. However, the United Kingdom and the United States appear to be experiencing stagnating or continued declines in life expectancy, raising questions about future trends in these countries.

Changes in Life Expectancy at birth



Contribution of causes of death



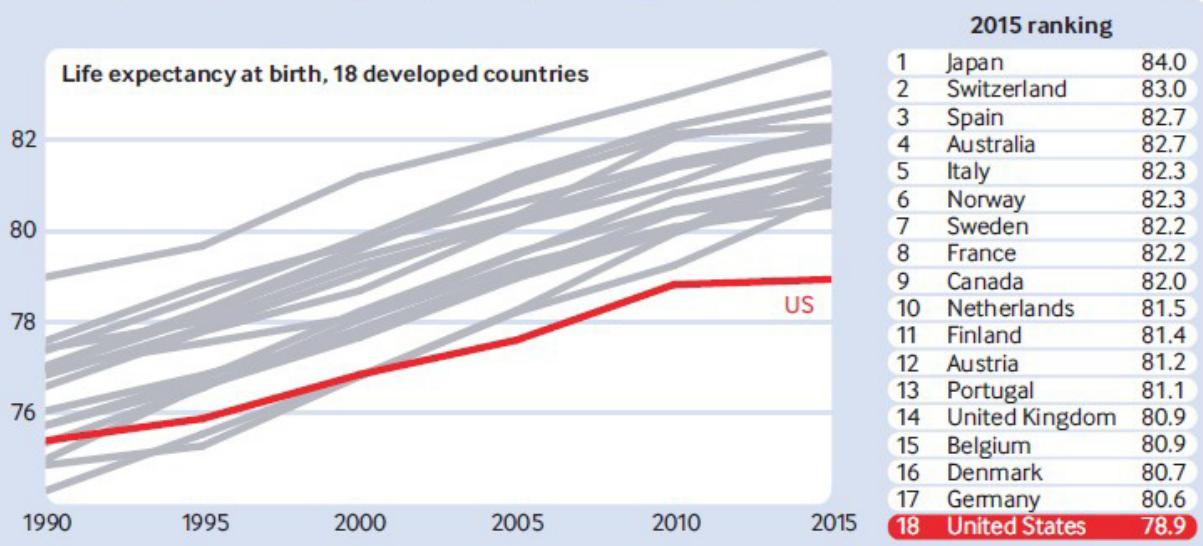
Why is US life expectancy falling behind?



Coming in last

The United States now ranks near the bottom of life expectancy rankings, when compared to other high income countries.

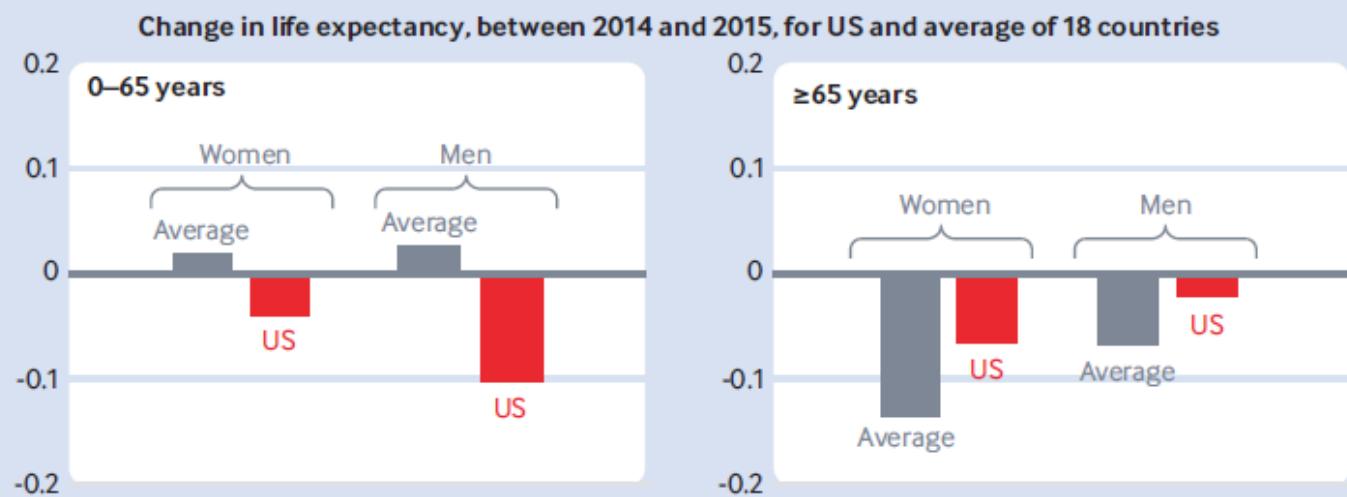
In a 2018 paper in *The BMJ*, authors Ho and Hendi compared life expectancy trends from 1990 to 2015 in 18 countries commonly used in cross national comparisons. These countries have all achieved high levels of development, and underwent changes in mortality associated with that development at roughly the same time. They also have large enough populations to produce reliable estimates of mortality.



Young vs old generations

Before their time

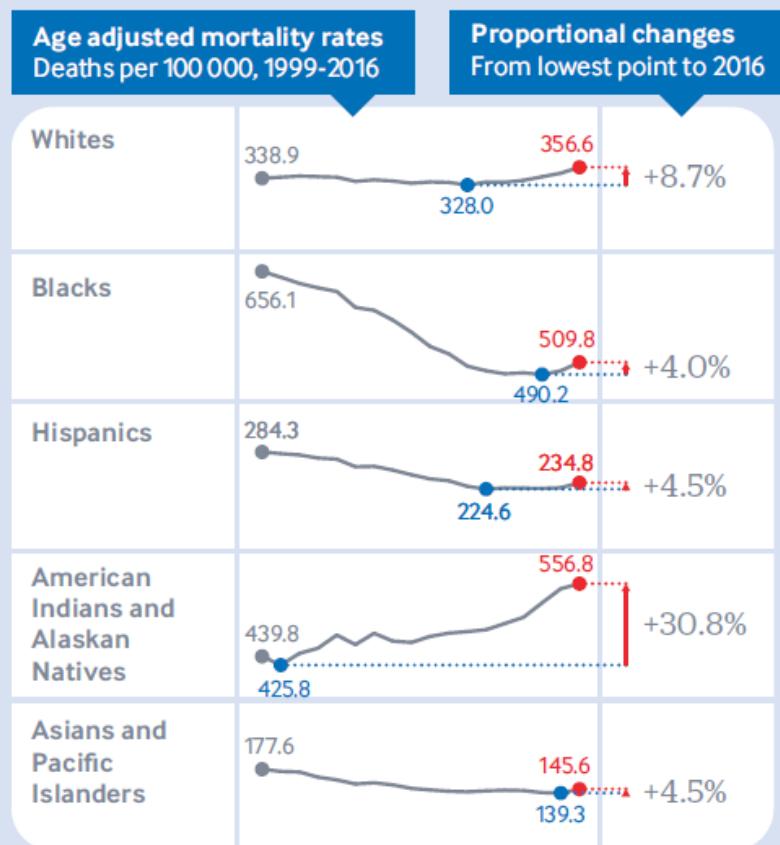
Ho and Hendi observed recent widespread life expectancy declines across the 18 high income countries. The decline in most countries was concentrated at ages ≥ 65 , and mostly attributable to diseases related to a severe influenza season. However, the US decline was largely concentrated at younger ages, particularly those in their 20s and 30s, and attributable to external causes like drug overdose.



Causes of concern for life expectancy in the US

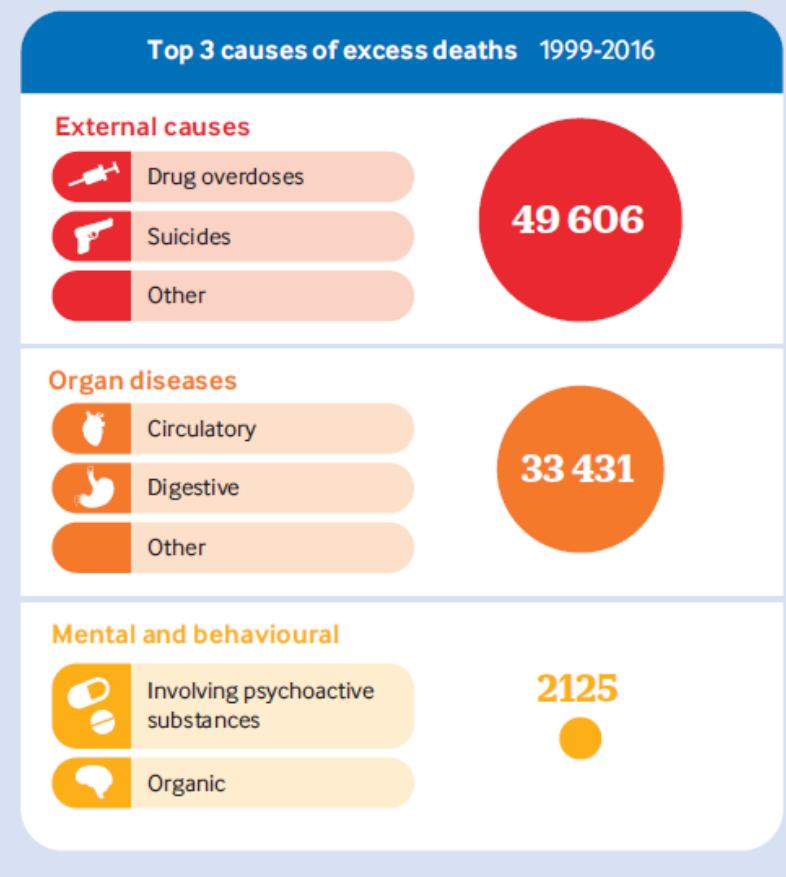
Who is affected?

Further detail is provided by Woolf et al, in their simultaneously published paper in *The BMJ*. They compared midlife mortality patterns in the US across racial and ethnic groups from 1999 to 2016. Among people aged 25-64 years, increases in mortality rates have been observed in all groups in recent years.



Cause for concern

Within these groups, there are a variety of different reasons for the observed changes in mortality. Changes were driven not only by external causes of death, but also by a variety of organ diseases and increases in mortality from mental and behavioral disorders.



Covid-19 and Life Expectancy

Reductions in 2020 US life expectancy due to COVID-19 and the disproportionate impact on the Black and Latino populations

Theresa Andrasfay^{a,1}  and Noreen Goldman^b 

^aLeonard Davis School of Gerontology, University of Southern California, Los Angeles, CA 90089; and ^bOffice of Population Research, Princeton University, Princeton, NJ 08544

Edited by James W. Vaupel, University of Southern Denmark, Odense, Denmark, and approved December 8, 2020 (received for review July 15, 2020)

COVID-19 has resulted in a staggering death toll in the United States: over 215,000 by mid-October 2020, according to the Centers for Disease Control and Prevention. Black and Latino Americans have experienced a disproportionate burden of COVID-19 morbidity and mortality, reflecting persistent structural inequalities that increase risk of exposure to COVID-19 and mortality risk for those infected. We estimate life expectancy at birth and at age 65 y for 2020, for the total US population and by race and ethnicity, using four scenarios of deaths—one in which the COVID-19 pandemic had not occurred and three including COVID-19 mortality projections produced by the Institute for Health Metrics and Evaluation. Our medium estimate indicates a reduction in US life expectancy at birth of 1.13 y to 77.48 y, lower than any year since 2003. We also project a 0.87-y reduction in life expectancy at age 65 y. The Black and Latino populations are estimated to experience declines in life expectancy at birth of 2.10 and 3.05 y, respectively, both of which are several times the 0.68-y reduction for Whites. These projections imply an increase of nearly 40% in the Black–White life expectancy gap, from 3.6 y to over 5 y, thereby eliminating progress made in reducing this differential since 2006. Latinos, who have consistently experienced lower mortality than Whites (a phenomenon known as the Latino or Hispanic paradox), would see their more than 3-y survival advantage reduced to less than 1 y.

are younger than the White population and, all else being equal, would have fewer deaths (4, 5).

In the period preceding the COVID-19 pandemic, annual improvements in US life expectancy had been small—for example, an increase from 76.8 y to 78.9 y or an average annual increase of 0.15 y between 2000 and 2014—but overall life expectancy has rarely declined (6).§ The recent declines that have taken place have attracted enormous attention from researchers and the media. Annual reductions of 0.1 y for each of three consecutive years (2015, 2016, and 2017) (7–9), attributed partly to increases in “deaths of despair” (6), made repeated headlines as the longest period of decrease since the 1918 influenza pandemic. Conversely, a 0.1-y recovery in life expectancy in 2018 was greeted with substantial relief (10).

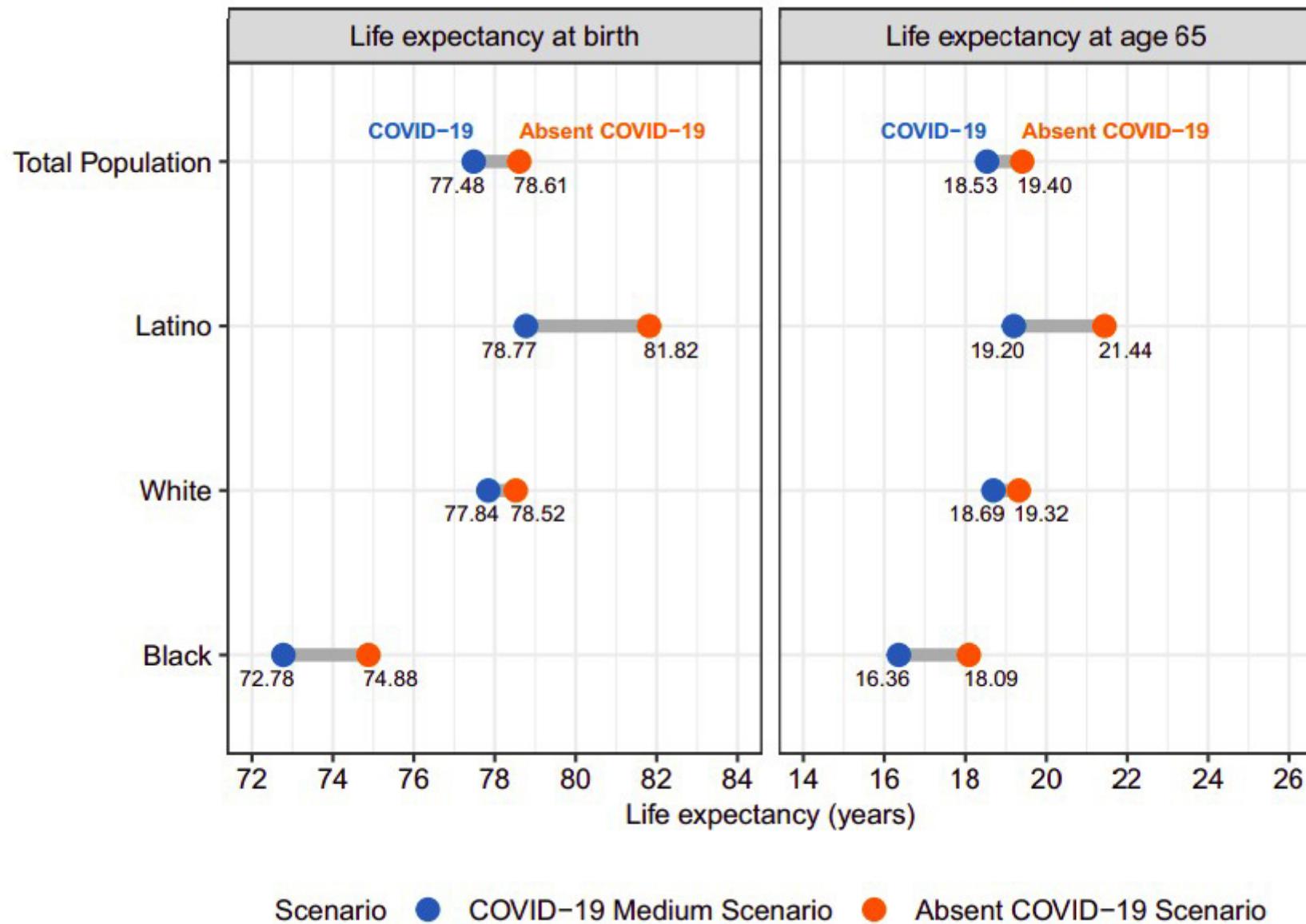
Black Americans have consistently had lower life expectancy than Whites, but relative gains in life expectancy over the past two decades have been greater in the Black population than among Whites, thereby narrowing the Black mortality disadvantage (11,

Significance

COVID-19 has generated a huge mortality toll in the United States, with the highest rates among Black and Latino Americans.

SOCIAL SCIENCES

Covid-19 and Life Expectancy Stratified



A political measure



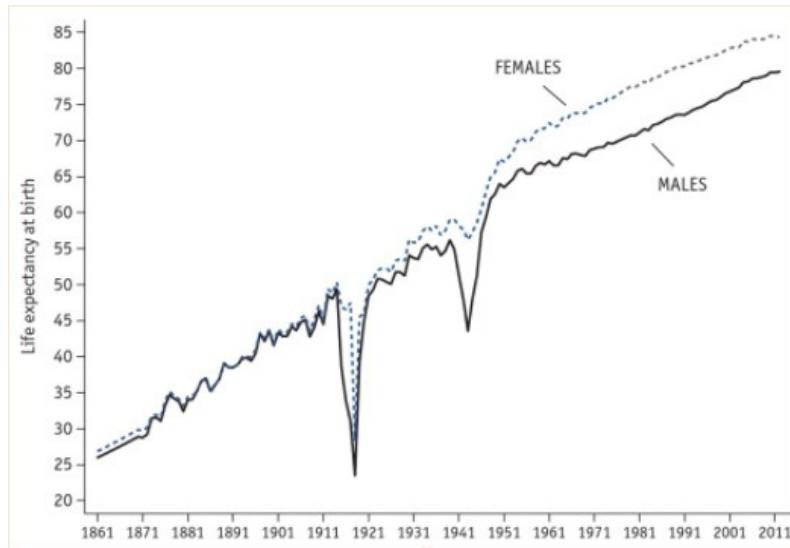
“Life expectancy, due to the pandemic, has decreased: up to 4 – 5 years in areas of greatest contagion: a year and a half – two less for the entire Italian population. A similar decline has not been recorded in Italy since the two world wars”

M.Draghi addressing the Italian Senate, 17th February 2021

Preliminary estimates

Based on incomplete data for year 2020, not including “second wave”

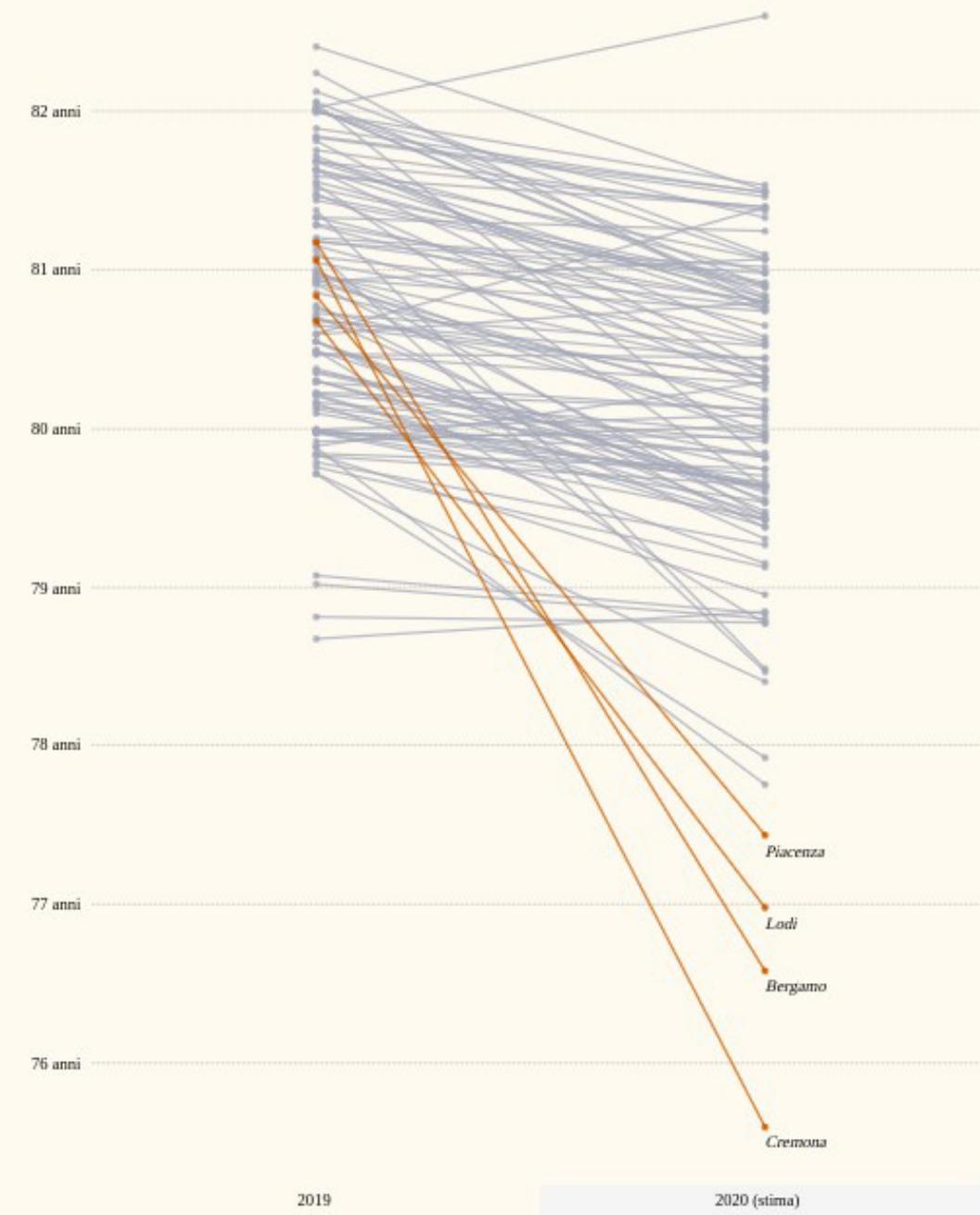
Prof. Stefano Mazzucco, University of Padova



Source: Il Sole 24 Ore, 7 December 2020

https://www.infodata.ilsole24ore.com/2020/12/07/demografia-nel-2020-la-speranza-di-vita-degli-italiani-calera-a-causa-del-covid-19/?refresh_ce=1

La speranza di vita maschile alla nascita valori per provincia, 2019 e stima 2020 (nota bene: la seconda ondata di COVID-19 non è ancora inclusa)



Islam N et al. BMJ. 2021 Nov 3;375:e066768. doi: 10.1136/bmj-2021-066768.



OPEN ACCESS



Check for updates

Effects of covid-19 pandemic on life expectancy and premature mortality in 2020: time series analysis in 37 countries

Nazrul Islam,¹ Dmitri A Jdanov,^{2,3} Vladimir M Shkolnikov,^{2,3} Kamlesh Khunti,^{4,5} Ichiro Kawachi,⁶ Martin White,⁷ Sarah Lewington,^{1,8} Ben Lacey¹

WHAT IS ALREADY KNOWN ON THIS TOPIC

Reported numbers of deaths with covid-19 are subject to changes within and across countries as well as some degrees of delays, inaccuracy, and incompleteness

Excess deaths (difference between observed and expected numbers of deaths from all causes) allows the assessment of the full impact of the pandemic, including the direct effect on deaths with covid-19, and the indirect effect of the pandemic on deaths from other diseases

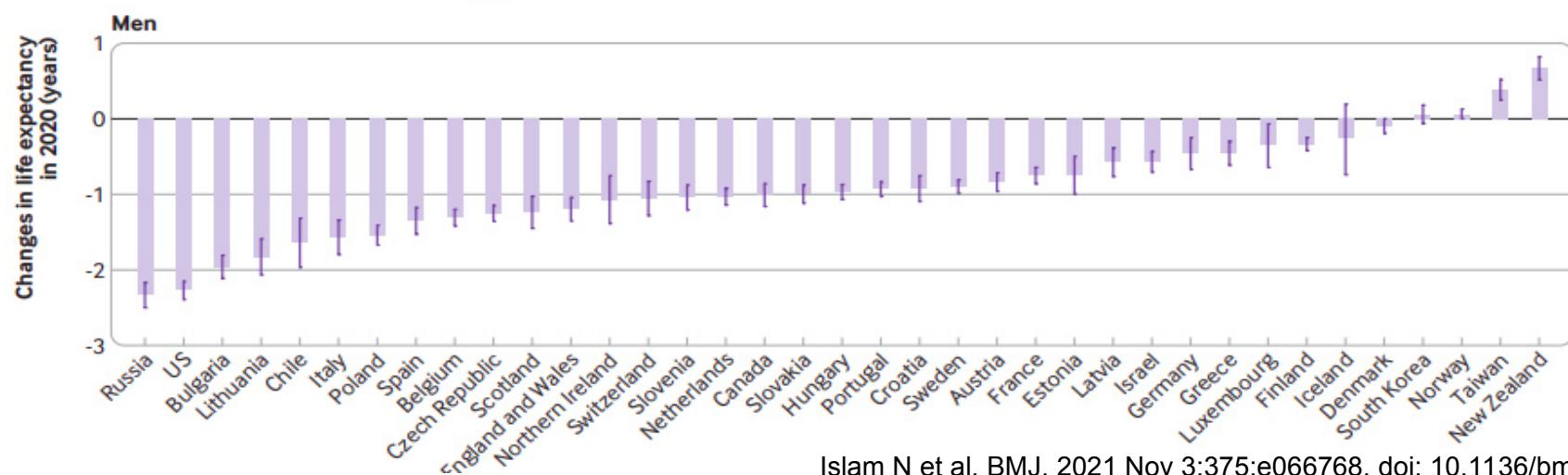
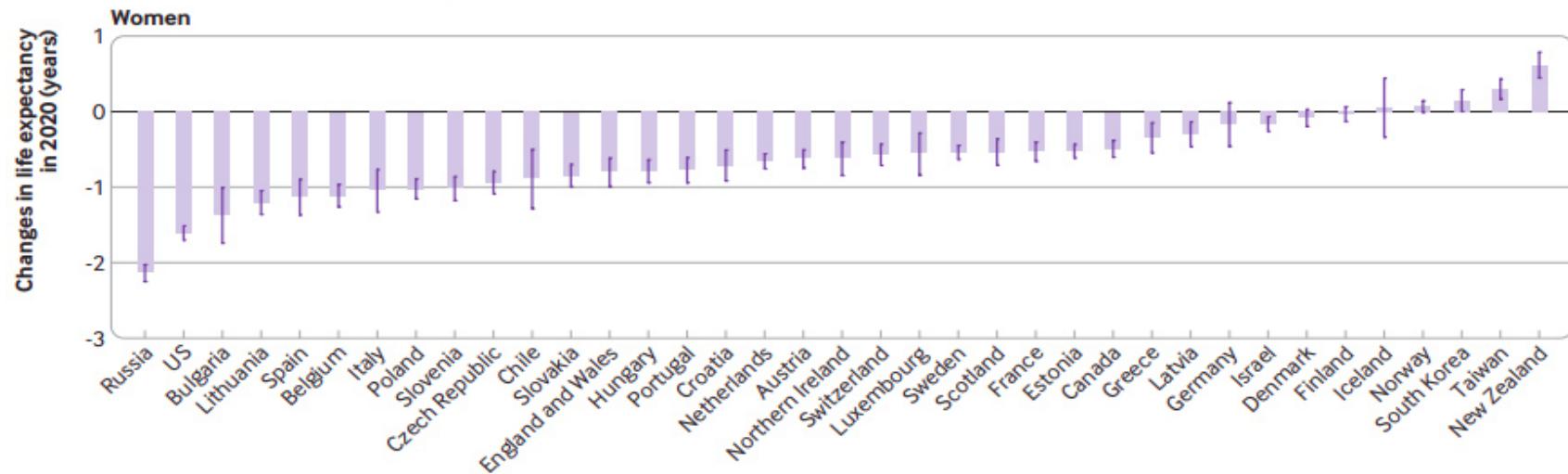
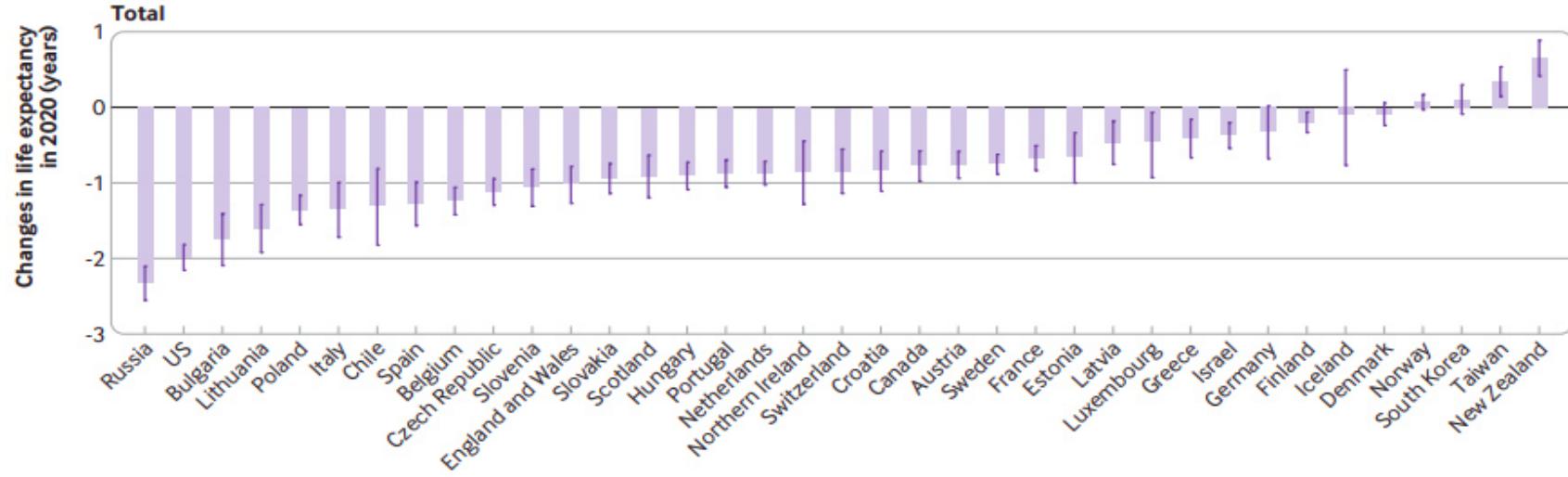
Estimation of excess deaths does not, however, consider the age at death, and therefore does not quantify the impact of the pandemic on premature deaths as years of life lost (YLL)

WHAT THIS STUDY ADDS

In 2020, life expectancy was lower and YLL higher than expected in all countries except New Zealand, Taiwan, Iceland, South Korea, Denmark, and Norway—in the remaining 31 countries, >28 million excess years of life were lost

Highest reduction in life expectancy in 2020 was observed in Russia (men, -2.33 years; women, -2.14), the US (men, -2.27; women, -1.61), Bulgaria (men -1.96; women, -1.37), Lithuania (men, -1.83; women, -1.21), Chile (men, -1.64; women, -0.88), and Spain (men, -1.35; women, -1.13)

Excess YLL rates associated with the covid-19 pandemic in 2020 were more than five times higher than those associated with the seasonal influenza epidemic in 2015



Key messages

- The life table method is the canonical basis for the calculation of life expectancy by EUROSTAT and international organizations. The estimates of life expectancy depend from the assumption that mortality rates in the table will remain constant.
- Life expectancy can be conveniently used to compare the state of health of populations around the world. A word of caution is needed when the measure is used to address national policy, since the conditions may rapidly evolve.
- Covid-19 is likely to have a huge impact on life expectancy during the pandemics. However, the recovery should be fast enough to gain the years lost in a relatively short period of time.

Materials

- Health Expectancy Calculation by the Sullivan Method: A Practical Guide.
3rd Edition
- Ho JY et al, Recent trends in life expectancy across high income countries,
<https://www.bmjjournals.org/content/362/bmj.k2562>



ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Second cycle degree programme (LM) in Statistical Sciences

85278 – Methods and tools for Health Statistics

**85277 – Methods and tools for Official Statistics: Population
and Health Statistics**

a.a. 2022/2023

Stream 1. Regional, national and international health statistics

Topic 1.2.2

Health systems performance assessment

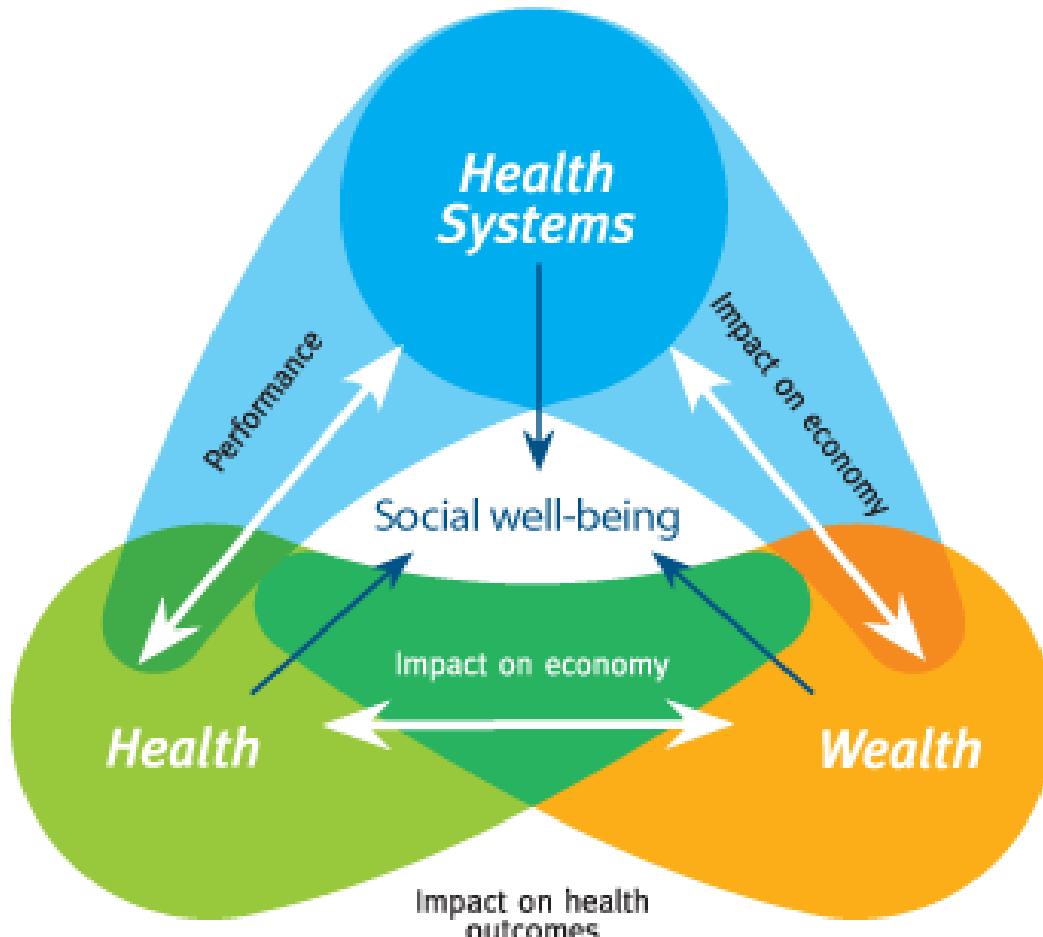
Fabrizio Carinci

fabrizio.carinci@unibo.it

Monday, 20th February 2023

Health System Performance Assessment

The WHO Tallinn Charter, 2008



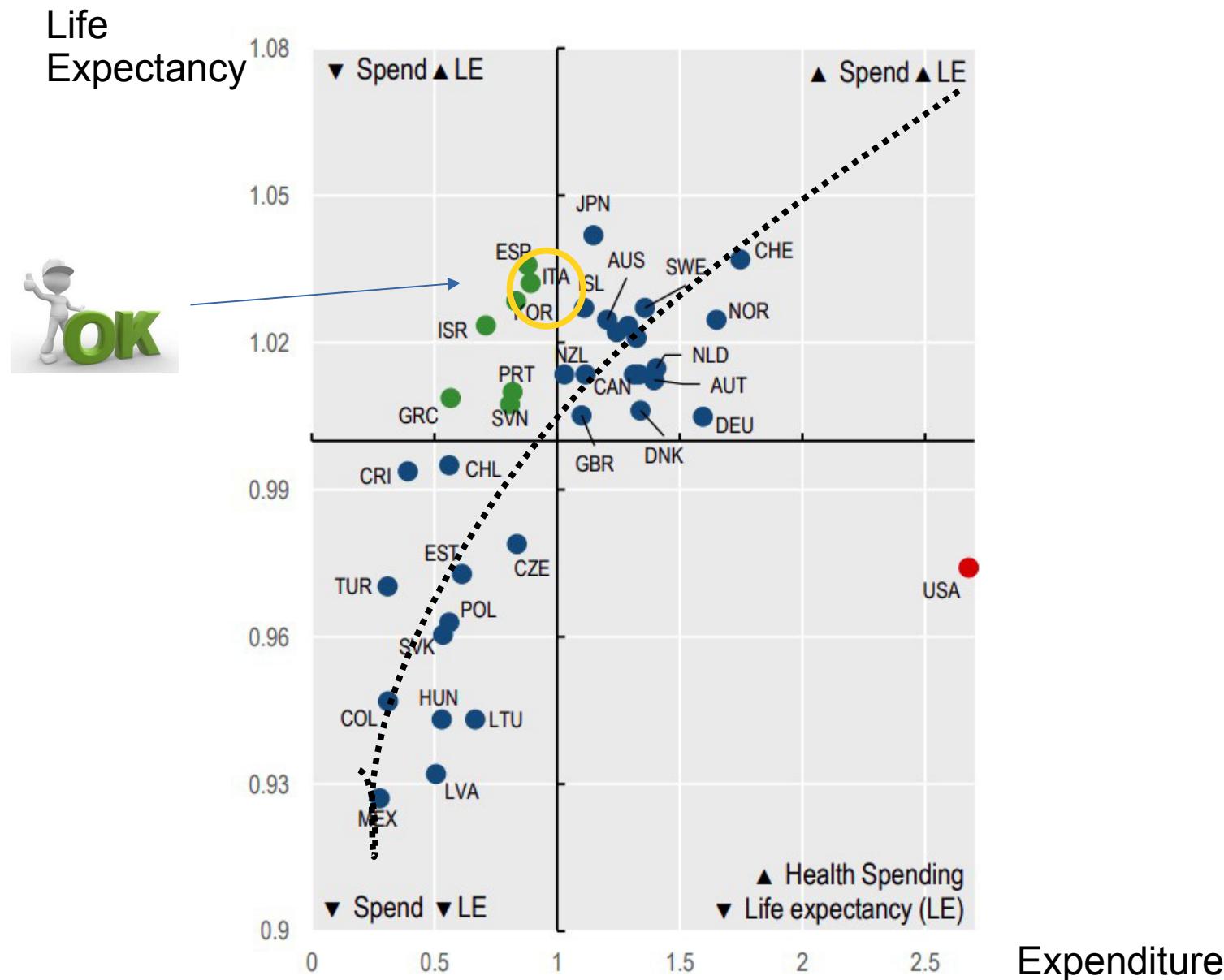
EUROPE

WHO European Ministerial Conference on Health Systems
Tallinn, Estonia 25-27 June 2008

http://www.euro.who.int/_data/assets/pdf_file/0008/88613/E91438.pdf

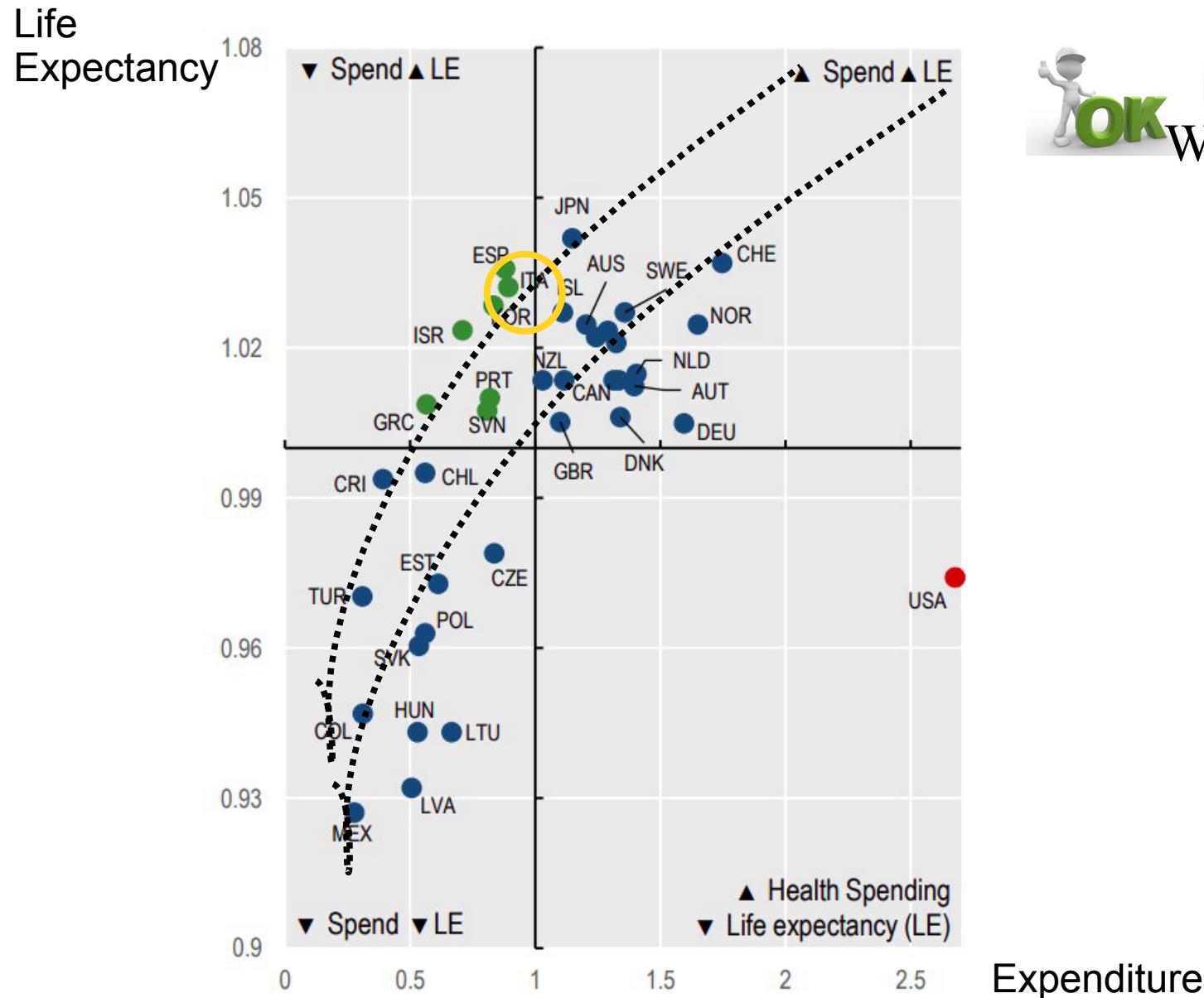
Life Expectancy vs Health Expenditure – 2019

Source: OECD Health at a Glance 2021



Life Expectancy vs Health Expenditure – what if?

Source: OECD Health at a Glance 2021

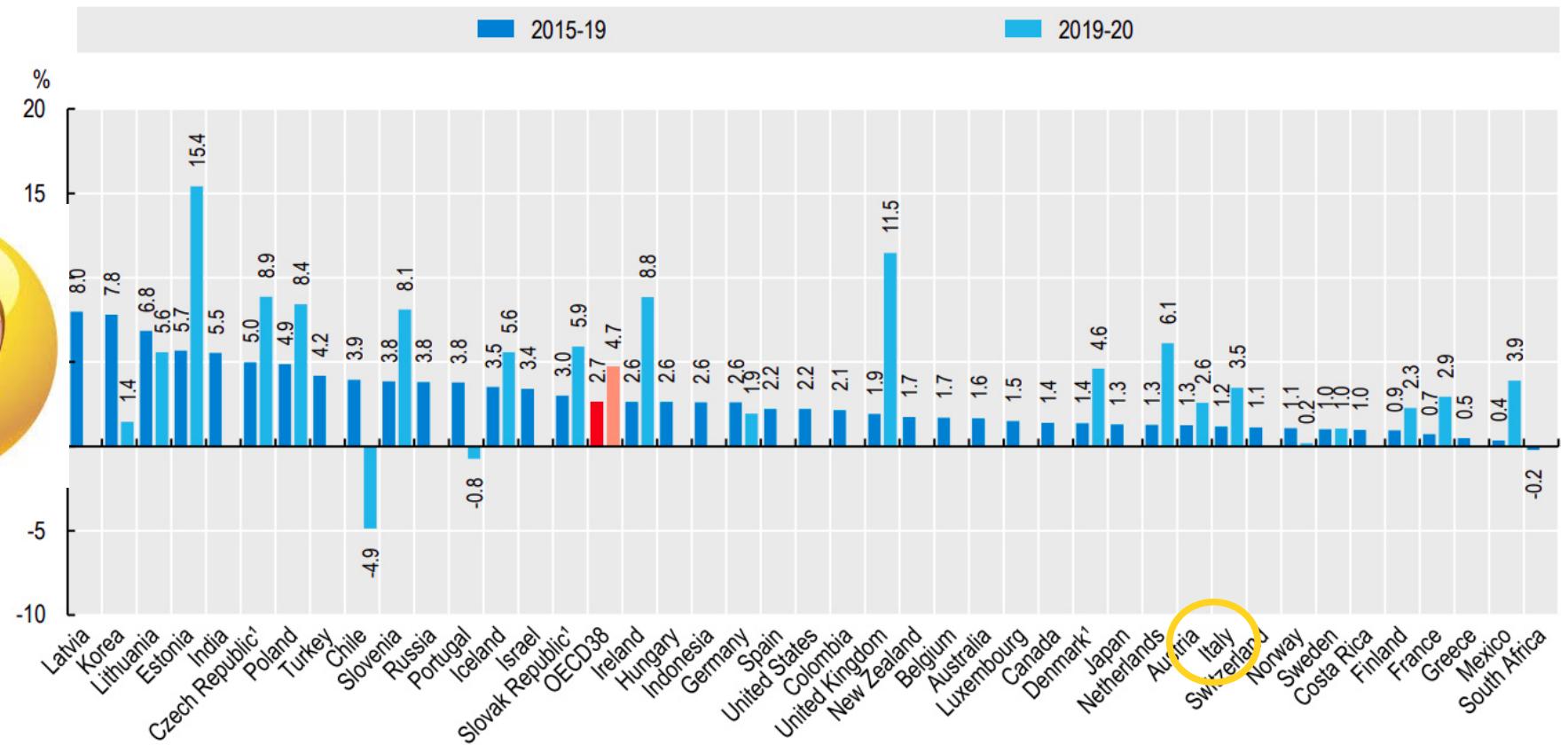


Annual average growth rate in per capita health expenditure 2008-2018

Source: OECD Health at a Glance 2019



Figure 7.5. Annual growth in per capita health expenditure (real terms), 2015-19 (or nearest year) and 2019-20



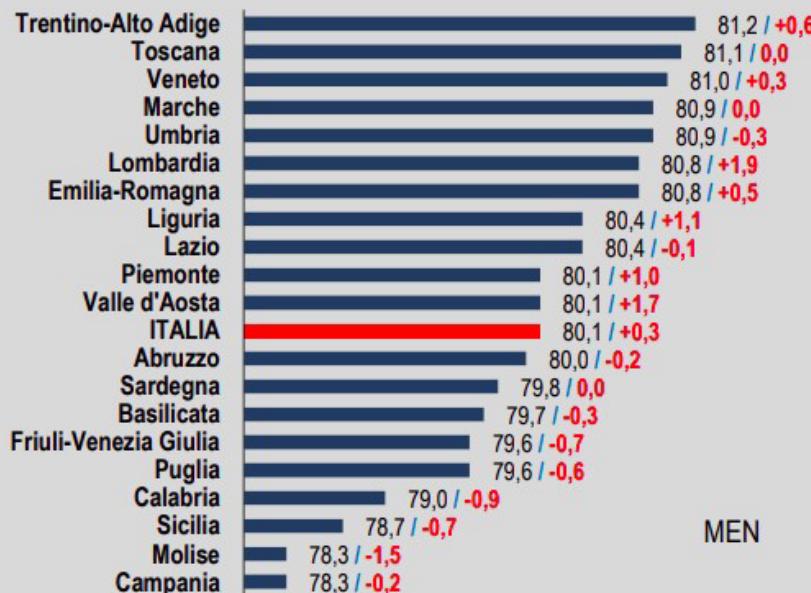
Life expectancy at birth in Italy, by Region and Sex

https://www.istat.it/it/files//2022/04/Demographic-indicators_year_2021.pdf

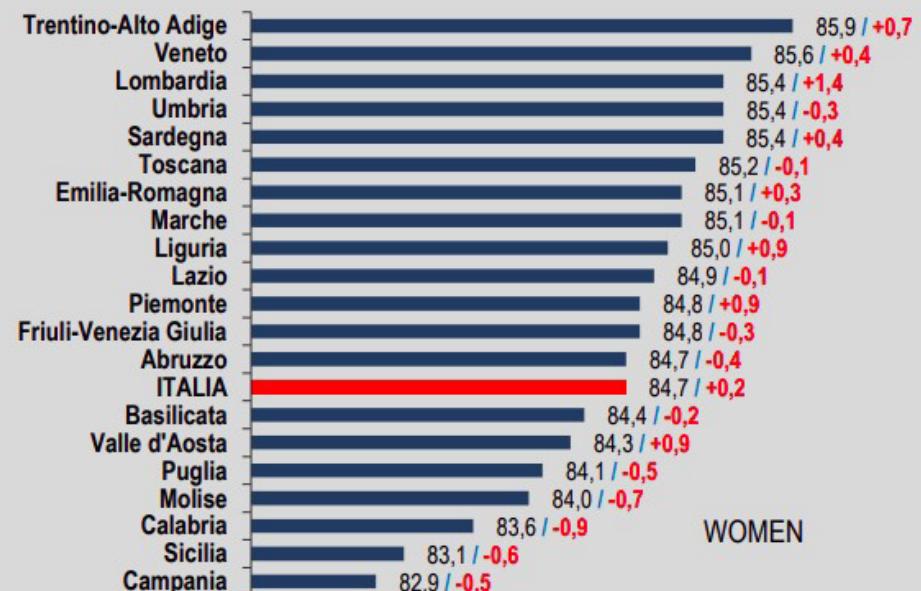


FIGURE 1. LIFE EXPECTANCY AT BIRTH BY GENDER AND REGION

Year 2021 and difference with 2020 (red figures), in years and tenths of year, estimate.



MEN

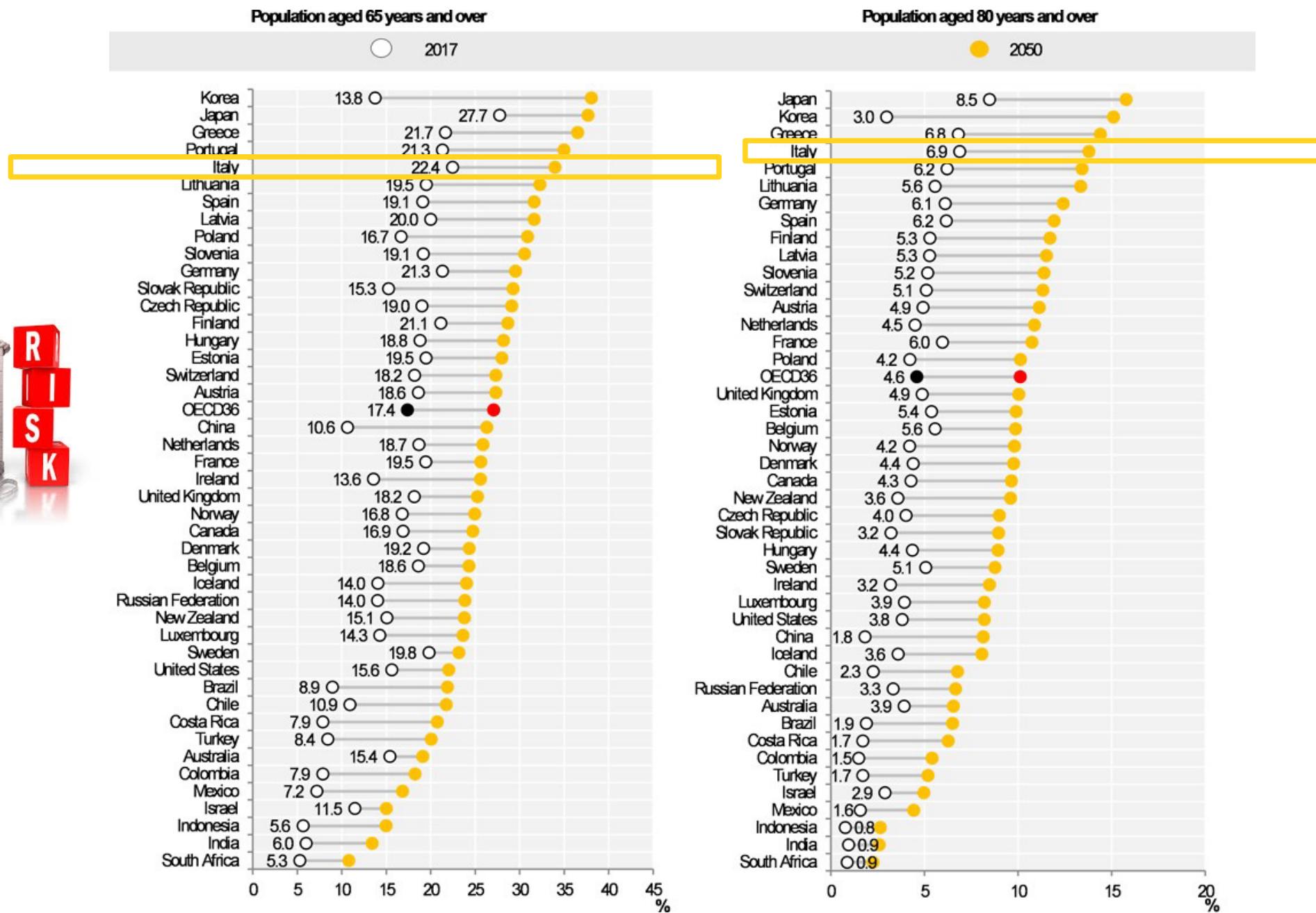


WOMEN

Source: Istat, Tavole di mortalità della popolazione residente (2020), Sistema di nowcasting per indicatori demografici (2021).

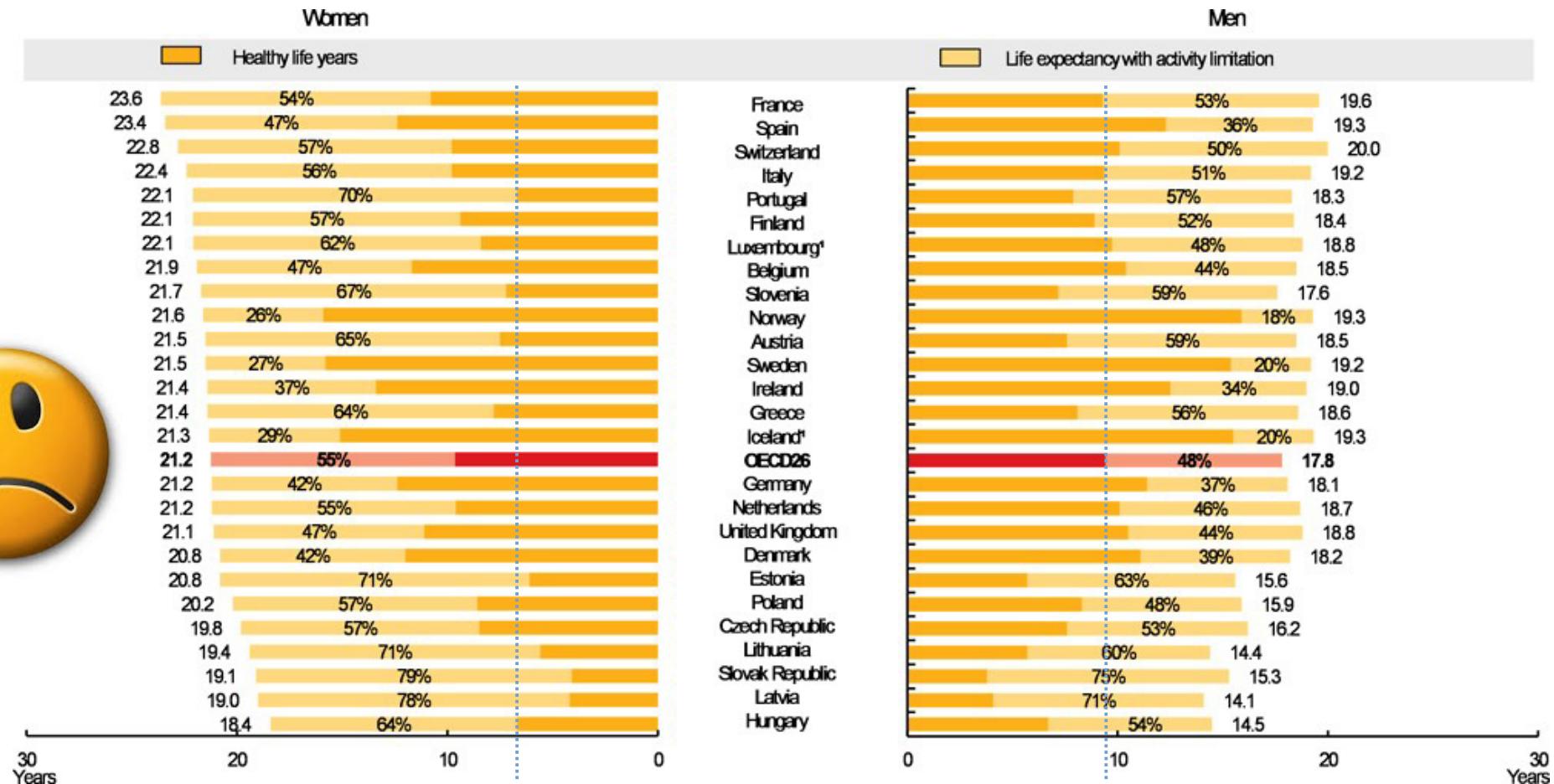
Share of the population aged over 65 and 80 years, 2017 and 2050

Source: OECD Health at a Glance 2019



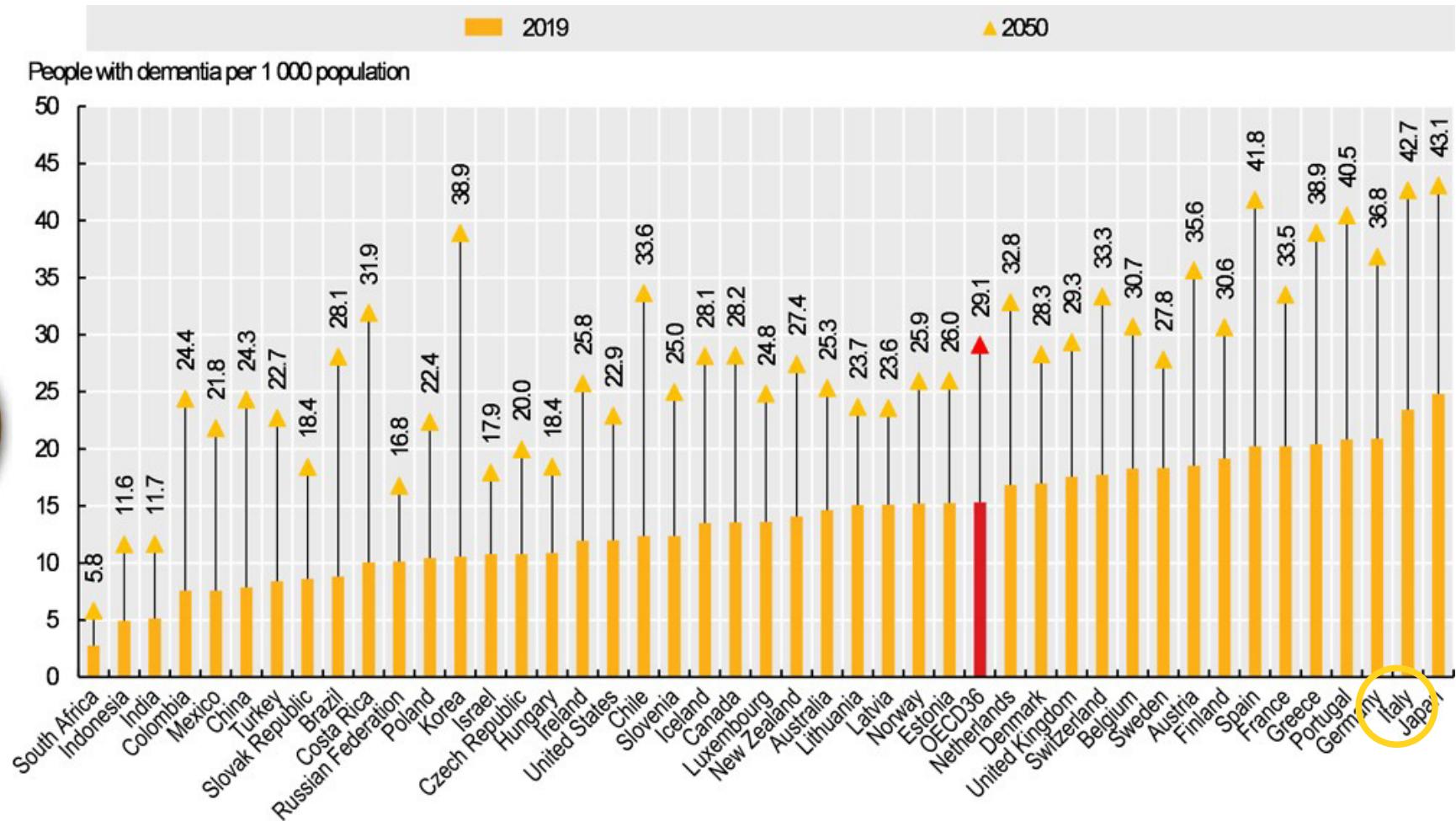
Life expectancy in good health / without limitation

Source: OECD Health at a Glance 2019



Prevalence of dementia: estimates 2050

Fonte: OECD Health at a Glance 2019



Controlling the evolution of life expectancy: how to monitor and manage future results?

- Long term estimates depend from the future demographics as a result of improved **health care systems worldwide**
- In the long run, the progress made in terms of life expectancy can create new problems. It is essential to evaluate the results and predict the impact on high risk strata of the aged population, e.g. those affected by chronic diseases
- The effect on selected strata will need to be governed, evaluating the relative of new **systems of care and technologies** (e.g. "digital health")
- Covid-19 has changed the scenario on a global level

QUESTIONS:

- To what extent the results achieved can be **attributed** to the health care system (methodological and statistical problem)?
- How should we evaluate **quality of care** in view of future challenges?

Reference textbook on health care quality

Joint publication
OECD-WHO
December 2019

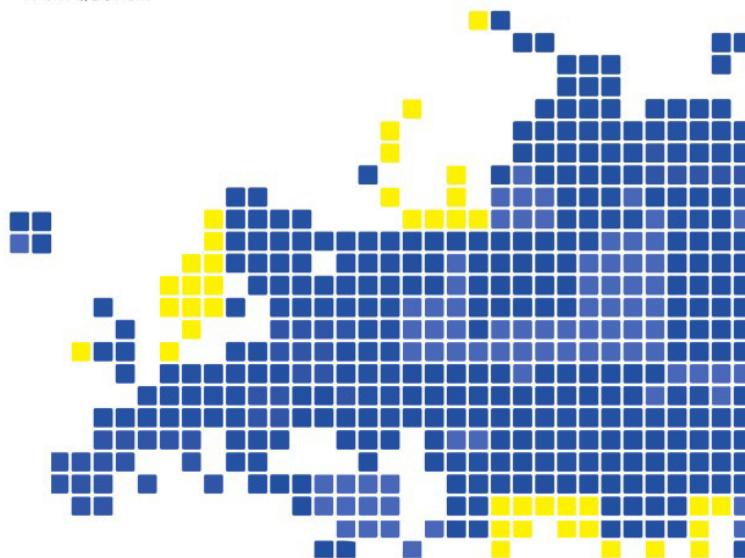
Freely available in PDF at:

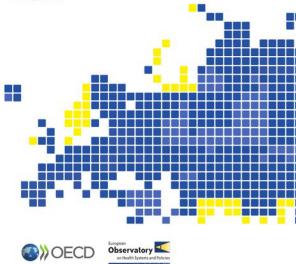
<https://apps.who.int/iris/bitstream/handle/10665/327356/9789289051750-eng.pdf>

Improving healthcare quality in Europe

Characteristics, effectiveness and implementation of different strategies

Edited by
Reinhard Busse
Niek Klazinga
Dimitra Panteli
Wilm Quentin





OECD

Health Observatory
OECD Health Statistics**Table 1.1** Selected definitions of quality, 1980–2018

Donabedian (1980) In: <i>"Explorations in quality assessment and monitoring. The definition of quality and approaches to its assessment"</i>	Quality of care is the kind of care which is expected to maximize an inclusive measure of patient welfare, after one has taken account of the balance of expected gains and losses that attend the process of care in all its parts. <i>[More generally, quality in this work is "the ability to achieve desirable objectives using legitimate means".]</i>
Institute of Medicine, IOM (1990) In: <i>"Medicare: A Strategy for Quality Assurance"</i>	Quality of care is the degree to which health services for individuals and populations increase the likelihood of desired health outcomes and are consistent with current professional knowledge.
Council of Europe (1997) In: <i>"The development and implementation of quality improvement systems (QIS) in health care. Recommendation No. R (97) 17"</i>	Quality of care is the degree to which the treatment dispensed increases the patient's chances of achieving the desired results and diminishes the chances of undesirable results, having regard to the current state of knowledge.
European Commission (2010) In: <i>"Quality of Health care: policy actions at EU level. Reflection paper for the European Council"</i>	[Good quality care is] health care that is effective, safe and responds to the needs and preference of patients. <i>The Paper also notes that "Other dimensions of quality of care, such as efficiency, access and equity, are seen as being part of a wider debate and are being addressed in other fora."</i>
WHO (2018) In: <i>"Handbook for national quality policy and strategy"</i>	Quality health services across the world should be: <ul style="list-style-type: none"> • Effective: providing evidence-based health care services to those who need them. • Safe: avoiding harm to people for whom the care is intended. • People-centred: providing care that responds to individual preferences, needs and values. <p>In order to realize the benefits of quality health care, health services must be timely [...], equitable [...], integrated [...], and efficient [...]</p>

Coherent with
OECD 2015

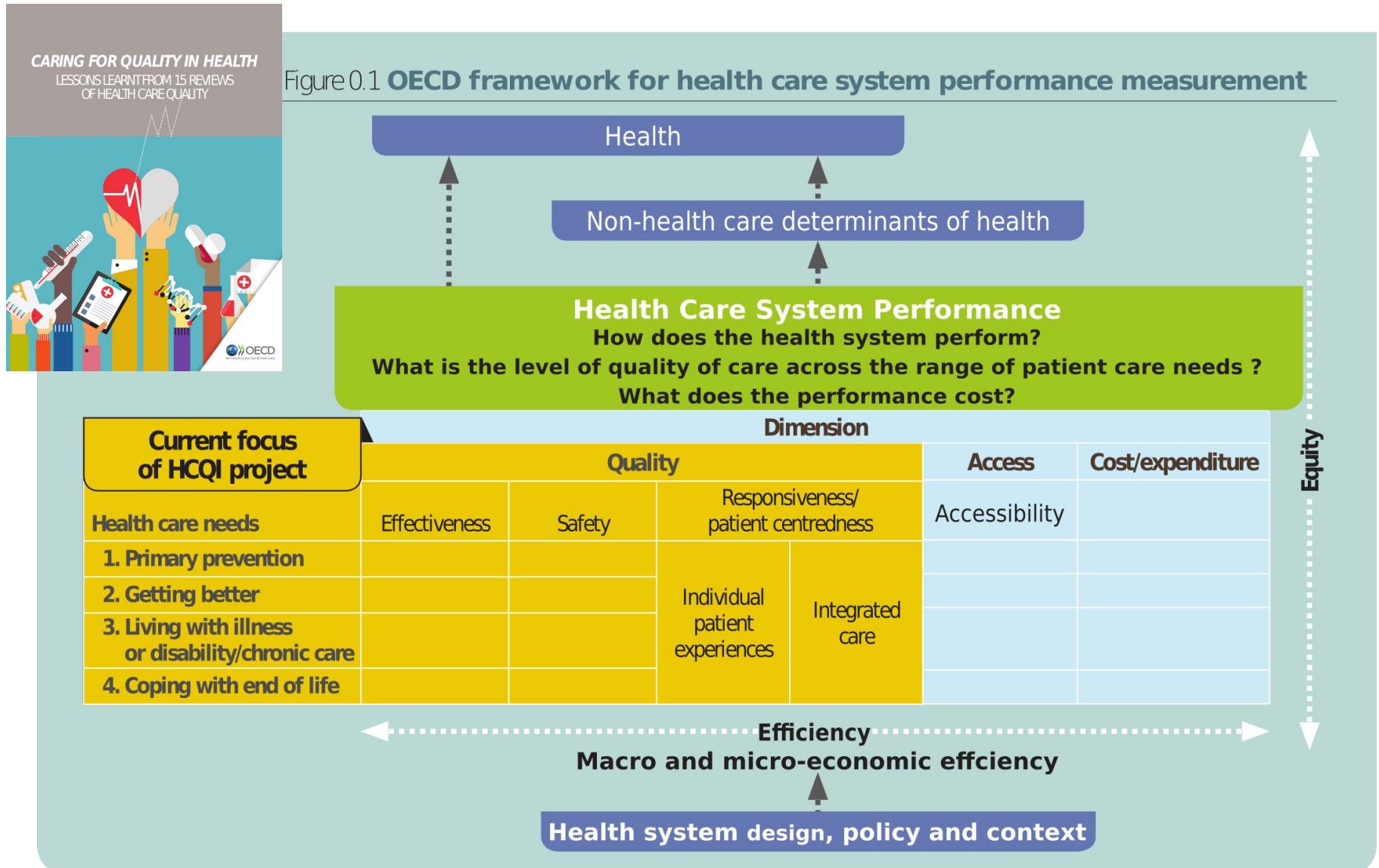
Dimensions of quality according to 10 definitions

Table 1.2 Quality dimensions in ten selected definitions of quality, 1980–2018

	Donabedian (1980)	IOM (1990)	Council of Europe (1997)	IOM (2001)	OECD (2006)	WHO (2006b)	EC (2010)	EC (2014)	WHO (2016)	WHO (2018)
Core dimensions of healthcare quality	Effectiveness	X	X	X	X	X	X	X	X	X
	Safety		X	X	X	X	X	X	X	X
	Responsiveness		X	Patient-centredness	X	Patient-centredness	X	Patient-centredness	Patient-centredness	Patient-centredness
	Acceptability					X				
	Appropriateness		X					X		
	Continuity									
	Timeliness			X					X	X
	Satisfaction	X	X							
	Health improvement	X	X							
	Other	Patient Welfare		Assessment of care process		Patient's preferences		Integration		Integration
Other dimensions of health systems performance	Efficiency		X	X		X	X	X	X	X
	Access		X			X				
	Equity			X		X	X	X	X	X

How to measure healthcare quality

<https://www.oecd.org/els/health-systems/Caring-for-Quality-in-Health-Final-report.pdf>



Source: Carinci, F. et al. (2015), "Towards Actionable International Comparisons of Health System Performance: Expert Revision of the OECD Framework and Quality Indicators", *International Journal for Quality in Health Care*, Vol. 27, No. 2, pp. 137-146, <http://dx.doi.org/10.1093/intqhc/mzv004>.

Performance indicators

International Journal for Quality in Health Care Advance Access published March 10, 201



International Journal for Quality in Health Care, 2015, 1–10

doi: 10.1093/intqhc/mzv004

Article

Article

Towards actionable international comparisons of health system performance: expert revision of the OECD framework and quality indicators

**F. CARINCI^{1,2}, K. VAN GOOL^{3,4}, J. MAINZ⁵, J. VEILLARD⁶, E. C. PICHORA⁶,
J. M. JANUEL⁷, I. ARISPE⁸, S. M. KIM⁹, and N.S. KLAZINGA³, ON BEHALF
OF THE OECD HEALTH CARE QUALITY INDICATORS EXPERT GROUP***

Source: Carinci, F. et al. (2015), "Towards Actionable International Comparisons of Health System Performance: Expert Revision of the OECD Framework and Quality Indicators", *International Journal for Quality in Health Care*, Vol. 27, No. 2, pp. 137-146, <http://dx.doi.org/10.1093/intqhc/mzv004>.

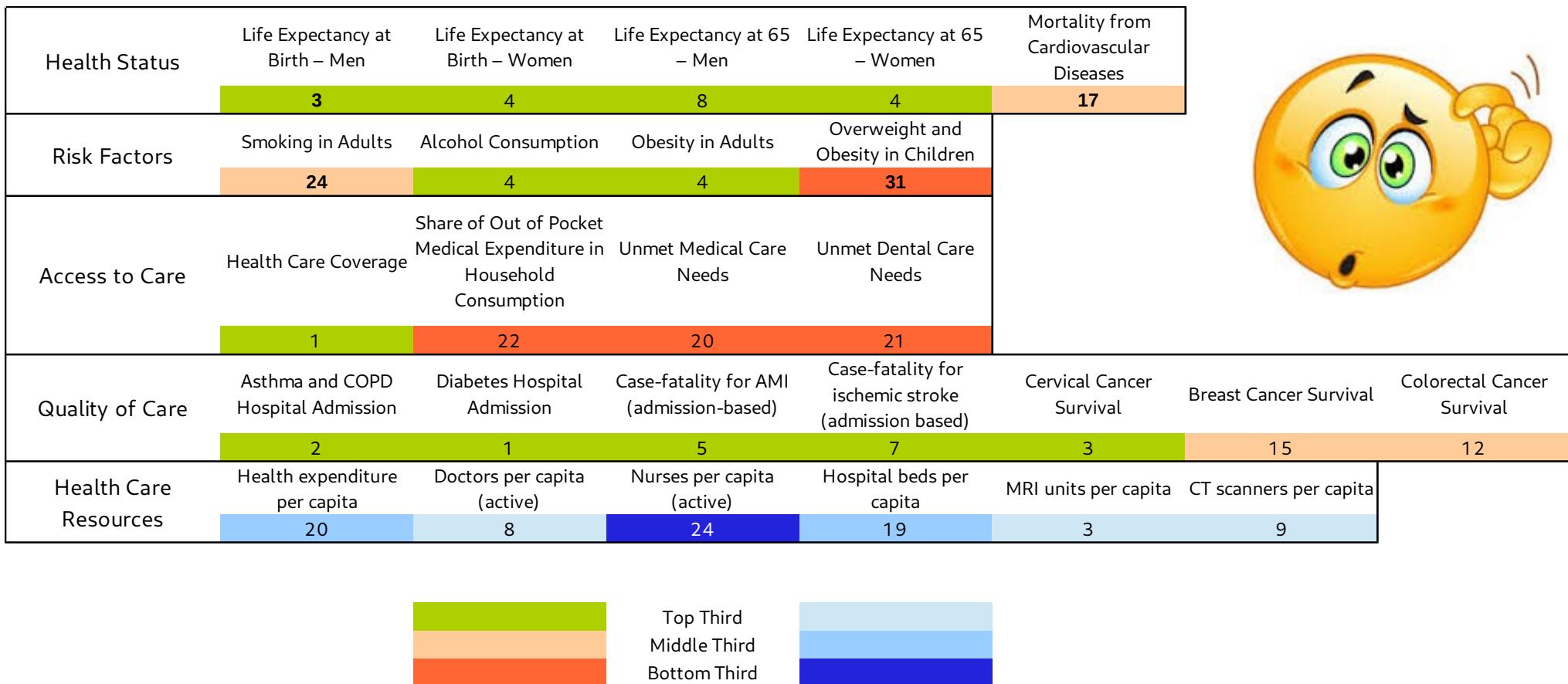
Criteria for selecting performance indicators

Table 1 Criteria used to score HCQI

Criterion	Definition
Validity	Sufficient scientific evidence exists to support a link between the value of an indicator and one or more aspects of health care quality
Reliability	Repeated measurements of a stable phenomenon get similar results
Relevance	An indicator measures an aspect of quality with high clinical importance, a high burden of disease or high health care use
Actionability	An indicator measures an aspect of quality that is subject to control by providers and/or the health care system and is actually used at a national level for policy making, monitoring or strategy development
International feasibility	An indicator can be derived for international comparisons without substantial additional resources
International comparability	Reporting countries comply with the relevant data definition and where differences in the indicator values between countries reflect issues in quality of care rather than differences in data collection methodologies, coding or other non-quality of care reasons

Performance Dashboards

Source: OECD Health at a Glance 2015



How does Italy compare?

Source: OECD Health at a Glance 2021

<https://www.oecd.org/italy/health-at-a-glance-Italy-EN.pdf>

● Italy ● Highest performer
● OECD ● Lowest performer

Health status is good in Italy, which has one of the oldest populations across OECD countries

Life expectancy (2019 or nearest year)

Years of life at birth



Avoidable mortality (2019 or nearest year)

Deaths per 100 000 population (age-standardised)



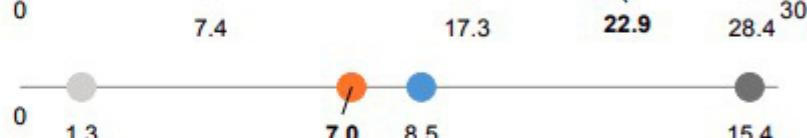
Population ageing (2019 or nearest year)

Share of population 65 or older



Self-rated health (2019 or nearest year)

Population in poor health (% population 15+)



Risk factors for health are mixed, with higher-than-average smoking rates but lower alcohol consumption and overweight/obesity than the OECD average

Smoking (2019 or nearest year)

Daily smokers (% population 15+)



Alcohol (2019 or nearest year)

Litres consumed per capita (population 15+)



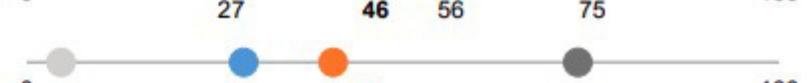
Overweight/obese (2019 or nearest year)

Population with BMI>=25 (% population 15+)



Air pollution (2019 or nearest year)

Deaths due to ambient particulate matter pollution (per 100 000 population)



How does Italy compare?

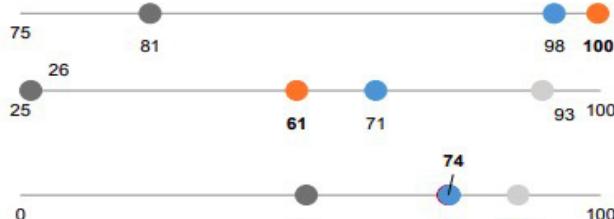
Source: OECD Health at a Glance 2021

<https://www.oecd.org/italy/health-at-a-glance-Italy-EN.pdf>



Population coverage is high, though satisfaction with quality of care is below the OECD average

Population coverage, eligibility (2019 or nearest year)
Population eligible for core services (% population)



Population coverage, satisfaction (2019 or nearest year)
Population satisfied with availability of quality health care (% population)

Financial protection (2019 or nearest year)
Expenditure covered by compulsory prepayment (% total expenditure)



Many indicators of quality care are good, and primary care has helped keep avoidable hospital admissions low

Safe primary care (2019 or nearest year)
Antibiotics prescribed (defined daily dose per 1 000 people)



Effective primary care (2019 or nearest year)
Avoidable COPD admissions (per 100 000 people, age-sex standardised)



Effective preventive care (2019 or nearest year)
Mammography screening within the past two years (% of women 50+)

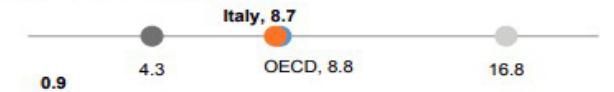


Effective secondary care (2019 or nearest year)
30 day mortality following AMI (per 100 000 people, age-sex standardised)



Many indicators of health resources are at or somewhat below the OECD average

Health spending (2019 or nearest year)
Per capita (USD based on PPPs)



Long-term care spending (2019 or nearest year)
% GDP



Doctors (2019 or nearest year)
Practicing physicians (per 1 000 population)



Nurses (2019 or nearest year)
Practicing nurses (per 1 000 population)



Hospital beds (2019 or nearest year)
Per 1 000 population

Covid dashboard



How do you judge it?

- Better than OECD average
- Close to OECD average
- Worse than OECD average

	Excess deaths		COVID-19 deaths		COVID-19 cases		Vaccination rates	
	Per 1 million population		Per 1 million population		Per 100 000 population		Share of population fully vaccinated	
OECD	1 499		1 285		8 392		60.0	
Australia	211	☒	36	☒	437	☒	45.6	☒
Austria	1 270	●	1 180	●	8 368	●	60.1	●
Belgium	1 374	●	2 186	☒	10 867	●	72.6	●
Canada	1 125	●	699	●	4 347	●	71.2	●
Chile	2 138	●	1 739	●	8 669	●	73.7	☒
Colombia	2 323	●	2 151	☒	9 754	●	33.6	☒
Costa Rica			928	●	10 560	●	42.6	☒
Czech Republic	3 465	☒	2 838	☒	15 842	☒	55.7	●
Denmark	195	☒	436	☒	6 190	●	75.3	☒
Estonia	1 396		956	●	11 956	●	53.5	●
Finland	343	☒	176	☒	2 572	☒	63.4	●
France	1 374	●	1 652	●	10 438	●	66.1	●
Germany	925	●	1 095	●	5 117	●	64.2	●
Greece	1 402	●	1 188	●	6 170	●	59.4	●
Hungary	2 424	●	3 070	☒	8 443	●	58.7	●
Iceland	188	☒	82	☒	3 284	☒	80.5	☒
Ireland			1 007	●	7 929	●	74.2	☒
Israel	766	●	743	●	14 925	☒	64.4	●
Italy	2 151	●	2 140	☒	7 850	●	68.3	●
Japan	787	●	117	☒	1 347	☒	61.2	●
Korea	52	☒	40	☒	624	☒	52.7	●
Latvia	1 209	●	1 325	●	8 473	●	46.4	☒
Lithuania	1 928	●	1 573	●	12 171	●	60.3	●
Luxembourg	879	●	1 306	●	12 510	●	62.9	●
Mexico	4 456	☒	1 812	●	2 857	☒	35.4	☒
Netherlands	1 384	●	1 020	●	11 535	●	67.6	●
New Zealand	214	☒	5	☒	91	☒	41.5	☒
Norway	-277	☒	148	☒	3 550	☒	67.0	●
Poland	3 663	☒	1 978	●	7 670	●	51.7	●
Portugal	2 025		1 663	●	10 405	●	85.2	☒
Slovak Republic	3 133	☒	2 293	☒	14 828	☒	41.4	☒
Slovenia	2 320	●	2 268	☒	14 174	☒	48.3	●
Spain	1 841	●	1 710	●	10 490	●	78.6	☒
Sweden	545	●	1 420	●	11 177	●	64.2	●
Switzerland	1 069	●	1 197	●	9 810	●	58.4	●
Turkey			600	●	8 672	●	52.9	●
United Kingdom	1 599	●	2 232	☒	11 608	●	66.0	●
United States	2 559	●	1 824	●	13 197	☒	55.2	●

Programma Nazionale Esiti

<https://pne.agenas.it>

The screenshot shows the homepage of the PNE 2021 website. At the top, there is a blue header bar with the Agenzia Nazionale per i Servizi Sanitari Regionali logo and the Ministero della Salute logo. Below the header, the title "PNE²⁰₂₁ Programma Nazionale Esiti - edizione 2021" is displayed. The main content area features a large image of a hand holding a magnifying glass over a map of Italy, with the text "PNE è uno strumento di valutazione a supporto di programmi di audit clinico e organizzativo" overlaid. Below this, a quote reads: "PNE non produce classifiche, graduatorie, giudizi." There are two buttons: "Novità Edizione 2021 →" and "Report PNE 2021" with a file icon. The page is divided into three main sections: "Ospedale" (Hospital), "Territorio" (Territory), and "Equità" (Equity). Each section contains an illustration and a brief description.

PNE è uno strumento di valutazione a supporto di programmi di audit clinico e organizzativo

"PNE non produce classifiche, graduatorie, giudizi."

Novità Edizione 2021 → Report PNE 2021

Ospedale
Indicatori per ambito nosologico/struttura, flussi e treemap

Territorio
Tassi di accesso in P.S., ospedalizzazioni evitabili ed esiti territoriali

Equità
Risultati stratificati per genere / cittadinanza e dettaglio del titolo di studio

PNE Results

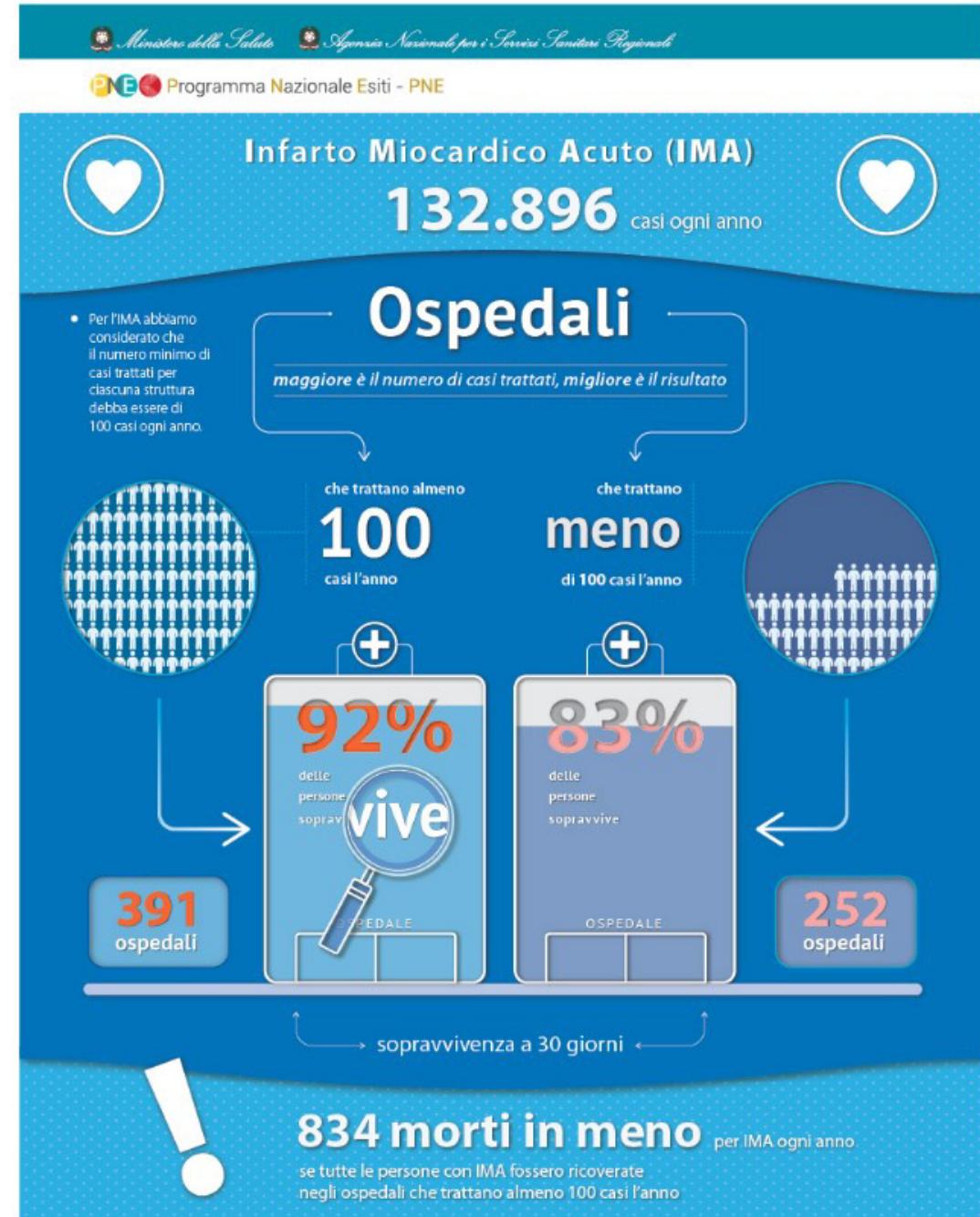
<https://pne.agenas.it>

The relation between hospital volume, performance and outcomes measured in terms of mortality rates

Among cases of Acute Myocardial Infarction (AMI), those admitted to hospitals that treat less cases (lower volumes) have a higher probability of dying

The EXCESS MORTALITY calculated for hospitals with lower volumes is based on a multivariate predictive model

The SOCIETAL IMPACT of such a simple statistical information is very HIGH



Fonte: PNE edizione 2016



Performance benchmarking

About CIHI [Home](#) > [Health System Performance](#) > [Indicators](#) > [International](#) > International Comparisons - Indicator Results

International Comparisons: A Focus on Quality of Care

[Indicator Results](#) [Methodology](#) [Peer Countries](#)

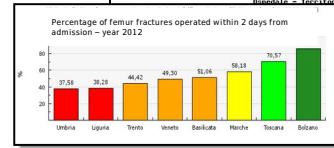
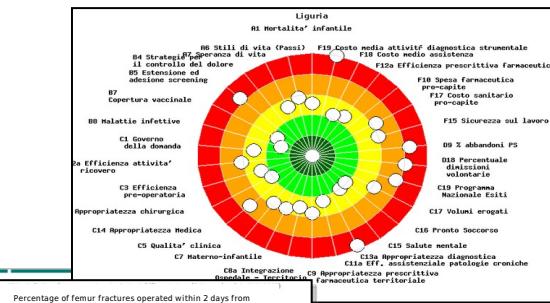


Figure A: How Provinces Compare With OECD Countries on Quality of Care Indicators

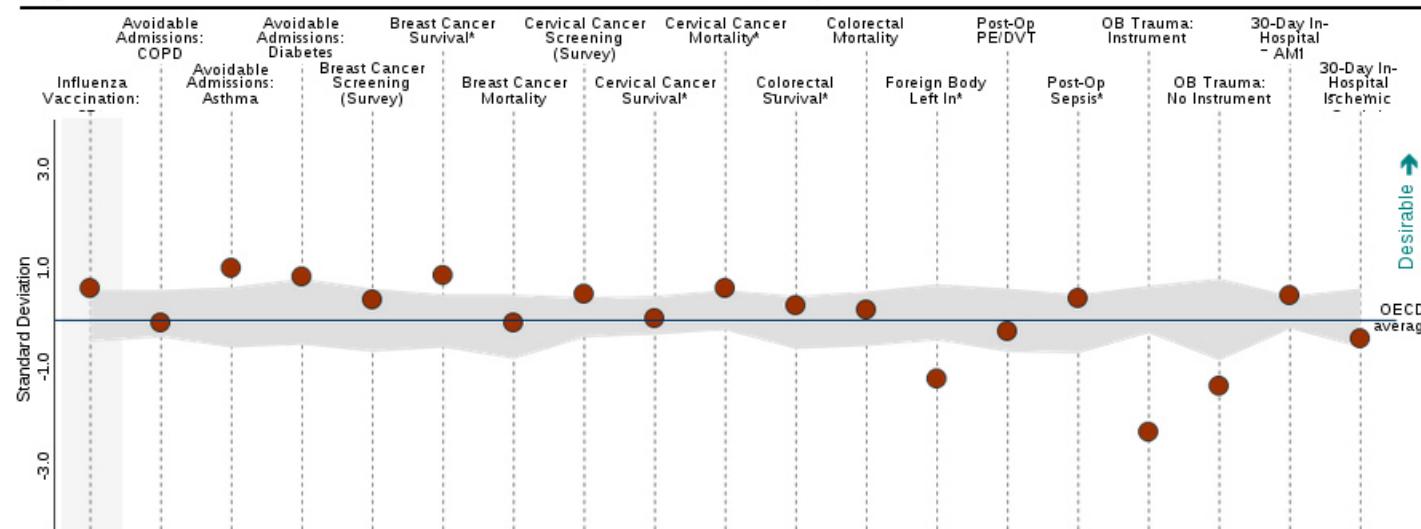
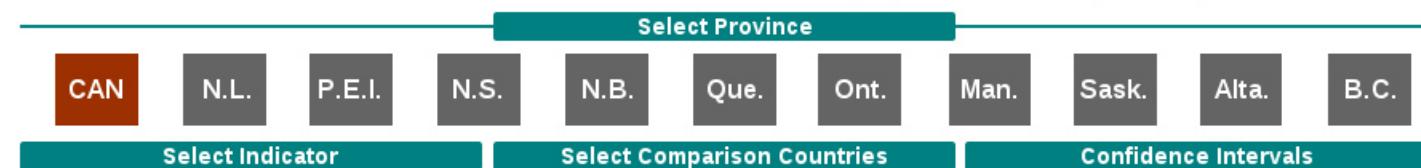


Figure A presents Canadian and provincial results relative to the OECD average, measured in standard deviations.

* Not available for all provinces.

The area between the top and bottom quarters of OECD countries.

Actionable
Public
Reporting



Matrix Dimensions: Quality

Effectiveness

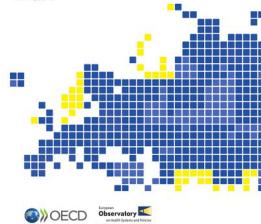
- Achieving desirable outcomes, given the correct provision of evidence-based health care services to all who could benefit

Safety

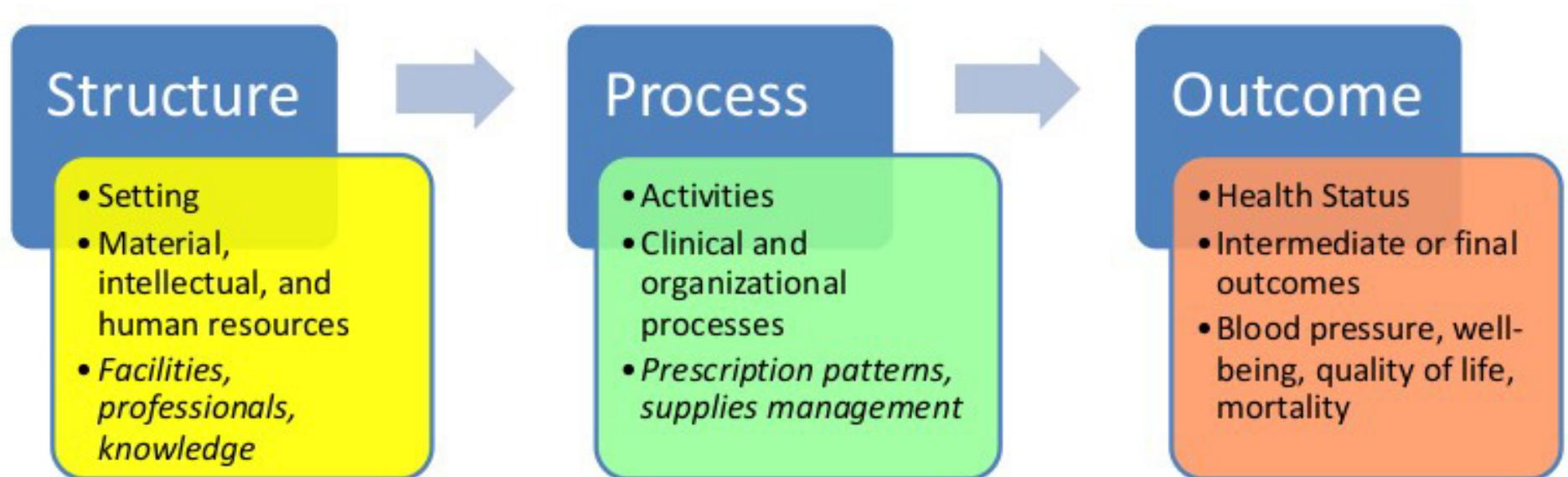
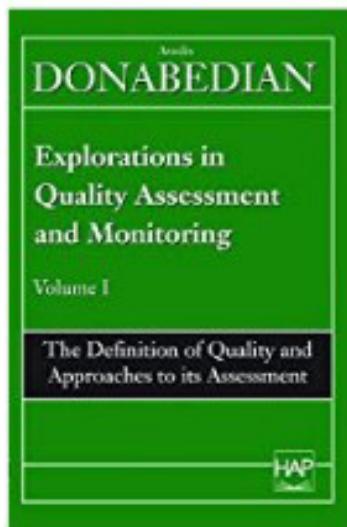
- System has the right structures, renders services and attains results in ways that prevent harm to the user, provider, or environment

Responsiveness/ Patient centeredness

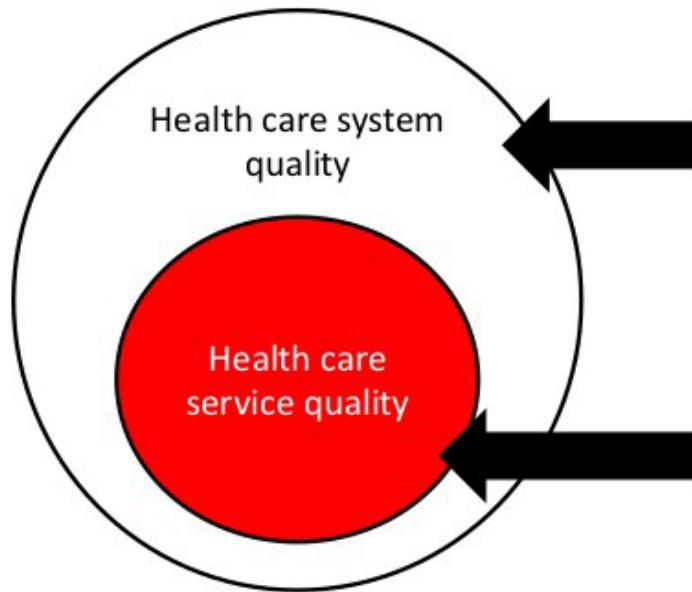
- System actually functions by placing the patient/user at the center of its delivery of health care



Donabedian's Triad



Levels of quality of care



Health Systems Performance

Achieving the global objectives in terms of health improvement, response to health needs, economic efficiency

Health Care Quality

"the degree to which health services for individuals and populations are effective, safe, and people-centred"

An evolving view of outcomes

Rationale, examples of measures and data sources

From Deaths	<ul style="list-style-type: none">Mortality and life-expectancy: classical parameters to measure health systems outcomesLook at outcome from a public health perspectiveNeed good death registries as an information source
To Diseases	<ul style="list-style-type: none">Prevalence and incidence of diseases are classical parameters to assess morbidity of diseases in a countryRelated outcome measures try to capture the reduction in morbidity and the outcomes of specific diseases (e.g. QALYs, SF36)Medical/clinical perspective is the dominant way of operationalizing outcome measures. Outcome measurement is dependent on clinical registries (such as on cancer and diabetes).Linking to costs (value) at system level (burden of diseases studies) and for specific services and interventions (cost-effectiveness studies)
To Disability	<ul style="list-style-type: none">Many chronic diseases come with long term disabilities and outcomes should also address the way a health system deals with disabilitiesAt system level DALY (Disability Adjusted Life Expectancy) most well-known measure; at health services level various instruments available to assess disabilities and their outcomes (e.g. inter RAI initiative)Administrative data-bases and surveys are the main data source
To Discomfort	<ul style="list-style-type: none">Increasingly outcomes experienced by citizens/patients seen as an important outcomePROMS (patient reported outcomes) mainly tested for clinical procedures and treatments and still under development for chronic conditions ; EQ5D a more generic measure used.PREMs with some limited international validation of instruments (CAHPS, Picker)

Properties of performance indicators

Indicators must be „SMART“ !

Specific with respect to the objective that must be measured

Measureable, quantitative and/or qualitative

Accessible, meaning that data can be obtained at a sustainable cost

Relevant with regards to the target information needed

Time-defined, with a clear indication of the relevant timeframe

Key messages

- Life expectancy is a key indicator to understand the results of health care systems internationally. However, it needs to be monitored very closely, by looking at various strata of the general population. The effect of Covid-19 will need to be closely investigated.
- The definition of quality of care has been evolving quite rapidly, with a possible convergence of the notion across institutions and jurisdictions
- Performance measurement includes multiple dimensions of quality of care that needs to be tackled using performance indicators that possess specific criteria e.g. actionability and being „SMART“

Materials

- *Course notes*
- OECD Health at a Glance 2019, p.1-40
- OECD-WHO, Improving Health Care Quality in Europe, p.3-30



ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Second cycle degree programme (LM) in Statistical Sciences

85278 – Methods and tools for Health Statistics

**85277 – Methods and tools for Official Statistics: Population
and Health Statistics**

a.a. 2022/2023

Stream 1. Regional, national and international health statistics

Topic 1.2.3

OECD Health Care Quality Indicators Project

Fabrizio Carinci

fabrizio.carinci@unibo.it

Monday, 20th February 2023

OECD Health Care Quality Indicators Project

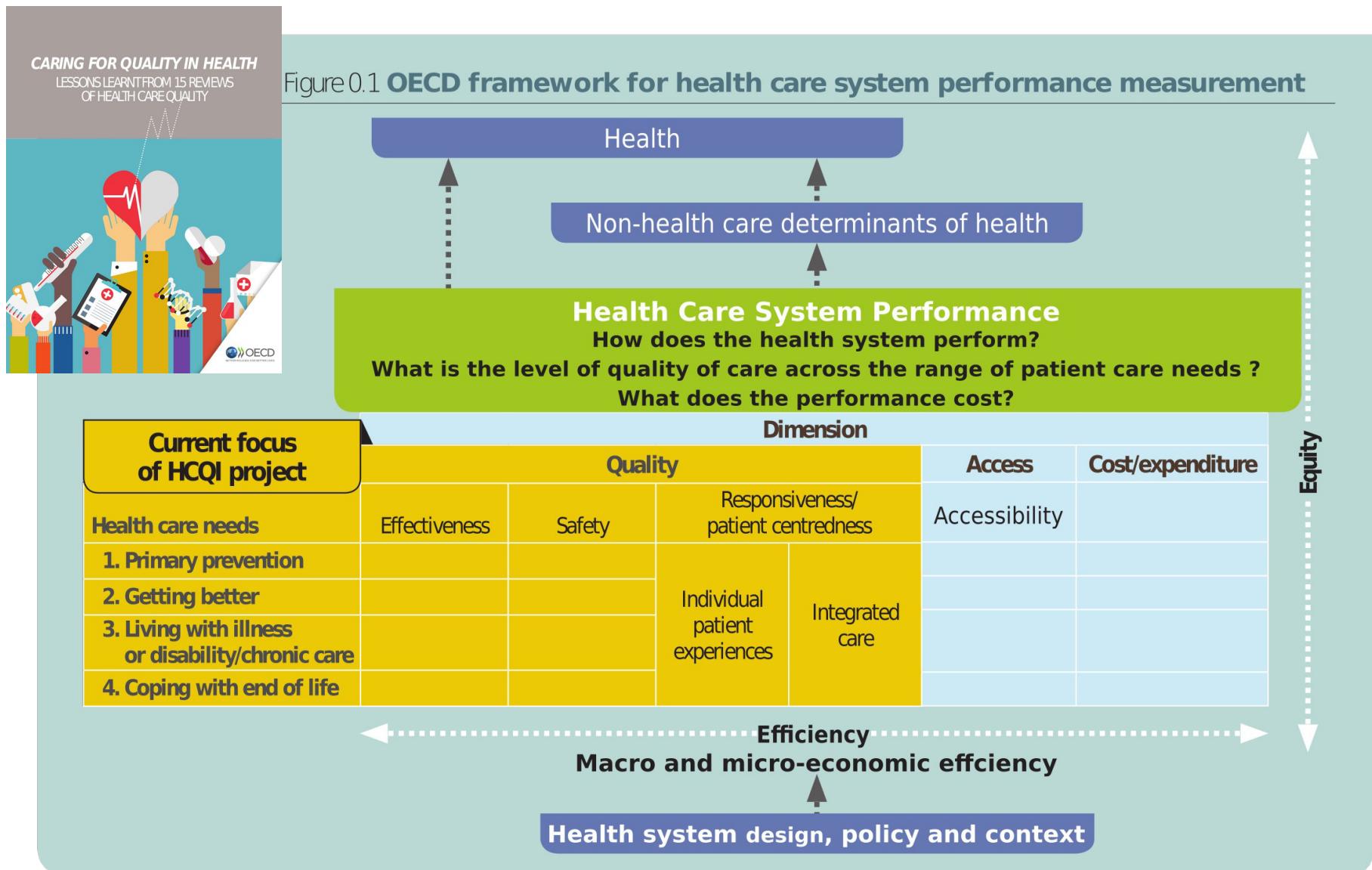
- Under the umbrella of the Organisation for Economic Cooperation and Development (OECD), the ‘Health Care Quality Indicators’ (HCQI) Project was initiated in 2001
- The general objective was to help Member States (MS) identify **priority areas for quality improvement** to provide **achievable standards** by examining results among best performing countries
- In 2006, the OECD released a **common conceptual framework for health system performance**. Nested “quality matrix” with **vertical dimensions** of ‘effectiveness’, ‘patient safety’ and ‘responsiveness/patient-centeredness’, **horizontally subdivided according to levels of health care needs over the life cycle**: ‘staying healthy’ for healthy subjects, ‘getting better’ for people affected by a disease, ‘living with illness or disability’ for those with a chronic condition and ‘coping with end of life’ for terminal patients.

HCQI Outputs

- Since 2007, results of the HCQI project have routinely contributed to international comparisons through the publication of the OECD series ‘Health at a Glance’ and the release of OECD Health Statistics alongside other international health data on expenditure, resources, utilization and outcomes.
- In 2013, the OECD HCQI data collection process included a total of **70 indicators** covering the following ‘themes’: Primary Care (PC); Acute Care (AC); Mental Health (MH); Cancer Care (CC); Patient Safety (PS) and Patient Experiences (PEs). The collection reports data from 34 countries, including non-OECD member countries eg Singapore and Latvia.

OECD Health Systems Performance Assessment Framework 2015

<https://www.oecd.org/els/health-systems/Caring-for-Quality-in-Health-Final-report.pdf>



Source: Carinci, F. et al. (2015), "Towards Actionable International Comparisons of Health System Performance: Expert Revision of the OECD Framework and Quality Indicators", *International Journal for Quality in Health Care*, Vol. 27, No. 2, pp. 137-146, <http://dx.doi.org/10.1093/intqhc/mzv004>.

Performance indicators

International Journal for Quality in Health Care Advance Access published March 10, 201



International Journal for Quality in Health Care, 2015, 1–10

doi: 10.1093/intqhc/mzv004

Article

Article

Towards actionable international comparisons of health system performance: expert revision of the OECD framework and quality indicators

**F. CARINCI^{1,2}, K. VAN GOOL^{3,4}, J. MAINZ⁵, J. VEILLARD⁶, E. C. PICHORA⁶,
J. M. JANUEL⁷, I. ARISPE⁸, S. M. KIM⁹, and N.S. KLAZINGA³, ON BEHALF
OF THE OECD HEALTH CARE QUALITY INDICATORS EXPERT GROUP***

Source: Carinci, F. et al. (2015), "Towards Actionable International Comparisons of Health System Performance: Expert Revision of the OECD Framework and Quality Indicators", *International Journal for Quality in Health Care*, Vol. 27, No. 2, pp. 137-146, <http://dx.doi.org/10.1093/intqhc/mzv004>.

Strengths of quality indicators

	Structure indicators	Process indicators	Outcome indicators
STRENGTHS	Easily available. Many structural factors are evident and easily reportable Stable. Structural factors are relatively stable and often easy to observe	Easily available. Utilization of health technologies is often easily measured Easily interpreted. Compliance with process indicators can often be interpreted as good quality without the need for case-mix adjustment or inter-unit comparisons Attribution. Processes are directly dependent on actions of providers Smaller sample size needed. Significant quality deficiencies can be detected more easily	Focus. Directs attention towards the patient and helps nurture a “whole system” perspective Goals. Represent the goals of care more clearly Meaningful. More meaningful to patients and policy-makers Innovation. Encourages providers to experiment with new modes of delivery Far-sighted. Encourages providers to adopt long-term strategies (for example, health promotion) that may realize long-term benefits Resistant to manipulation. Less open to manipulation but providers may engage in risk-selection or upcoding to influence risk-adjustment

Weaknesses of quality indicators

	Structure indicators	Process indicators	Outcome indicators
WEAKNESSES	Link to quality is very weak. Can only indicate potential capacity for providing quality care Subject to response bias. Over-reporting of resources or idealizing organizational aspects (for example, having a quality management system in place)	Salience. Processes of care may have little meaning to patients unless the link to outcomes can be explained Specificity. Processes indicators are highly specific to single diseases or procedures and numerous indicators may be required to represent quality of care provided Ossification. May stifle innovation and the development of new modes of care Obsolescence. Usefulness may dissipate as technology and modes of care change Adverse behaviour. Can be manipulated relatively easily and may give rise to gaming and other adverse behaviours	Measurement definition. Relatively easy to measure some outcome aspects validly and reliably (for example, death) but others are notoriously difficult (for example, wound infection) Attribution. May be influenced by many factors outside the control of a healthcare organization Sample size. Requires large sample size to detect a statistically significant effect Timing. May take a long time to observe Interpretation. Difficult to interpret if the processes that produced them are complex or occurred distant from the observed outcome Ambiguity. Good outcomes can often be achieved despite poor processes of care (and vice versa)

Structure of an indicator

Indicator = Numerator/Denominator (per 100; 1,000; 100,000 etc)

- Numerator: **Outcome**

Most important, as this is targeting the phenomenon under study. It allows answering our main questions related to the results delivered by a hospital (provider), local health care authority (organization), geographical area (region), or an entire national health system. Can be used to compare across units.

- Denominator: **Reference population**

This defines the target population, which can be the general population or those who have experienced an "index" condition, e.g. admission with acute myocardial infarction

- Inclusion/Exclusion criteria: *Used to target the indicator to a selected population. Can avoid bias introduced by heterogeneous strata, e.g. outcomes among subjects with very complex conditions (too young or too old, oncology, traumas, poisoning, etc)*

Structure of an indicator: Example

*Lower Extremity Amputations in Diabetes
(OECD Manual for Data Collection 2017)*

Numerator: All non-maternal/non-neonatal admissions with a procedure code of major lower extremity amputation in any field and a diagnosis code of diabetes in any field in a specified year.

Exclude: Cases resulting from a transfer from another acute care institution (transfers-in). Cases with MDC 14 or specified pregnancy, childbirth, and puerperium codes in any field; Cases with MDC 15 or specified Newborn and other neonates codes in any field; Cases with trauma diagnosis code in any field; Cases with tumour-related peripheral amputation code in any field; Cases that are same day/day only admissions

Denominator 1: Population count.

Denominator 2: Estimated population with diabetes. Countries are requested to provide the diabetes prevalence (%) estimates for each age cohort. It is recognised that countries may not have prevalence estimates for the specified age cohorts, in which case, countries may apply the average or a linear estimate across the cohorts. The population with diabetes will be calculated by applying the estimated proportion (%) of the general population in each age cohort that has diabetes.

Effectiveness – Primary and secondary prevention

PRIMARY/SECONDARY PREVENTION	CD	<i>Vaccination against diphtheria, tetanus and pertussis, children aged 1</i> <i>Vaccination against measles, children aged 1</i> <i>Vaccination against hepatitis B, children aged 1</i> <i>Influenza vaccination coverage, population aged 65 and over</i>
PC	Hypertension hospital admission Annual retinal exam for diabetics	
CC	<i>Mammography screening in women aged 50-69</i> <i>Cervical cancer screening in women aged 20-69</i>	

Source: Carinci, F. et al. (2015), "Towards Actionable International Comparisons of Health System Performance: Expert Revision of the OECD Framework and Quality Indicators", International Journal for Quality in Health Care, Vol. 27, No. 2, pp. 137-146, <http://dx.doi.org/10.1093/intqhc/mzv004>.

Effectiveness – Getting better

GETTING BETTER	AC	Admission-based AMI 30 day in-hospital (same hospital) mortality
		Patient-based AMI 30 day (in-hospital and out of hospital) mortality
		Patient-based ischemic stroke 30 day (in-hospital and out of hospital) mortality
		Admission-based ischemic stroke 30 day in-hospital (same hospital) mortality
		Admission-based hemorrhagic stroke 30 day in-hospital (same hospital) mortality
		Patient-based hemorrhagic stroke 30 day (in-hospital and out of hospital) mortality
		Hip-fracture surgery initiated within 48 hours after admission to the hospital
		Patient-based AMI 30 day in-hospital (any hospital) mortality
		Patient-based ischemic stroke 30 day in-hospital (any hospital) mortality
	CC	Patient-based hemorrhagic stroke 30 day in-hospital (any hospital) mortality
		Breast cancer five year relative survival
		Cervical cancer five year relative survival
		Colorectal cancer five year relative survival
		Breast cancer mortality in women
		Cervical cancer mortality
	PC	Colorectal cancer mortality
		Overall volume of antibiotics for systemic use prescribed
		Volume of cephalosporins/quinolones as proportion of all systemic antibiotics prescribed

Source: Carinci, F. et al. (2015), "Towards Actionable International Comparisons of Health System Performance: Expert Revision of the OECD Framework and Quality Indicators", International Journal for Quality in Health Care, Vol. 27, No. 2, pp. 137-146, <http://dx.doi.org/10.1093/intqhc/mzv004>.

Effectiveness – Chronic care

LIVING WITH ILLNESS OR DISABILITY / CHRONIC CARE	
PC	Asthma hospital admission Chronic Obstructive Pulmonary Disease (COPD) hospital admission Diabetes hospital admission (uncomplicated, short and long-term complications) Diabetes lower extremity amputation Congestive Heart Failure (CHF) hospital admission Adequate use of cholesterol lowering treatment in diabetic patients First choice antihypertensives for diabetes patients
MH	Excess mortality for patients with schizophrenia Excess mortality for patients with bipolar disorder Deaths after discharge from suicide among people diagnosed with a mental disorder Deaths after discharge from suicide among people diagnosed with schizophrenia/bipolar disorder In-patient suicides among people diagnosed with a mental disorder In-patient suicides among people diagnosed with schizophrenia or bipolar disorder Hospital (same) re-admissions within 30 days for patients discharged with schizophrenia Hospital (same) re-admissions within 30 days among patients discharged with schizophrenia Hospital (any) re-admissions within 30 days for patients discharged with schizophrenia Hospital (any) re-admissions within 30 days among patients discharged with schizophrenia Hospital (same) re-admissions within 30 days for patients discharged with bipolar disorder Hospital (same) re-admissions within 30 days among patients discharged with bipolar disorder Hospital (any) re-admissions within 30 days for patients discharged with bipolar disorder Hospital (any) re-admissions within 30 days among patients discharged with bipolar disorder

Source: Carinci, F. et al. (2015), "Towards Actionable International Comparisons of Health System Performance: Expert Revision of the OECD Framework and Quality Indicators", International Journal for Quality in Health Care, Vol. 27, No. 2, pp. 137-146, <http://dx.doi.org/10.1093/intqhc/mzv004>.

Safety – Primary and secondary prevention

PS

Obstetric trauma vaginal delivery with instrument

Obstetric trauma vaginal delivery without instrument

Source: Carinci, F. et al. (2015), "Towards Actionable International Comparisons of Health System Performance: Expert Revision of the OECD Framework and Quality Indicators", International Journal for Quality in Health Care, Vol. 27, No. 2, pp. 137-146, <http://dx.doi.org/10.1093/intqhc/mzv004>.

Safety – Getting better

PS	<p>Retained surgical item or unretrieved device fragment (15+ yrs)</p> <p>Postoperative PE or DVT (all surgical discharges)</p> <p>Postoperative PE or DVT (hip and knee discharges)</p> <p>Postoperative sepsis (all surgical discharges)</p> <p>Postoperative sepsis (all abdominal discharges)</p> <p>Postoperative wound dehiscence (15+ yrs)</p> <p>Retained surgical item or unretrieved device fragment (0-14 yrs)</p> <p>Accidental puncture or laceration (0-14 yrs)</p> <p>Accidental puncture or laceration (15+ yrs)</p> <p>Postoperative haemorrhage or haematoma (0-14 yrs)</p> <p>Postoperative wound dehiscence (0-14 yrs)</p> <p>Postoperative haemorrhage or haematoma (15+ yrs)</p>
PC	<p>Long-term use of benzodiazepines/benzodiazepine-related drugs in elderly patients</p> <p>Use of long-acting benzodiazepines in elderly patients</p> <p>Pilot of prescription safety indicators (6 indicators)</p>

Source: Carinci, F. et al. (2015), "Towards Actionable International Comparisons of Health System Performance: Expert Revision of the OECD Framework and Quality Indicators", International Journal for Quality in Health Care, Vol. 27, No. 2, pp. 137-146, <http://dx.doi.org/10.1093/intqhc/mzv004>.

Responsiveness / Patient centredness

Primary and Secondary Prevention

PE

- Regular doctor spending enough time with patients during the consultation**
- Other doctor spending enough time with patients during the consultation**
- Other doctor providing easy-to-understand explanations**
- Regular doctor providing easy-to-understand explanations**
- Regular doctor giving opportunity to ask questions or raise concerns**
- Other doctor giving opportunity to ask questions or raise concerns**
- Regular doctor involving patients in decisions about care or treatment**
- Other doctor involving patients in decisions about care or treatment**
- Waiting time of more than 4 weeks for getting appointment with a specialist**
- Medical tests, treatment or follow-up skipped due to costs**
- Consultation skipped due to costs**
- Prescribed medicines skipped due to costs**
- Consultation skipped due to difficulties in travelling**
- Waiting time of more than 1 hour on the day of consultation with a doctor**

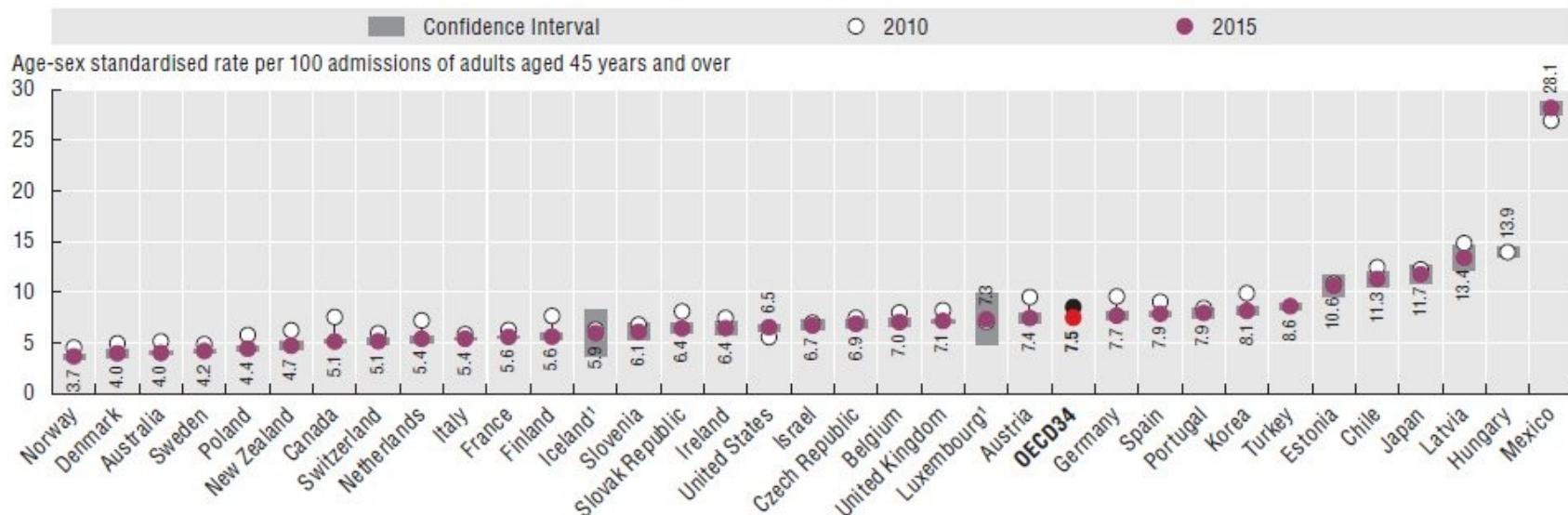
Source: Carinci, F. et al. (2015), "Towards Actionable International Comparisons of Health System Performance: Expert Revision of the OECD Framework and Quality Indicators", International Journal for Quality in Health Care, Vol. 27, No. 2, pp. 137-146, <http://dx.doi.org/10.1093/intqhc/mzv004>.

Mortality after AMI

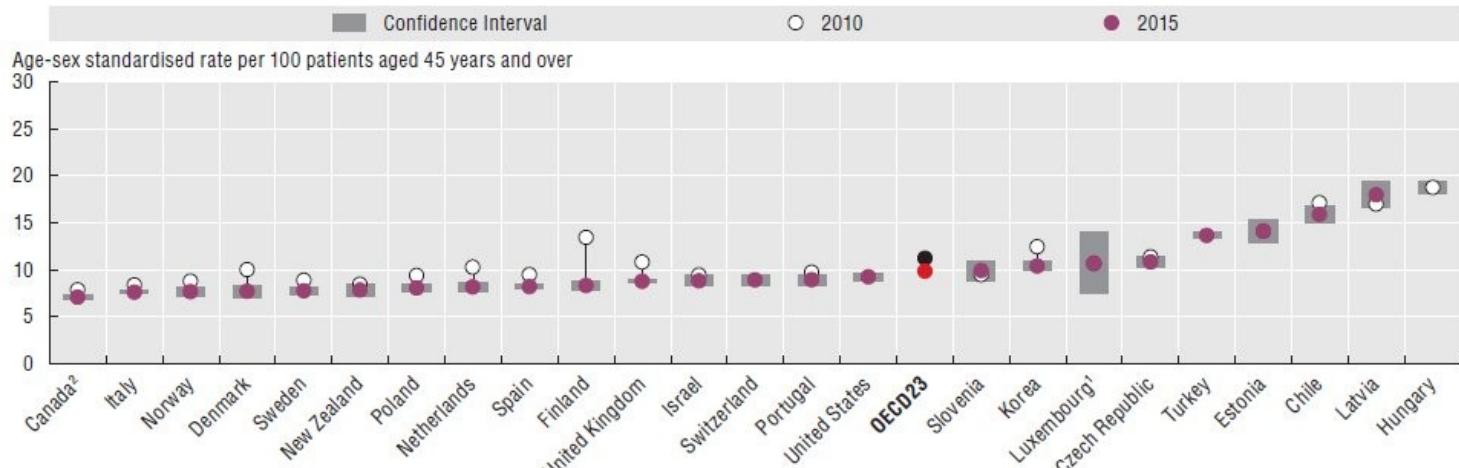
Source: OECD Health at a Glance 2017

http://www.oecd-ilibrary.org/social-issues-migration-health/health-at-a-glance-2017_health_glance-2017-en

6.17. Thirty-day mortality after admission to hospital for AMI based on unlinked data, 2010 and 2015
(or nearest years)



6.18. Thirty-day mortality after admission to hospital for AMI based on linked data, 2010 and 2015
(or nearest years)

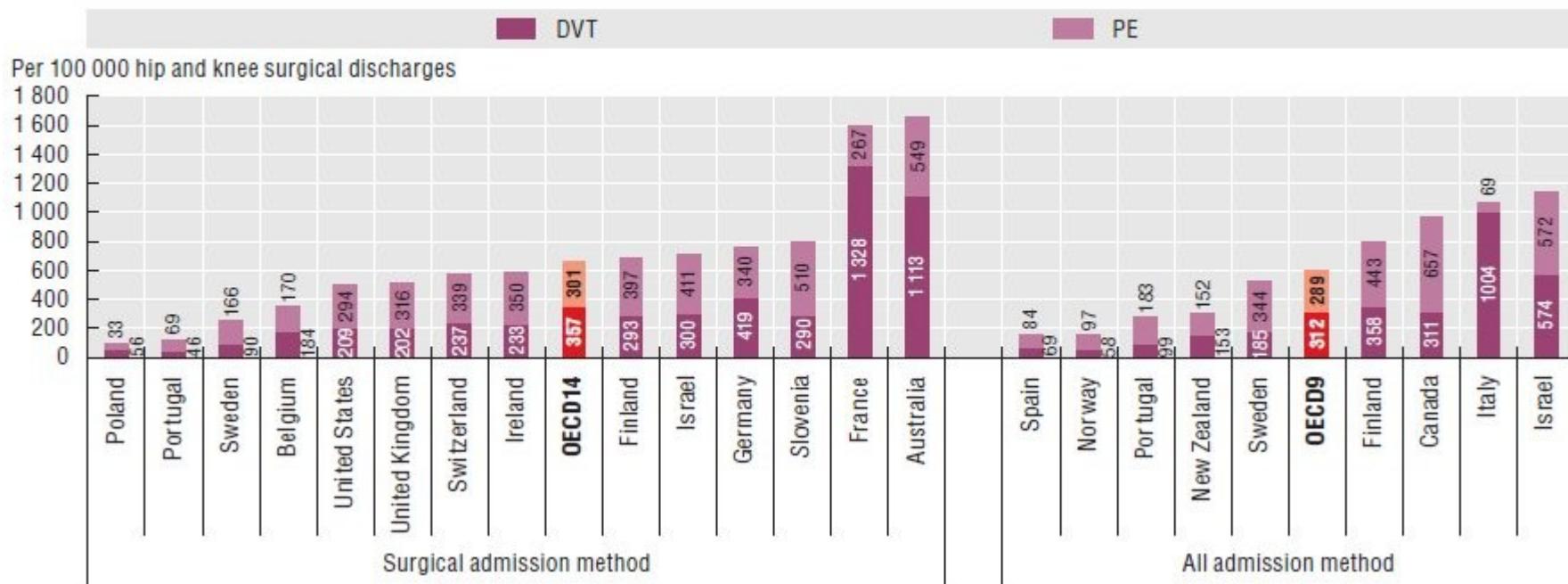


Patient Safety Indicators

Source: OECD Health at a Glance 2017

http://www.oecd-ilibrary.org/social-issues-migration-health/health-at-a-glance-2017_health_glance-2017-en

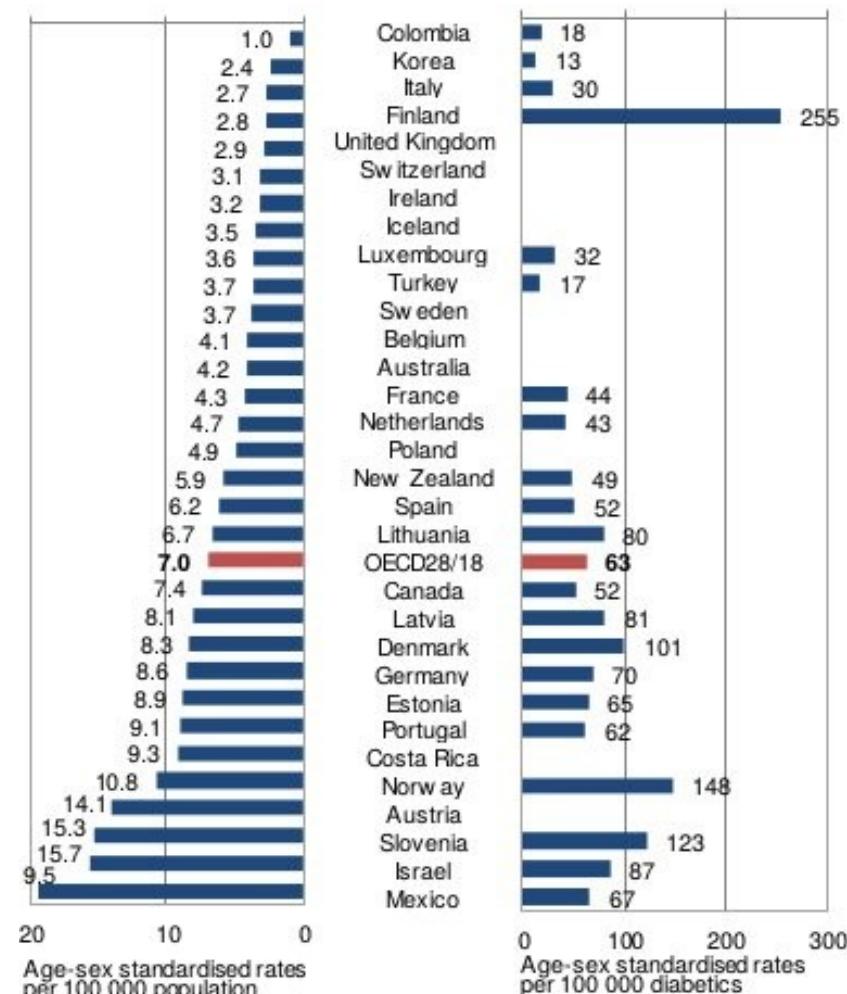
6.25. Postoperative pulmonary embolism (PE) or deep vein thrombosis (DVT) in hip and knee surgeries, 2015 (or nearest year)



Lower extremity amputations in diabetes, 2015

Source: OECD Health at a Glance 2017

6.13. Major lower extremity amputation in adults with diabetes,
2015 (or nearest year)

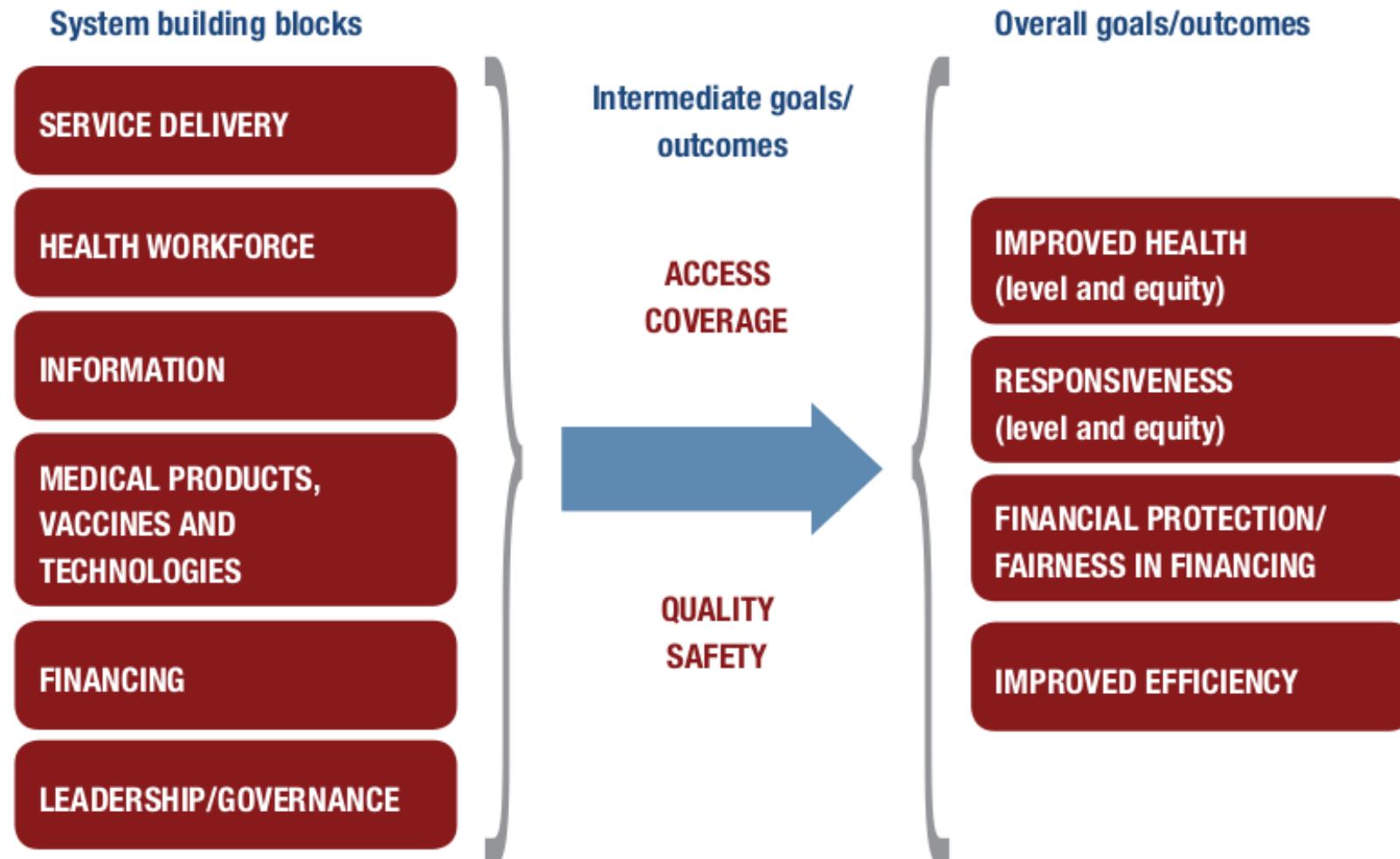


Note: Three-year average for Iceland and Luxembourg.

Source: OECD Health Statistics 2017.

Quality as a means to an end

Fig. 1.1 *Quality is an intermediate goal of health systems*



Source: WHO, 2006

Quality pursued actively with targeted strategies

System level strategies	Organizational/institutional level strategies	Patient/community level interventions
Legal framework for quality assurance and improvement	Clinical quality governance systems	Formalized patient and community engagement and empowerment
Training and supervision of the workforce	Clinical decision support tools	Improving health literacy
Regulation and licensing of physicians and other health professionals	Clinical guidelines	Shared decision-making
Regulation and licensing of technologies (pharmaceuticals and devices)	Clinical pathways and protocols	Peer support and expert patient groups
Regulation and licensing of provider organizations/institutions	Clinical audit and feedback	Monitoring patient experience of care
External assessments: accreditation, certification and supervision of providers	Morbidity and mortality reviews	Patient self-management tools
Public reporting and comparative benchmarking	Collaborative and team-based improvement cycles	Self-management
Quality-based purchasing and contracting	Procedural/surgical checklists	
Pay-for-quality initiatives	Adverse event reporting	
Electronic Health Record (HER) systems	Human resource interventions	
Disease Management Programmes	Establishing a patient safety culture	

Source: authors' own compilation based on Slawomirksi, Auraen & Klazinga, 2017, and WHO, 2018.

Triple aim





RECOMMENDATIONS TO OECD MINISTERS
OF HEALTH FROM THE HIGH LEVEL
REFLECTION GROUP ON THE FUTURE
OF HEALTH STATISTICS

*Strengthening the international comparison
of health system performance through
patient-reported indicators*

January 2017



<https://www.oecd.org/health/ministerial/>



Value-based health care (M.Porter)

<https://www.oecd.org/health/ministerial/>

Perspective
DECEMBER 23, 2010

Solving the Health Care Problem

- The fundamental **goal and purpose** of health care is to improve **value for patients**

$$\text{Value} = \frac{\text{Health outcomes that matter to patients}}{\text{Costs of delivering these outcomes}}$$

- Delivering high value health care is the **definition of success**
- Value is the only goal that can **unite the interests** of system participants
- Improving value is the **only real solution**
- The question is how to design health care delivery systems and organizations that **substantially improve patient value**



Source: "What is Value in Health Care" (Michael Porter, New England Journal of Medicine, 2010)

What Is Value in Health Care?

Michael E. Porter, Ph.D.

In any field, improving performance and accountability depends on having a shared goal that unites the interests and activities of all stakeholders. In health care, however, stakeholders have

myriad, often conflicting goals, including access to services, profitability, high quality, cost containment, safety, convenience, patient-centeredness, and satisfaction. Lack of clarity about goals has led to divergent approaches, gaming of the system, and slow progress in performance improvement.

Achieving high value for patients must become the overarching goal of health care delivery, with value defined as the health outcomes achieved per dollar spent.¹ This goal is what matters for patients and unites the interests of all actors in the system. If value improves, patients, payers, providers, and suppliers can all benefit while the economic sustainability of the health care system increases.

Value — neither an abstract ideal nor a code word for cost reduction — should define the framework for performance improvement in health care. Rigorous, disciplined measurement and improvement of value is the best way to drive system progress. Yet value in health care remains largely unmeasured and misunderstood.

Since value is defined as outcomes relative to costs, it encompasses efficiency. Cost reduction without regard to the outcomes achieved is dangerous and self-defeating, leading to false "savings" and potentially limiting improvement.

Outcomes, the numerator of the value equation, are inherently condition-specific and multidimensional. For any medical condition, no single outcome captures the results of care. Cost, the equation's denominator, refers to the total costs of the full cycle of care for the patient's medical condition, not the cost of individual services. To reduce cost, the best approach is often to spend more on some services to reduce the need for others.

N ENGL J MED 363;26 NEJM.ORG DECEMBER 23, 2010

The New England Journal of Medicine

Downloaded from nejm.org on September 13, 2017. For personal use only. No other uses without permission.

2477

Medical conditions

<https://www.oecd.org/health/ministerial/>

Principles of Value-Based Health Care Delivery

- Value **cannot be understood** at the level of a hospital, specialty, intervention, or for overall primary care
- Value is created in caring for a patient's **medical condition** over the **full cycle of care**

$$\text{Value} = \frac{\text{Set of outcomes that matter to patients for the condition}}{\text{Total costs of delivering them over the full care cycle}}$$



- In **primary and preventive care**, value is created in serving **segments of patients** with similar primary and preventive needs



- The most powerful single lever for reducing cost and improving value is **improving outcomes**



CREATE INTEGRATED PRACTICE UNITS (IPUs)

Organize care around patient medical conditions and distinct patient segments.



MEASURE OUTCOMES

Measure health outcomes for every patient.



MEASURE COSTS

Measure the actual costs of patient care.



BUNDLED PRICES

Reimburse the full care cycle for medical conditions.



SYSTEMS INTEGRATION

Clinically integrate care across separate units and facilities using an IPU structure.



GEOGRAPHIC EXPANSION

Increase the geographic reach of leading providers in their areas of excellence.



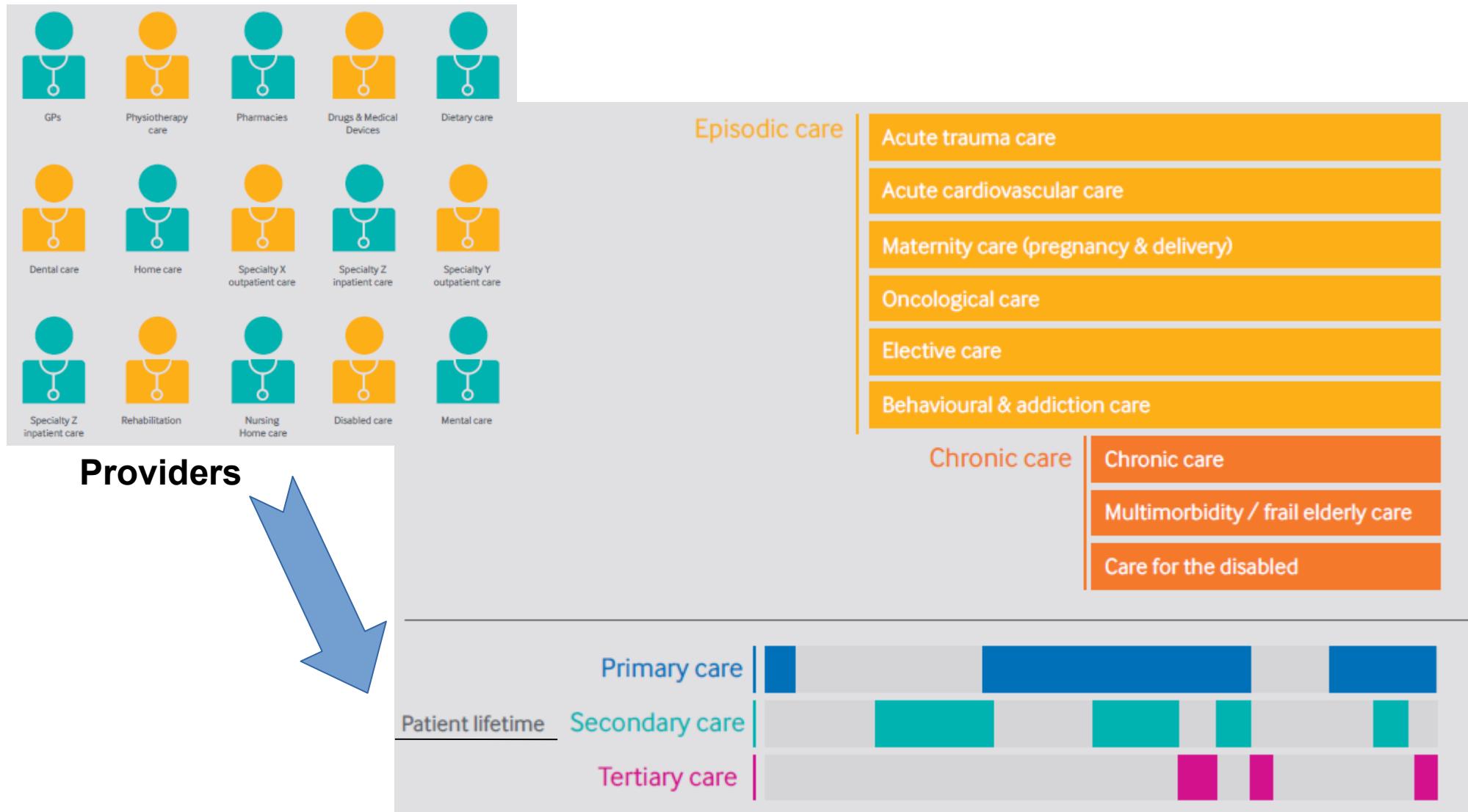
BUILD AN ENABLING INFORMATION TECHNOLOGY PLATFORM

Use information technology to help restructure care delivery and accurately measure results.

Value-Based Health Care Delivery

<http://www.isc.hbs.edu/health-care/vbhcd/Pages/default.aspx>

Person-centred care



The future of health statistics

- Patient-Reported Outcome Measures” (PROMs)
 - Measure patients’ perceptions of their health status, clinical outcomes, mobility and quality of life. Examples: What was a patient’s mobility like before a hip replacement, and did it improve after the intervention? Does a patient’s condition limit their ability to do strenuous activities such as jogging, skiing or cycling?
- Patient-Reported Experience Measures (PREMs)
 - Measure patients’ perceptions of their experience of care by focusing on the process of care and how that has an impact on their experience. Examples: Did the patient wait long for treatment? Did the patient feel they were involved in decision making?
- Patient Activation Measures” (PAMs)
 - measure the extent to which patients are activated in improving and maintaining their health through self-management
- Patient-Reported Incident Measures” (PRIMs)
 - measure safety incidents

ICHOM (www.ichom.org)



ABOUT [STANDARD SETS](#) MEASURE TECHHUB EVENTS & MEDIA SUPPORT US LOGIN



COMPLETED CONDITIONS

All 21	Cardiovascular 3	Congenital anomalies 2	Digestive 1	Malignant neoplasms 5	Maternal and neonatal 1	Mental and behavioral disorders 1
Musculoskeletal 2	Neurological 2	Primary/preventative care 1	Sense organ 2	Urogenital 1		



Pregnancy and Childbirth
Maternal and neonatal



Inflammatory Bowel Disease
Digestive



Overactive Bladder
Urogenital

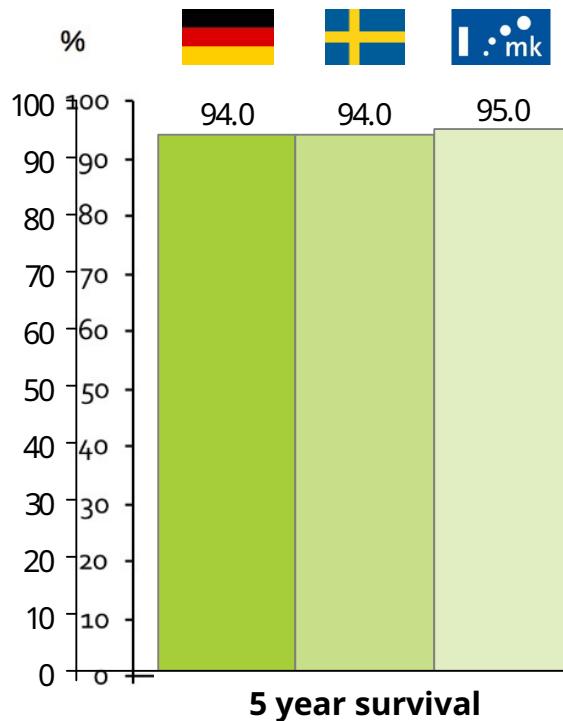


Colorectal Cancer
Malignant Neoplasms

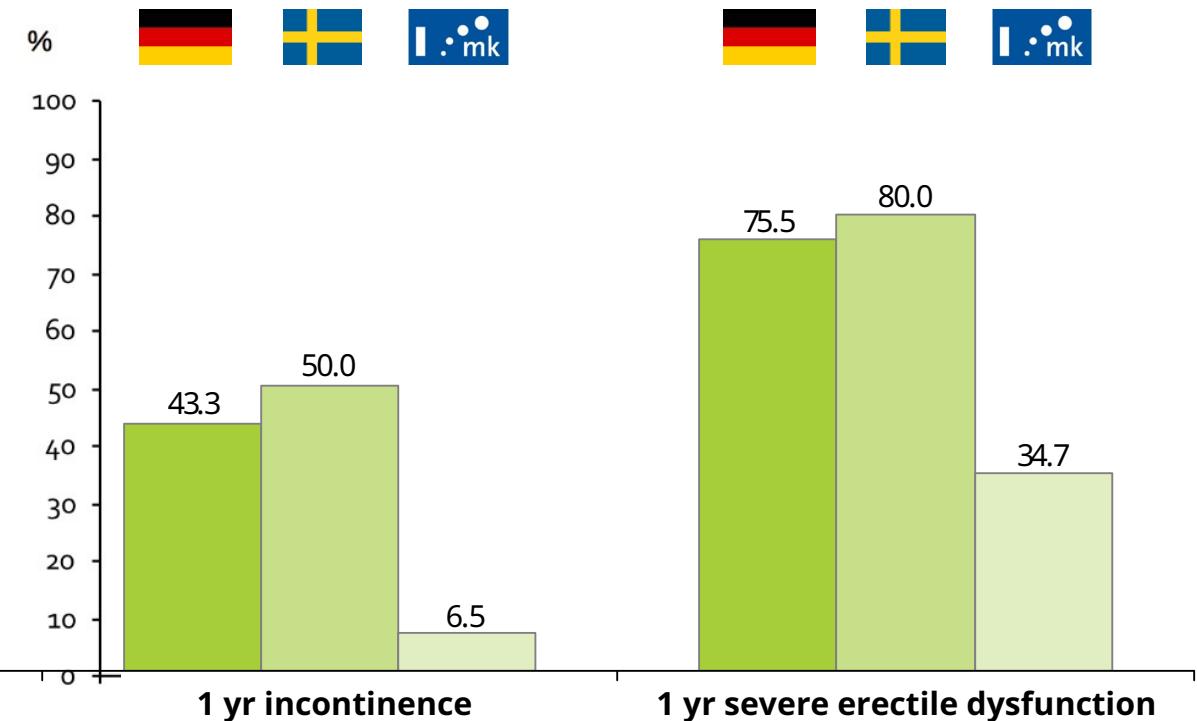
Why measuring and reporting meaningful outcomes matters

Comparing outcomes of prostate cancer care

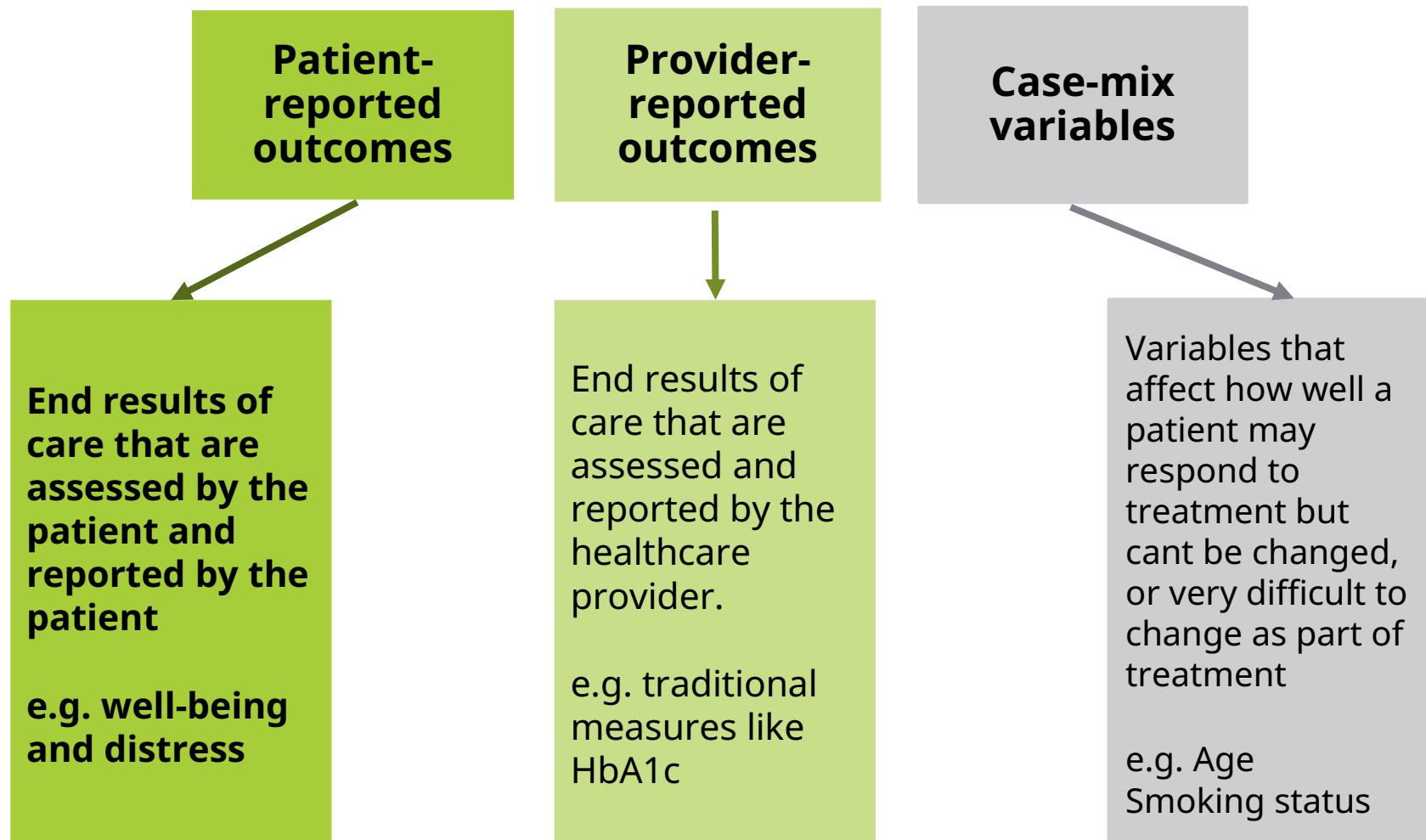
Focussing on mortality alone...



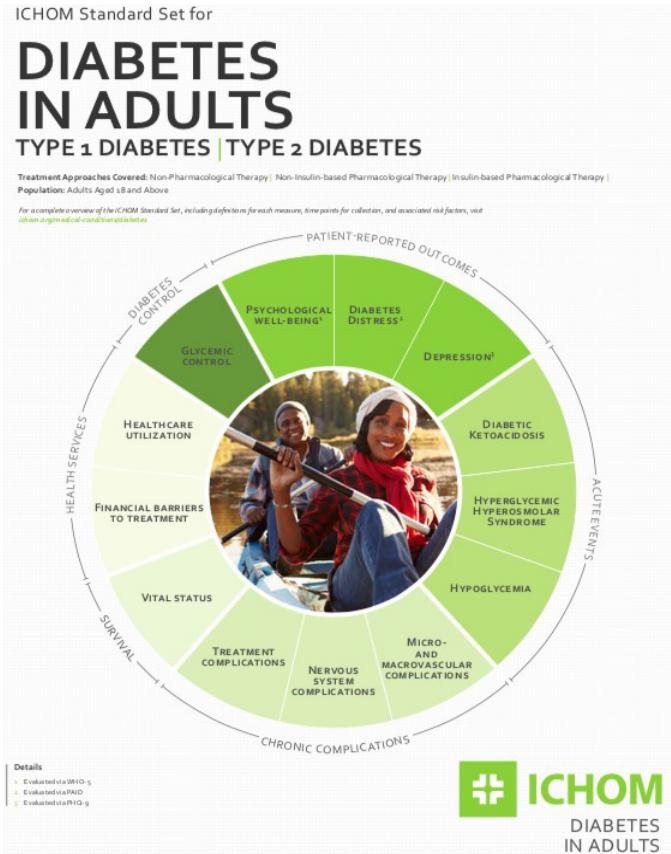
...may obscure large differences in outcomes that matter most to patients



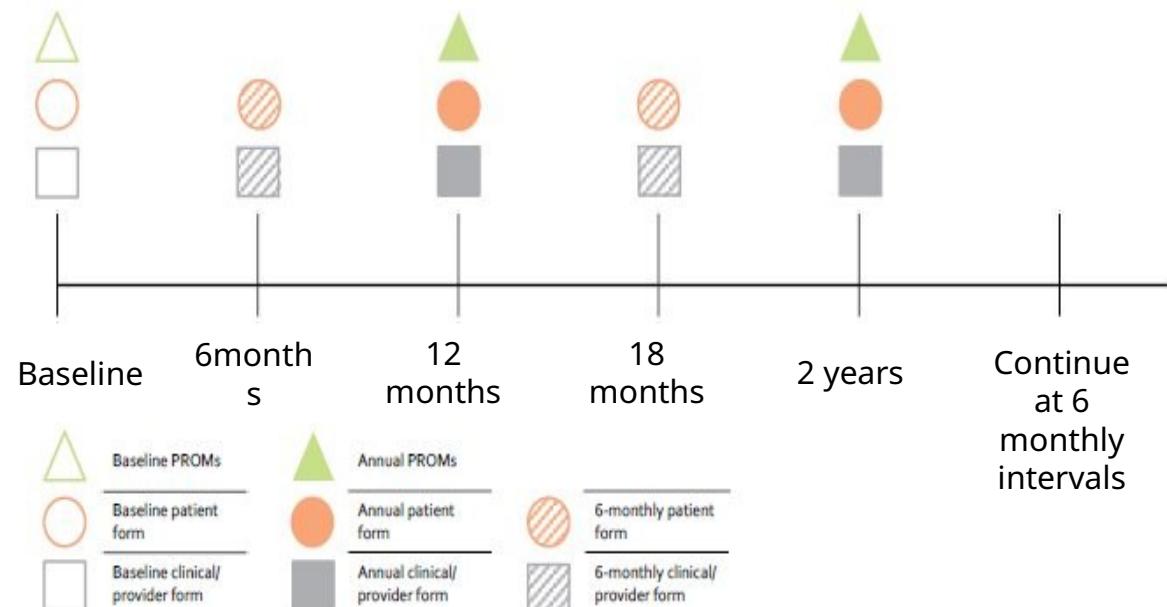
The Standard Set for Diabetes contains 3 categories of data



Example of ICHOM Standard Set: Diabetes

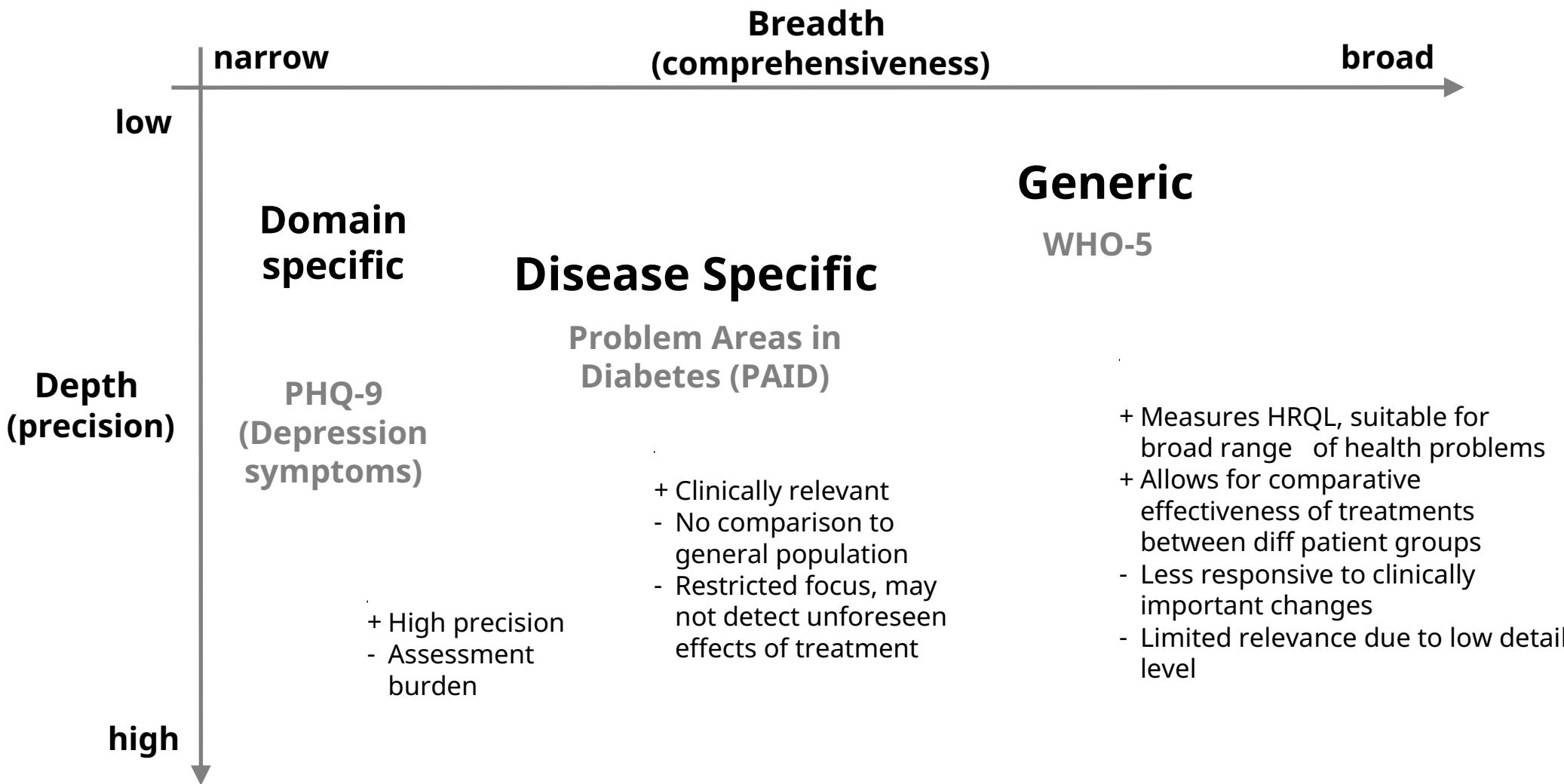


Follow-up algorithm



<http://www.ichom.org/medical-conditions/diabetes/>

The diabetes standard set includes 3 PROMs



Key messages

- The OECD health systems performance framework includes several dimensions through which we can evaluate health systems. Quality of care is a key element of the OECD framework.
- Performance indicators are used to populate the quality matrix, each with a numerator and denominator
- Various methods have been adopted to define standardized comparable indicators. The recent initiative by ICHOM allowed using common definitions for many different clinical conditions, particularly for PROMs.
- The future of health statistics involve more user centred approaches, e.g. Patient Reported Experience Measures (PREMs) and Patient Reported Incidence Measures (PRIMs).

Materials

- *Course notes*
- OECD-WHO, Improving Health Care Quality in Europe, p.31-62
- OECD Health at a Glance 2019, p.41-60
- M.Porter, What is Value in Health Care, NEJM, 2010



ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Second cycle degree programme (LM) in Statistical Sciences

85278 – Methods and tools for Health Statistics

**85277 – Methods and tools for Official Statistics: Population
and Health Statistics**

a.a. 2022/2023

Stream 2. Risk stratification and standardization

Topic 2.1.1

Standardization methods in AHRQ and OECD indicators

Fabrizio Carinci

fabrizio.carinci@unibo.it

Tuesday, 21st February 2023

Performance measurement

- To make **decision makers and professionals accountable** for the results achieved in an organization (from a single unit to a whole system)
- To evaluate **adherence to agreed processes** and set new targets in a structured way
- To share **best practices** and avoid common mistakes
- To **benchmark** the results across different units and/or organizations, using standardized criteria

Performance Indicators

- Performance measurement is carried out using Key Performance Indicators (KPIs) focusing on specific aspects of health care (e.g. specific procedures implemented within an organization)
- In this way, continuous rounds of measurement and evaluation (“audit” and “feedback”) can be carried out to ensure that organizations and the system as a whole can continuously improve
- A range of statistical methods are used to compare KPIs and report on health systems performance at different levels of health care systems (hospitals, districts, regions, etc)

Why do we need advanced methods?

- To control for potential **bias** in the construction and estimation of performance indicators (*definitions, algorithms, data sources, type of study/data collection*)
- To **adjust** comparisons **for imbalances** in the composition of different populations to be compared (*risk adjustment and standardization*)
- To avoid drawing conclusions from **random variation**, which tends to be critical when the number of cases arising from smaller populations tends to be limited by definition (*confidence intervals, graphical representation*)

Sources of bias in HCQIs

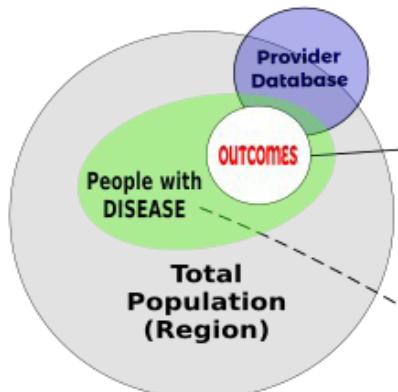
DATA

STATISTICAL OUTPUT

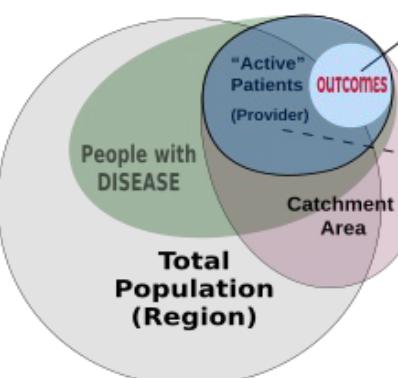
POPULATION-BASED DISEASE REGISTER



POPULATION-BASED DISEASE REGISTER LINKED TO A PROVIDER SOURCE



PROVIDER-BASED SOURCE



INDICATOR

Numerator

Denominator

General solutions for bias

- Controlling for bias in any analysis of routine data may **not be simple**
- Definitions can be improved to address more targeted outcomes and populations
- Background information on data sources can be used to enhance data standardization and improve comparability of results
- Quasi-experimental methods can be applied on top of routine data e.g. resampling (bootstrapping, cross-validation)
- Selection bias for known confounders (e.g. sicker patients in one hospital when comparing vs another) can be *partially* solved using **risk adjustment techniques**
- Methods e.g. propensity scores can be included in the regression models to enhance the reliability of routine data (observational) compared to randomised trials

Example: mortality rates

- Mortality rates are an apparently straightforward case of performance indicator that is valid across health system organizations
- Fraction of cases dying in a hospital coronary unit after being admitted for acute myocardial infarction (or within 30 days from discharge)
- Even when SMART principles are duly implemented, comparisons across hospitals may not be simple
- Cases in one hospital can be older or more severe, resulting in a higher likelihood of death. Comparisons should be based on equal terms.
- A comparison between **crude rates** can be *misleading*. That is why we need to “standardize” results, so that the evaluation of potential causes, and specific improvements, may be triggered within the organization
- Different methods are available, some of which are quite sophisticated and resource intensive

Standardized rates

- **Crude rates** are calculated based on the population under study **as a whole**
- **Standardized rates** are based on **specific characteristics** whose distribution is taken as a standard
- Age and Sex are two of the most common variables used for standardization
- There are two methods for calculating standardized rates:
 - ***Direct standardization***
 - ***Indirect standardization***

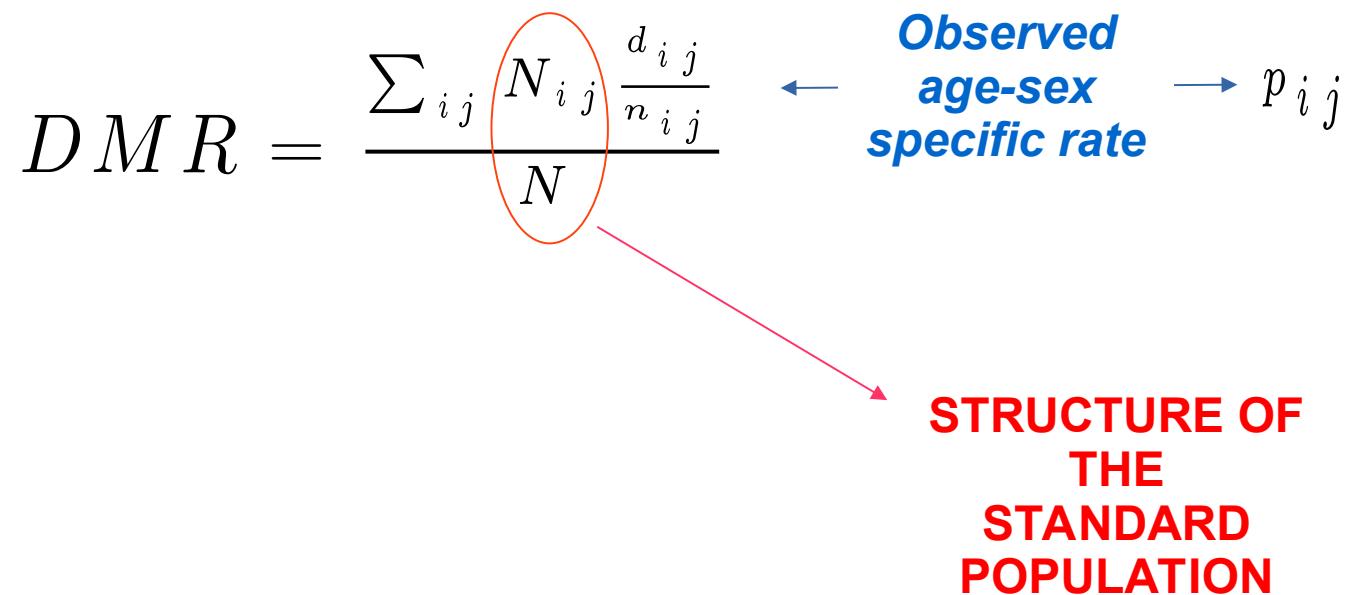
Direct standardized rates

- **Direct standardization** is obtained by dividing the total expected number of cases in a standard population by the standard population size

$$DMR = \frac{\sum_{i,j} N_{i,j} \frac{d_{i,j}}{n_{i,j}}}{N}$$

*Observed
age-sex
specific rate* $\rightarrow p_{i,j}$

STRUCTURE OF
THE
STANDARD
POPULATION



Direct standardized rates: confidence interval

- ***The standard error of a directly standardized rate is given by:***

$$DMR \pm 1.96 \sqrt{\sum_{ij} N_{ij}^2 \frac{p_{ij}(1-p_{ij})}{n_{ij}}}$$

Important property

- ***Direct standardization*** allows direct comparisons between two different rates obtained from different units or populations
- That is because standardized rates are calculated using the same standard population

Indirect standardized rates

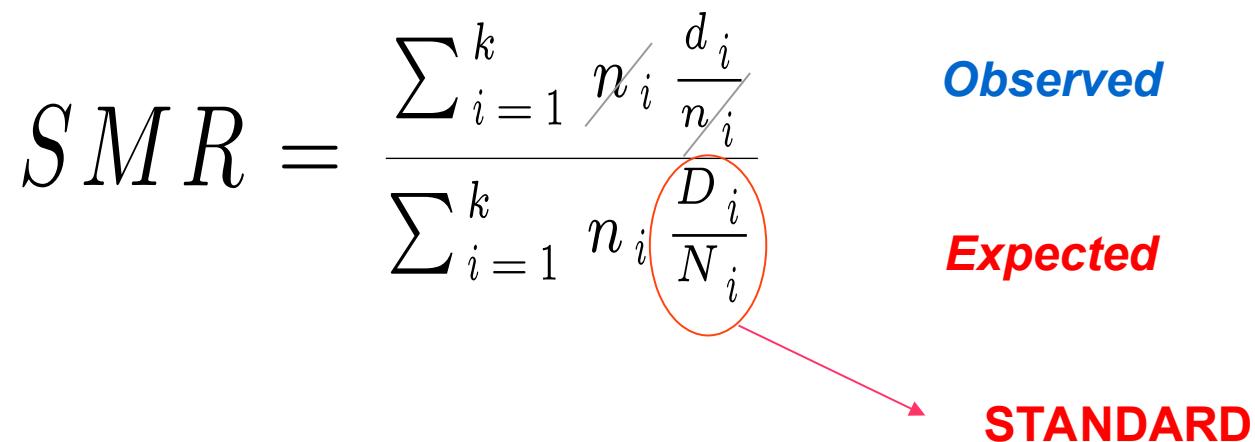
- **Indirect standardization** leads to the so-called “Standardized Mortality Ratio”, defined as the *ratio of the number of deaths observed in a population (in a specific time period) to those expected if it had the same age- and sex- specific death rates as a “standard” population*

$$SMR = \frac{\sum_{i=1}^k n_i \frac{d_i}{n_i}}{\sum_{i=1}^k n_i \frac{D_i}{N_i}}$$

Observed

Expected

STANDARD



Indirect standardized rates: confidence interval

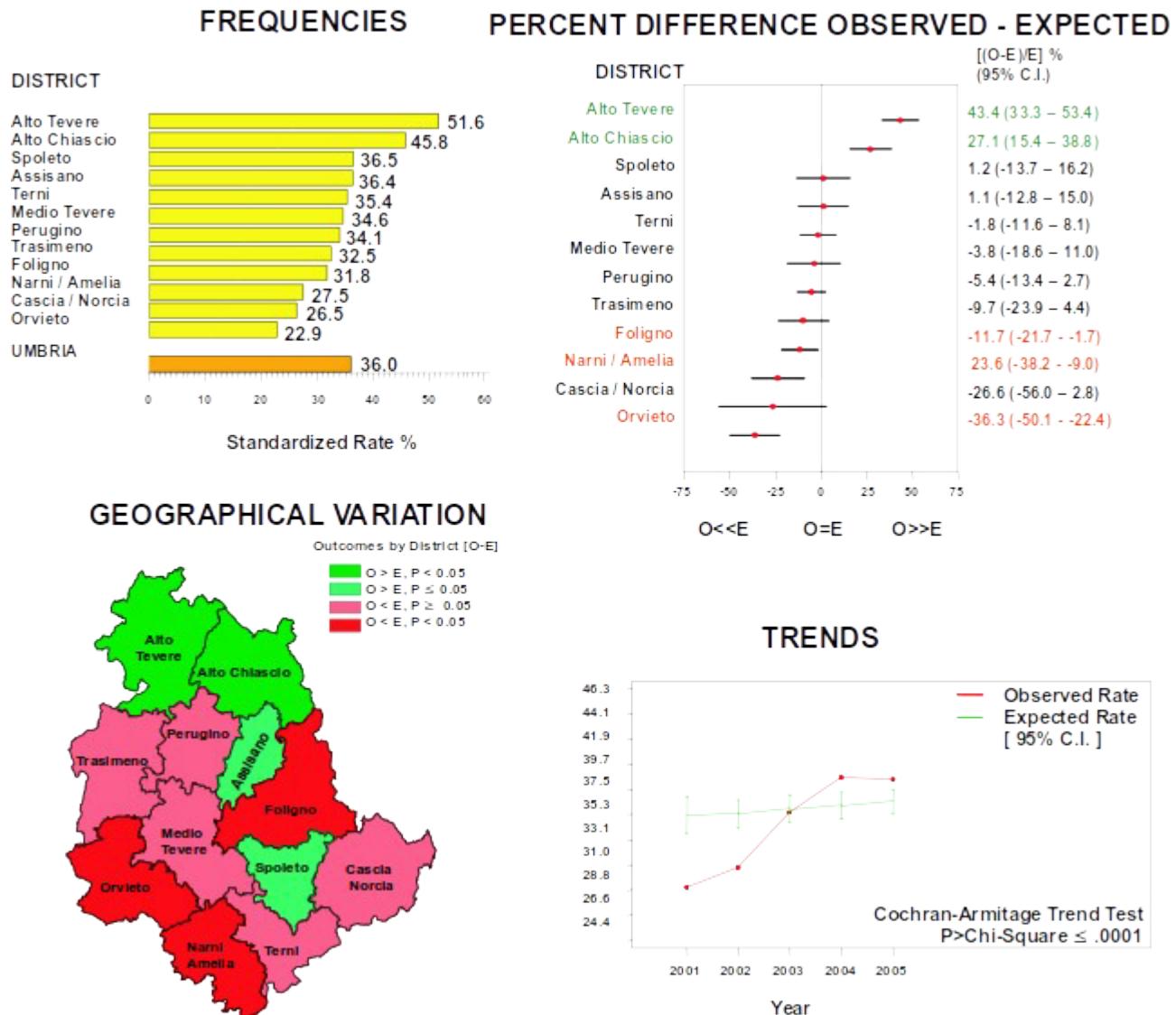
- ***The confidence interval of a standardized mortality ratio is given by:***

$$\frac{S M R}{\exp\left(\frac{1.96}{\sqrt{\sum_{i=1}^k d_i}}\right)} \text{ to } S M R * \exp\left(\frac{1.96}{\sqrt{\sum_{i=1}^k d_i}}\right)$$

Limitations of indirect standardization

- ***Indirect standardization*** does not allow direct comparisons between two different rates obtained from different units or populations
- That is because adjusted rates for different populations are not calculated using exactly the same standard population: weighting factors are different
- It is possible to compare the observed rate in a specific unit or population against the standard population **only**
- **Although technically incorrect, this is a commonly miskaken interpretation frequently due to graphical outputs**

Summary Results using AHRQ methodology



Direct standardized rates: example (1)

AGE	Population A			Population B		
	D	N	Rate	D	N	Rate
0-24	35	18000	1.94	30	13000	2.31
25-49	60	11000	5.45	50	7000	7.14
50-74	370	9000	41.11	400	11000	36.36
75+	250	3000	83.33	380	4000	95.00
TOTAL	715	41000	17.44	860	35000	24.57

CRUDE MORTALITY RATE (B) >> CRUDE MORTALITY RATE (A)

Direct standardized rates: example (2)

AGE	Population A			Population B		
	D	N	Rate	D	N	Rate
0-24	35	18000	1.94	30	13000	2.31
25-49	60	11000	5.45	50	7000	7.14
50-74	370	9000	41.11	400	11000	36.36
75+	250	3000	83.33	380	4000	95.00
TOTAL	715	41000	17.44	860	35000	24.57

AGE	Population A			Population B		
	Reference Population	Rate	Expected Deaths	Reference Population	Rate	Expected Deaths
0-24	11000	1.94	21.34	11000	2.31	25.41
25-49	17000	5.45	92.65	17000	7.14	121.38
50-74	20000	41.11	822.20	20000	36.36	727.20
75+	3000	83.33	249.99	3000	95.00	285.00
TOTAL	51000		1186.18	51000		1158.99

Direct standardized rates: example (3)

AGE	Population A			Population B		
	Reference Population	Rate	Expected Deaths	Reference Population	Rate	Expected Deaths
0-24	11000	1.94	21.34	11000	2.31	25.41
25-49	17000	5.45	92.65	17000	7.14	121.38
50-74	20000	41.11	822.20	20000	36.36	727.20
75+	3000	83.33	249.99	3000	95.00	285.00
TOTAL	51000		1186.18	51000		1158.99

Age adjusted death rate for population A = $1186.18/51000=23.3 \times 1000$

Age adjusted death rate for population B = $1158.99/51000=22.7 \times 1000$

The standardized mortality rates are very close!

Indirect standardized rates: example (1)

	Population A				Population B			
AGE	n	D/N	Expected d	n	D/N	Expected d		
0-24	2000	4.0	8.0	1000	4.0	4.0		
25-49	2500	7.0	17.5	1500	7.0	10.5		
50-74	3500	10.0	35.0	2500	10.0	25.0		
75+	4500	30.0	135.0	1000	30.0	30.0		
TOTAL	12500		195.5	6000		69.5		

SMR for population A = $120 / 195.5 = 0.61$

SMR for population B = $30 / 69.5 = 0.43$

The standardized mortality RATIOS are rather different but cannot be compared

Key messages

- Rigorous methods are required to measure and compare performance through a multidimensional set of indicators.
- Advanced methods can control for bias, using different techniques that allow standardizing indicators in different ways
- The methods of direct and indirect standardization are the two main techniques used to compare indicators across different organizational units. Different properties and limitations are present in both approaches.

Materials

- *Course notes*
- OECD Direct Standardization method (technical guide)
- Rolfe KA et al, Standardisation of rates using logistic regression: a comparison with the direct method, BMC Health Services Research 2008, 8:275



ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Second cycle degree programme (LM) in Statistical Sciences

85278 – Methods and tools for Health Statistics

**85277 – Methods and tools for Official Statistics: Population
and Health Statistics**

a.a. 2022/2023

Stream 2. Risk stratification and standardization

Topic 2.1.2

*Indirect standardization using Multivariate Risk Adjustment
and Model building strategies*

Fabrizio Carinci

fabrizio.carinci@unibo.it

Monday, 27th February 2023

US AHRQ Quality Indicators

- The AHRQ Quality Indicators (QIs) have been developed to assess health care quality.
- Four modules measuring various aspects of quality:
 - Inpatient Quality Indicators (IQIs)
 - Patient Safety Indicators (PSIs)
 - Pediatric Indicators (PDIs)
 - Prevention Quality Indicators (PQIs).
- The AHRQ QIs are available for public use at no charge. Resource materials on the QIs can be downloaded at:
www.qualityindicators.ahrq.gov

AHRQ QI Fundamental definitions

- *Counts:*
 - **Eligible population** = total number of hospital discharges that qualified for the eligible population for that specific indicator
 - **Observed events** = total sum of events occurred in the eligible population for that specific indicator
 - **Expected events** = total sum of events expected to occur for that specific indicator **if the hospital had average performance comparable to the reference population, considering its case mix**
- *Rates:*
 - **Observed rate** = Observed events / Eligible population
 - **Expected rate** = Expected events / Eligible population
 - **Risk-adjusted rate** =
(Observed events / Expected events) * reference population rate



Agency for Healthcare
Research and Quality

Careers | Contact Us | Español | FAQs | Email Updates

AHRQ QI ▾

Search...



AHRQ Quality Indicators™

Home ▾

Measures ▾

Software ▾

News

Resources ▾

FAQs

Archives ▾

Quality Improvement and monitoring at your fingertips

Get to know the AHRQ Quality Indicators

PQI

Prevention Quality
Indicators

[Learn about PQI >](#)

IQI

Inpatient Quality
Indicators

[Learn about IQI >](#)

PSI

Patient Safety
Indicators

[Learn about PSI >](#)

PDI

Pediatric Quality
Indicators

[Learn about PDI >](#)

AHRQ INDIRECT Standardization method

- Indirect standardization: the results that each hospital would obtain with own population ("case-mix") if it had the same "performance" of the reference population
- This does not allow comparisons BETWEEN hospitals

Risk adjustment model (based on national sample)

$$Y(\%) = \beta_0 + \beta_1(\text{females}) + \beta_2(\text{age}) + \beta_3(\text{comorbidity}) + \dots$$



Service unit (hospital, district,...)

$$Y_i \text{ expected} = \beta_0 + \beta_1(\text{females}) + \beta_2(\text{age}) + \beta_3(\text{comorbidity}) + \dots$$

$$\sum \text{Pred}_i \times 100 = \text{Expected Rate}$$

Standardized Rate=

(observed rate/expected rate)*population rate

AHRQ INDIRECT Standardization method

https://qualityindicators.ahrq.gov/Downloads/Resources/Publications/2022/Empirical_Methods_2022.pdf

Y_{ij} = 0 or 1, outcome for patient j in hospital i.

X_{ij} = covariates (e.g., gender, age, DRG, comorbidity)

P_{ij} = predicted probability from logit of Y on X

= $\exp(X_{ij}\beta)/[1+ \exp(X_{ij}\beta)]$

where β is estimated from logit on entire sample.

n_i = number of patients in sample at hospital i.

α = average outcome in the entire sample:
reference population rate

O_i = $(1/n_i) \sum(Y_{ij})$ OBSERVED RATE at hospital i

E_i = $(1/n_i) \sum(P_{ij})$ EXPECTED RATE at hospital i

RISK ADJUSTED RATE for hospital i: $RAR_i = \alpha(O_i / E_i)$

$VAR(RAR_i) = (\alpha/E_i)^2 (1/n_i)^2 \sum[P_{ij}(1-P_{ij})]$

Example of Indirect Standardization (AHRQ)

National rate
= 0.05

Type of Rate	Hospital A	Hospital B
Observed	0.02	0.06
Expected	0.04	0.10
Risk-adjusted	0.025	0.03

- It is not clear whether Hospital A or Hospital B has better or worse than average performance, **compared to the national rate**, because they may have different *case mixes* than the *national* population
- Hospital A has an expected rate of 0.04. Since its expected rate is lower than national, its **mix of patients is at lower risk** than the average case mix. Still, expected rate > observed rate, so the hospital is **performing better than expected** on its case mix of patients.
- Hospital B has an expected rate of 0.10. Since its expected rate is higher than national, its **mix of patients is at higher risk** than the average case mix. Since expected rate > observed rate, the hospital **also is performing better than expected** on its case mix of patients.

OECD DIRECT Standardization

- Direct standardization: the results that each country would obtain with own “performance” if it had the same population as the OECD average
- This allows comparisons BETWEEN countries

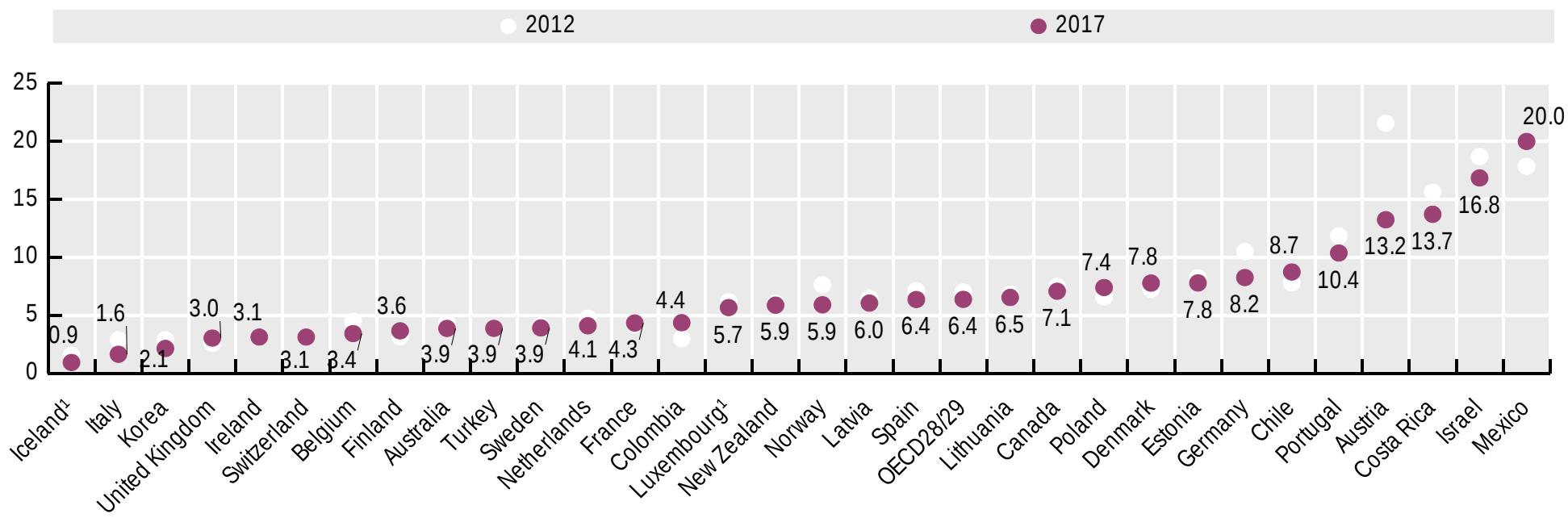
Sex	Age-group								
	0-14	15-19	20-24	25-29	30-34	35-39	40-44	45-49	
Male	Data are for adults only	41779971	42510958	43903155	43267382	44525464	43904781	43687912	
Female		39928651	41007212	42600096	42366599	43907385	43659988	44012322	
Total	81708622	83518170	86503251	85633981	88432849	87564769	87700234		
	50-54	55-59	60-64	65-69	70-74	75-79	80-84	85+	
Sex									
	Male	40224853	36052966	32284843	24545544	20135366	15164034	10097920	6845013
Female	41214826	37679220	34385226	27359105	24076627	20333960	16238522	15563267	
Total	81439679	73732186	66670069	51904649	44211993	35497994	26336442	22408280	

$$SR_{TOT} = \frac{\sum_{ij} (ASR_{ij} \times POP_{ij})}{POP_{TOT}}$$

Major lower extremity amputation in adults with diabetes

Source: OECD Health at a Glance 2019

https://www.oecd-ilibrary.org/social-issues-migration-health/major-lower-extremity-amputation-in-adults-with-diabetes-2012-and-2017-or-nearest-year_77b55d01-en



Caterpillar plot: limitations

EC Marshall, D Spiegelhalter, BMJ 316, 1701-1705

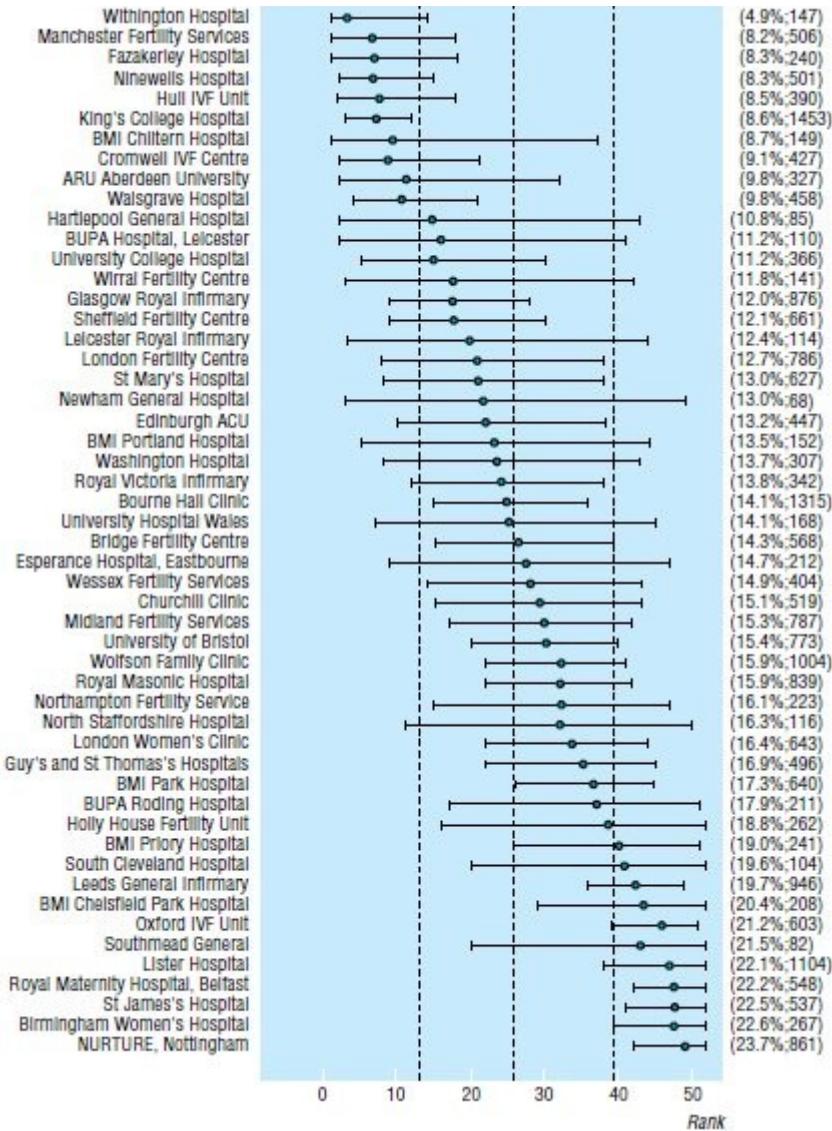


Figure 2 shows the point estimates and 95% confidence intervals for the ranks associated with these success rates (the higher the rank the better). The intervals are generally wide, illustrating the great uncertainty associated with the ranks. The 18 "significantly" outlying clinics can all be confidently placed in either the top or bottom half of the table, although we can conclude that only one clinic is in the lower quarter and five are in the upper quarter.

....Figure 3: change in years!

Funnel plots

- Funnel plots have been initially proposed by DJ Spiegelhalter as a valid robust alternative to caterpillar plots in comparing institutional performance (*DJ Spiegelhalter, Funnel plots for comparing institutional performance, Statist. Med. 2005; 24:1185–1202*). The idea was based on the use of funnel plots as a means for visual inspection of publication bias in meta-analysis.
- Funnel plots are a form of scatter plot in which observed rates are plotted against the total number of cases from which they have been calculated (e.g. the volume of cases in deaths after myocardial infarction).
- Overlays include a horizontal line representing the population average (optional) and binomial control limits at 95% and 99.8% confidence limits (consistently with 1.96,3 sigma deviations), generally calculated as:

$$\hat{p} \pm \Phi_{(1 - \frac{\alpha}{2})} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$$

cumulative inverse normal distribution

Using funnel plots in public health surveillance

Dover DC, Schopflocher DP, Popul Health Metr. 2011 Nov 10;9(1):58

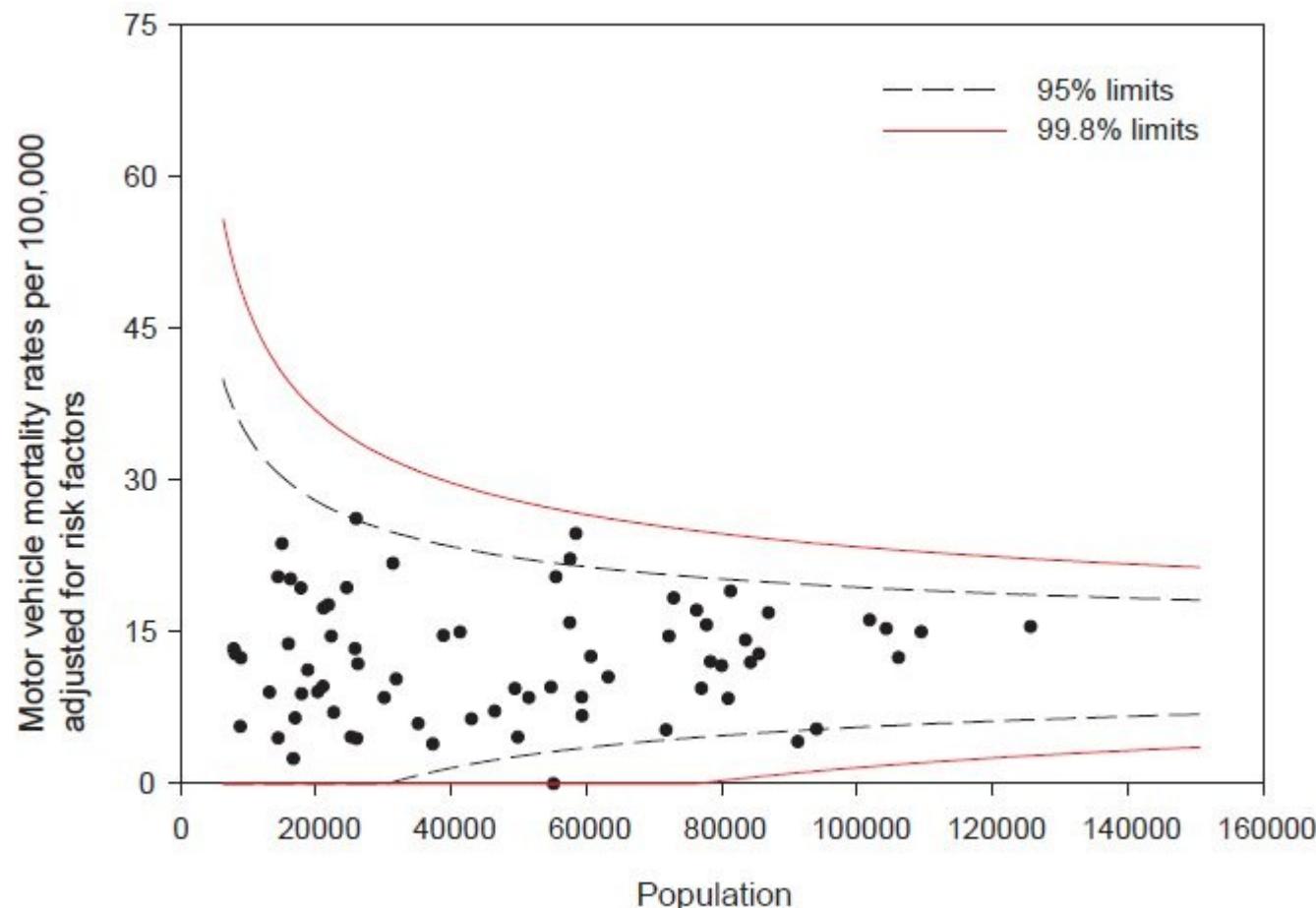
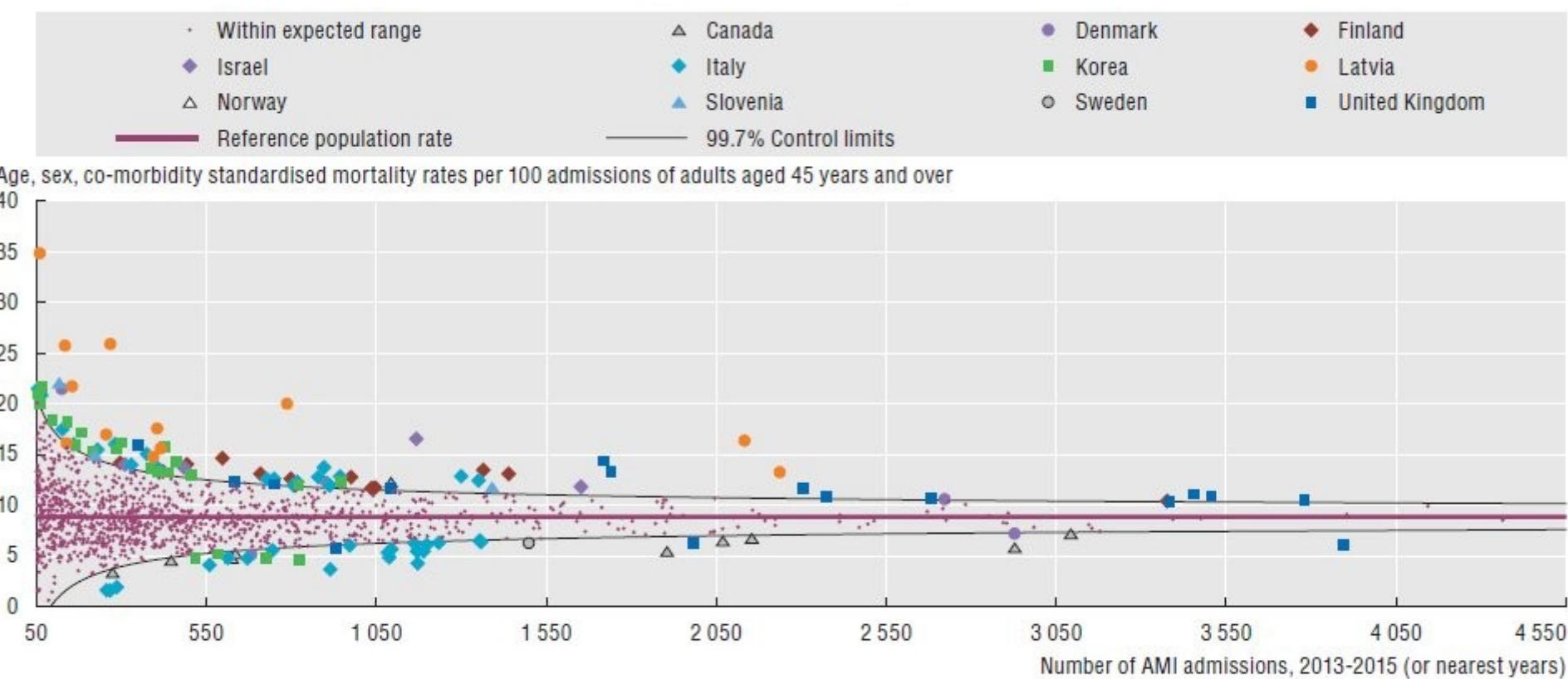


Figure 2 Funnel plot of adjusted motor vehicle traffic mortality rates. Motor vehicle mortality rates are adjusted for age, sex, seat belt use, and road type and utilization.

OECD Funnel Plot

6.20. Thirty-day mortality after admission to hospital for AMI based on linked data, 2013-2015
(or nearest years)

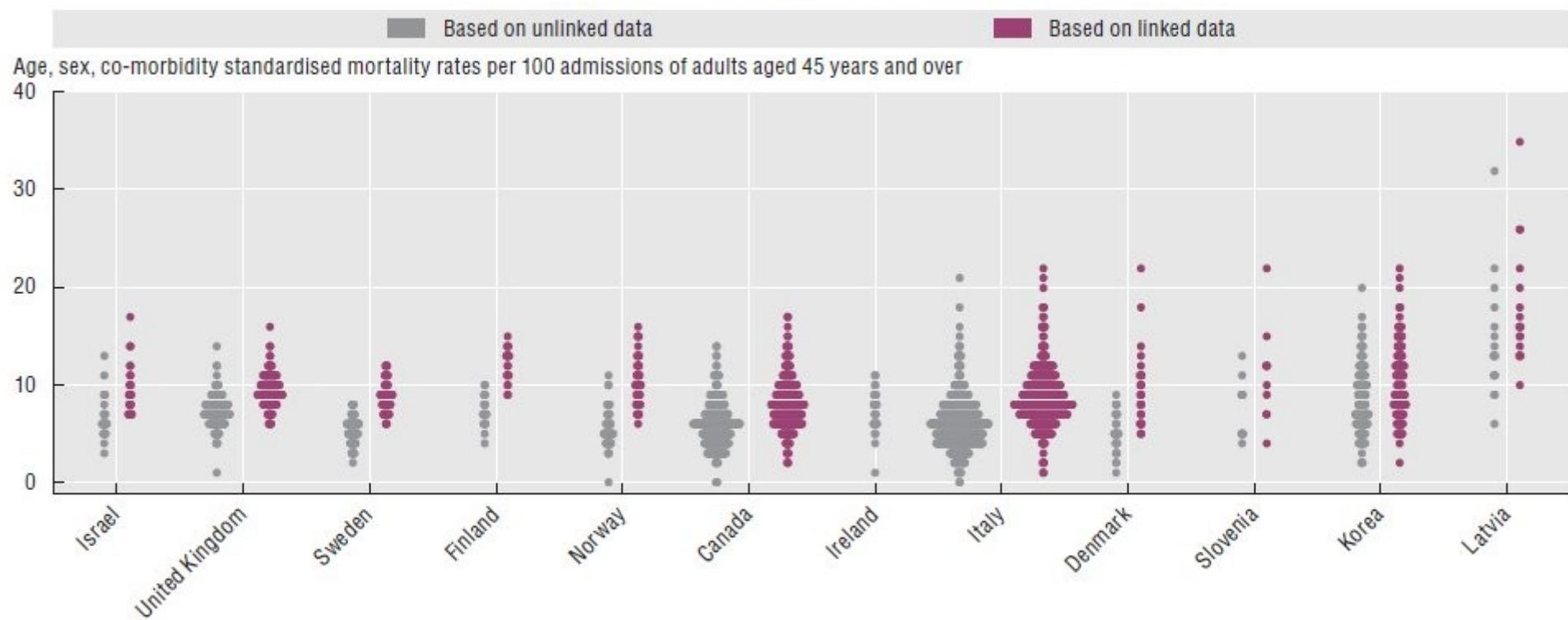


Note: Each dot in the figure represents a single hospital, unless otherwise stated. Results for Canada do not include deaths outside of acute care hospitals. UK data are limited to England and is presented at trust-level (i.e. multiple hospitals).

Source: OECD Hospital Performance Data Collection 2017.

OECD Turnip chart

6.21. Thirty-day mortality after admission to hospital for AMI based on linked and unlinked data, 2013-2015 (or nearest years)



Note: The width of each line in the figure represents the number of hospitals (frequency) with the corresponding rate. Data for Canada not linked to death statistics. UK data are limited to England and presented at trust level (i.e. multiple hospitals). Ordered by inter quartile range of admission-based data. Rates based on linked data are also standardised for previous AMI.

Source: OECD Hospital Performance Data Collection 2017.

Strategies for model building

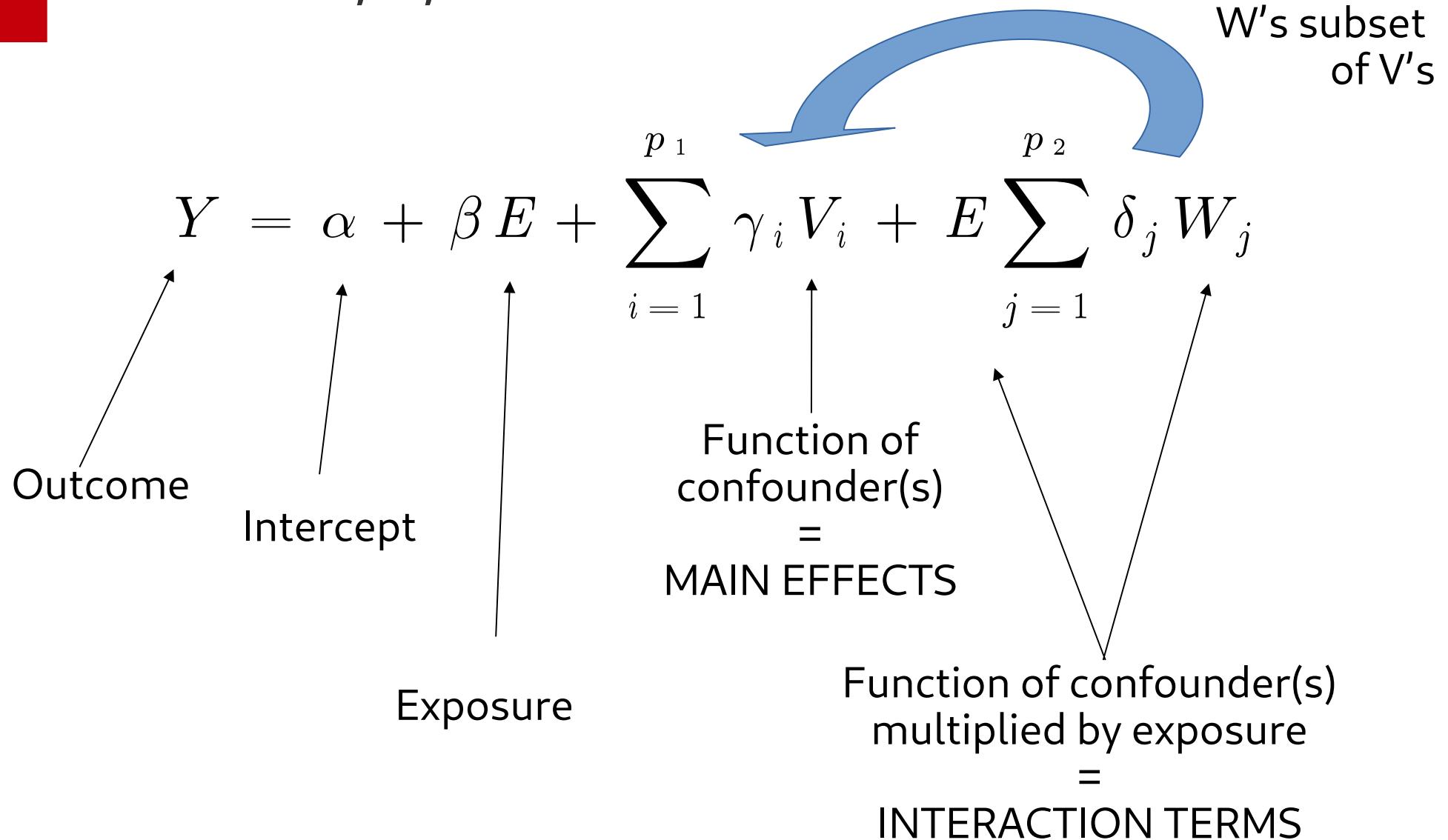
Y dependent variable = D or Y

E study factor, risk factor, exposure

C extraneous variable , confounder => $f(C) \Rightarrow V$

$E^* \text{subset}(V)$ effect modifier, interaction term => E^*W

General E, V, W model



Logistic regression: adjusted OR for binary exposure

$$\text{Adj OR}_{(E=1 / E=0)} = \exp(\beta + \sum_{j=1}^{p_2} \delta_j W_j)$$

Example:

$$\begin{aligned} \logit P(X) &= \alpha + \beta C A T + \\ &\quad \gamma_1 A G E + \gamma_2 C H L + \gamma_3 S M K + \gamma_4 E C G + \gamma_5 H P T + \\ &\quad C A T (\delta_1 C H L + \delta_2 H P T) \end{aligned}$$

$$\text{Adj OR}_{(C A T = 1 / C A T = 0)} = \exp(\beta + \delta_1 C H L + \delta_2 H P T)$$

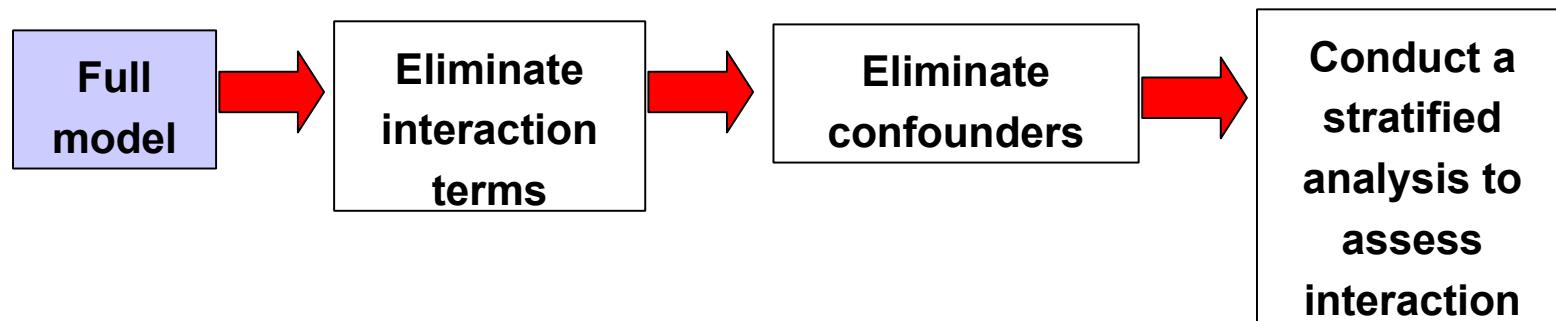
Hierarchically Well-Formulated Models (HWF)

- The investigator must ensure that the model has a certain structure to avoid misleading results
- **Well-Formulated Model:** *given any variable in the model, all lower-order components of the variable are also contained in the model*
- **Hierarchy Principle:** if a product variable is retained in the model, then all lower-order components of that variable must be retained in the model
- Main justification: **if the model is HWF, results (particularly of higher order interactions) will be independent from the coding of the variables in the model**

Backward elimination strategy

- Statistical significance of the interaction term
- Hierarchical backward elimination strategy (Bishop et al. 1975)
- If an interaction effect is retained in a model, than this precludes eliminating from that model any variable constituting that effect
- The validity of tests about lower-order terms depends on the outcome of previous tests of higher order terms

LR chain



Key messages

- Although many of the OECD indicators originate from AHRQ definitions, the standardization methods are different:
 - Standardization adopted by the AHRQ is **indirect**. It can be used to **compare results of each hospital vs a national average, but not against other hospitals**
 - Standardization adopted by the OECD for national estimates are normally based a **direct** standardization method, **making the comparison of results between countries possible**
- A well formulated modelling strategy shall follow fundamental rules that take into account the formal testing of effects of exposure, confounders and interactions within a hierarchical structure
- Despite of the relevant methodological criteria that can lay the ground for common guidelines, model building also relies a lot on the knowledge of the phenomenon and the level of experience and intuition of the investigator

Materials

- AHRQ Indirect Standardization method (technical guide)
- Dover DC et al, Using funnel plots in public health surveillance, Population Health Metrics 2011, 9:58
- Marshall EC et al, Reliability of league tables of in vitro fertilisation clinics: retrospective analysis of live birth rates, BMJ 1998;316:1701–5
- Kleinbaum, Logistic Regression. A self learning text. 3rd Edition. Modelling strategy guidelines, p.165-192



ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Second cycle degree programme (LM) in Statistical Sciences

85278 – Methods and tools for Health Statistics

**85277 – Methods and tools for Official Statistics: Population
and Health Statistics**

a.a. 2023/2023

Stream 2. Risk stratification and standardization

Topic 2.1.3

Model building: GEE Models and hierarchical formulation

Fabrizio Carinci

fabrizio.carinci@unibo.it

Tuesday, 27th February 2023

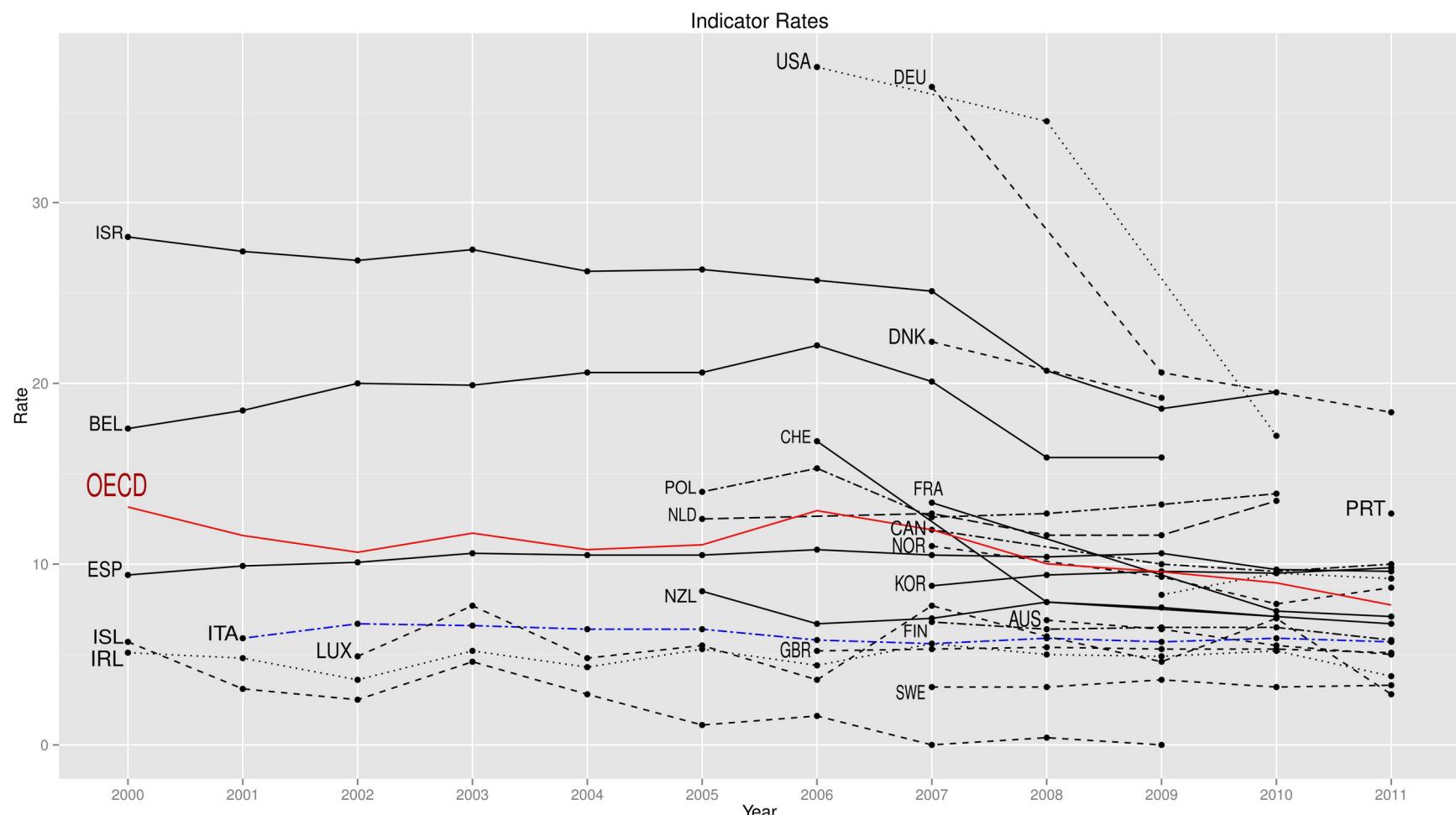
Study Case 2. GEE Model

Lower extremity amputation rates in people with diabetes as an indicator of health systems performance. A critical appraisal of the data collection 2000–2011 by the Organization for Economic Cooperation and Development (OECD)

F. Carinci¹ · M. Massi Benedetti² · N. S. Klazinga^{3,4} · L. Uccioli⁵

26 OECD countries (<http://stats.oecd.org/>)

AUS BEL CAN CHE DEU DNK ESP FIN FRA
GBR HUN IRL ISL ISR ITA KOR LUX MEX
NLD NOR NZL POL PRT SLV SWE USA



Lower extremity amputation rates in people with diabetes as an indicator of health systems performance. A critical appraisal of the data collection 2000–2011 by the Organization for Economic Cooperation and Development (OECD)

F. Carinci¹ · M. Massi Benedetti² · N. S. Klazinga^{3,4} · L. Uccioli⁵

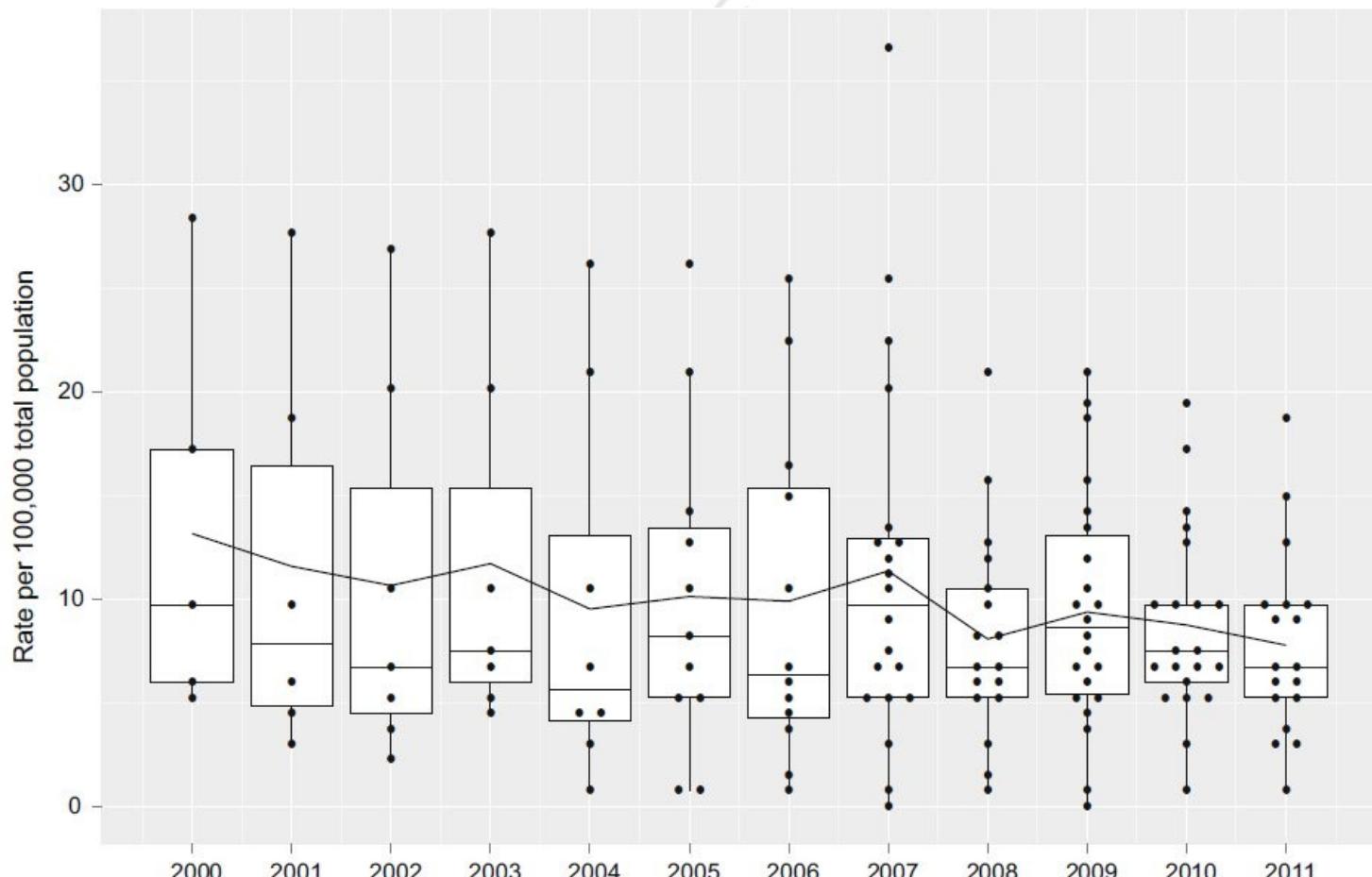
Table 1 Lower extremity amputation rates in diabetes: age-sex standardized rates per 100,000 population aged 15 or over, according to OECD definitions, 2000–2011. *Source* OECD health care quality indicators project (revised version, data collection 2013)

OECD Country	T	R	C	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011
Australia	Y	N	2									6.9	6.4	5.5	5.0
Belgium	N	N	1	17.5	18.5	20.0	19.9	20.6	20.6	22.1	20.1	15.9	15.9		
Canada	Y	N	2									11.9		10.0	9.6
Denmark	Y	Y	2									22.3		19.2	
Finland	Y	Y	3									6.8	6.4	6.5	6.5
France	N	N	3									13.4		7.4	7.1
Germany	N	N	2									36.4		20.6	18.4
Hungary	N	N	2					0.6	0.7	0.8	1.1	1.5	0.7	1.0	1.1
Iceland	Y	Y	3	5.7	3.1	2.5	4.6	2.8	1.1	1.6	0.0	0.4	0.0		
Ireland	Y	N	2	5.1	4.8	3.6	5.2	4.3	5.3	4.4	5.6	5.0	4.9	5.2	3.8
Israel	N	N	1	28.1	27.3	26.8	27.4	26.2	26.3	25.7	25.1	20.7	18.6	19.5	
Italy	Y	N	1		5.9	6.7	6.6	6.4	6.4	5.8	5.6	5.9	5.7	5.9	5.7
Korea	N	N	3									8.8	9.4	9.6	9.5
Luxembourg	N	N	3			4.9	7.7	4.8	5.5	3.6	7.7	6.0	4.6	7.0	2.8

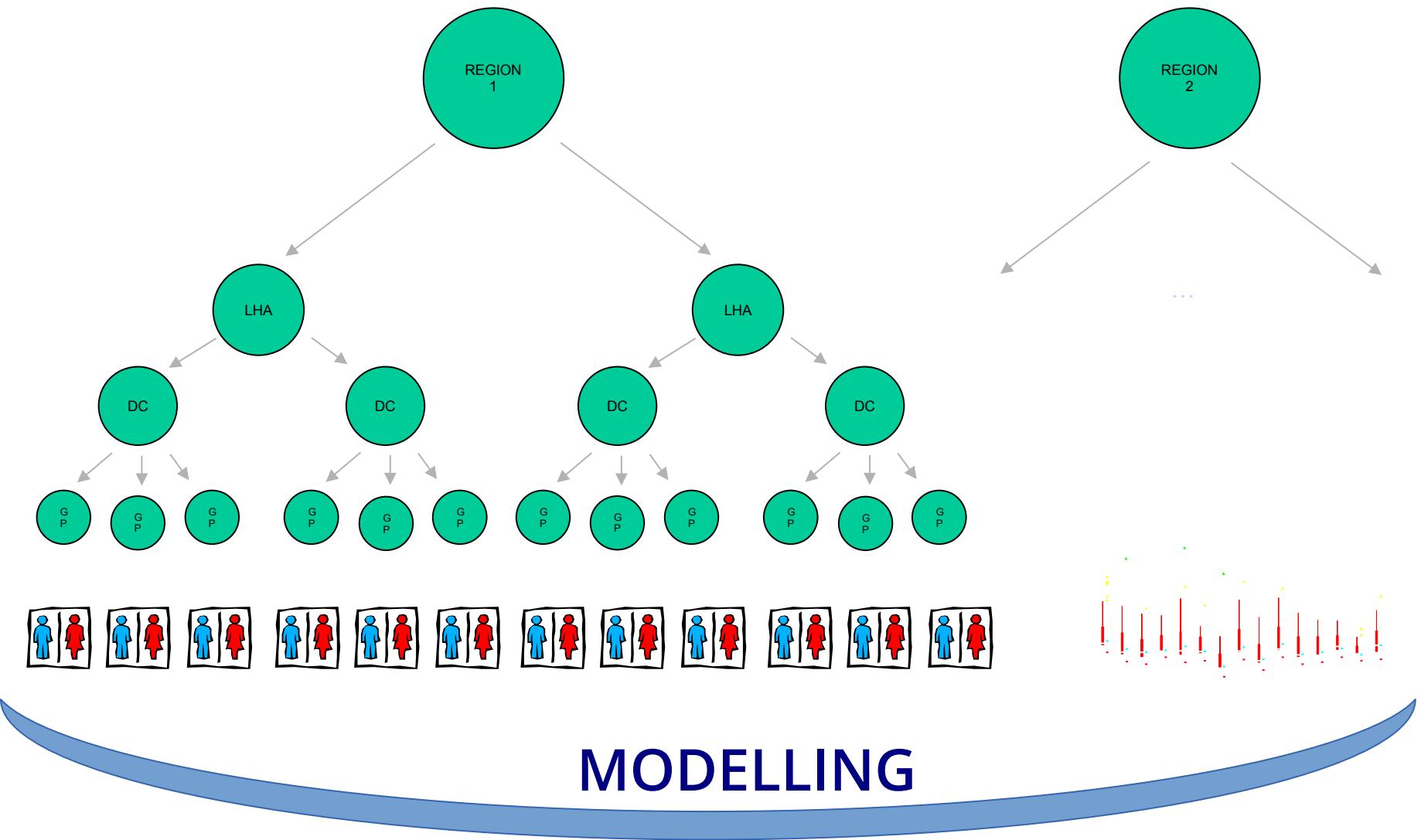
Lower extremity amputation rates in people with diabetes as an indicator of health systems performance. A critical appraisal of the data collection 2000–2011 by the Organization for Economic Cooperation and Development (OECD)

F. Carinci¹ · M. Massi Benedetti² · N. S. Klazinga^{3,4} · L. Uccioli⁵

Received: 15 January 2016



Clustered effects in health studies

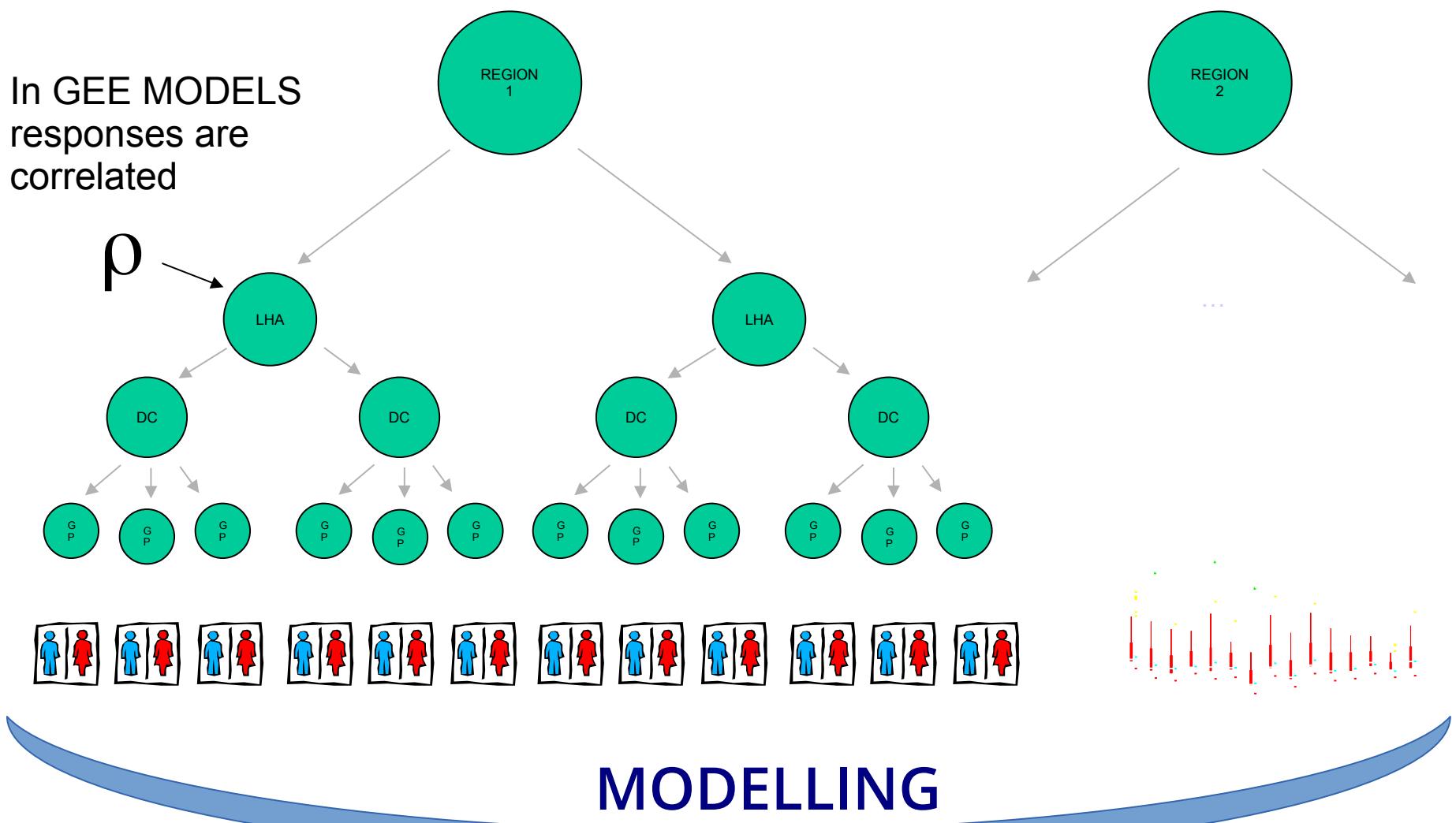


Generalized Estimating Equations

- Generalized Estimating Equation (GEE) is a general statistical approach for longitudinal/clustered data analysis applied in clinical trials and biomedical studies.
- In cases of repeated measurements e.g. HbA1c in the same subject or events occurring in the same hospitals, measures **are likely to be correlated**. This is a **violation from the assumption of independence** that can bias results, most often resulting in confidence intervals that are smaller and less conservative
- In these situations, two types of approaches are frequently used:
 - Generalised Estimating Equations (GEE): population-level approach using a quasi-likelihood function to estimate population-averaged estimates of the parameters.
 - Mixed-effect models: individual-level approach using random effects to capture the correlation between observations on the same subject (cluster).
- GEE is an extension of the Generalized Linear Model (GLM) and as such it can incorporate traditional multivariate regression eg linear and logistic. It adds a correlation structure that can be estimated in various ways. The interpretation of the results is similar to GLMs.

Clustered effects

In GEE MODELS
responses are
correlated



Generalized Estimating Equations

- In GEE, the variance-covariance matrix of responses is treated as a “nuisance parameter”. Therefore, this model fitting turns out to be easier than mixed-effect models. If the overall treatment effect is of primary interest, it might be preferable and also easier to interpret.
- Another advantage of GEE models is that they allow allocating *varying patterns of missing data* in the regression models in a flexible manner.
- Under mild regularity conditions, the parameter estimates are consistent and asymptotically normally distributed even when the “*working*” correlation structure of responses is misspecified, and the variance-covariance matrix can be estimated by a robust estimator that is also known (by its formula) as the “*sandwich variance estimator*”.
- GEE relaxes the distribution assumption and only requires the correct specification of marginal mean and variance as well as the link function which connects the covariates of interest and the marginal means.

Generalized Linear Models

Y follows a distribution from the exponential family (*random component*)
Response (marginal) $\mu_{ij} = E(y_{ij})$ linear combination of the covariates
(*systematic component*)

$$g(\mu_{ij}) = X_{ij}^t \beta$$

y_{ij} response for subject i at cluster j, $j = 1, 2, \dots, J$ X_{ij} $p \times 1$ vector of covariates

β $p \times 1$ vector of regression coefficients $g(\cdot)$ link function

Independent responses allow to express the likelihood as the product of each observation's contribution (L_i). The maximum likelihood estimator for GLM is the solution of the system of k score equations from n subjects:

$$S_k(\beta) = \sum_{i=1}^n \frac{\partial \mu_i}{\partial \beta_k} Var(Y_i)^{-1}(y_i - \mu_i) = 0$$

Score equations are specified by the mean and variance of the random response (entire distribution not needed).

Quasi-likelihood estimation

- Quasi-likelihood estimation of beta parameters is an extension of the classical model method implying the use of the score function where the **likelihood is not specified**.
- The variance of the response y_i is a function of the mean

$$\text{var}(y_i) = \phi V(\mu_i)$$

ϕ unknown scale parameter $v(\cdot)$ known variance function

LOGISTIC REGRESSION - Binary response: $g(\mu_i) = \log[\mu_i/(1 - \mu_i)]$ "Logit link"; $v(\mu_i) = \mu_i(1 - \mu_i)$; $\phi = 1$

- Suppose that the i -th subject has n_i (correlated) responses.
- If the mean is modeled using a link function $g(\mu)$, the estimator for the GEE model is the solution of the system of "score-like" equations:

$$\sum_{i=1}^K \frac{\partial \mu_i}{\partial \beta} \mathbf{W}_i^{-1} (\mathbf{Y}_i - \mu_i) = 0$$

with \mathbf{D}_i an $(n_i \times n_i)$ diagonal matrix with variance function $V(\mu_i)$ as the j -th element and the covariance matrix \mathbf{W}_i :

$$\mathbf{W}_i = \phi \mathbf{D}_i^{1/2} \mathbf{C}_i \mathbf{D}_i^{1/2}$$

Sandwich estimator and working correlation

- In GEE, the “sandwich” variance estimator is used as a robust estimator of $\text{var}(\beta)$ in GEE models:

$$\left[X^T \hat{W} X \right]^{-1} \left[\sum_i X_i^T (y_i - \hat{\mu}_i) (y_i - \hat{\mu}_i)^T X_i \right] \left[X^T \hat{W} X \right]^{-1}$$

- As a result, the GEE method requires specification of:
 - 1)Link function
 - 2)Variance function
 - 3)"Working Correlation Matrix":

$$C_i = R_i(\alpha)$$

where α expresses the *within cluster* correlation

Algorithm for fitting GEE Model

- Compute initial estimate of β from ordinary GLM
- Compute working correlation C_i from least square linear regression using residuals/SE
- Compute estimate of covariance matrix W_i
- Use to solve the system of equations and update β
- Iterate until convergence

Generalized Estimating Equations

Working Correlation structures

The form of the working correlation matrix must be specified at the outset.
Here are some of the possible choices:

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

Uncorrelated (Identity)

$$\begin{bmatrix} 1.00 & \alpha & \alpha & \alpha \\ \alpha & 1.00 & \alpha & \alpha \\ \alpha & \alpha & 1.00 & \alpha \\ \alpha & \alpha & \alpha & 1.00 \end{bmatrix}$$

Exchangeable

$$\begin{bmatrix} 1.00 & a_{12} & a_{13} & a_{14} \\ a_{21} & 1.00 & a_{23} & a_{24} \\ a_{31} & a_{32} & 1.00 & a_{34} \\ a_{41} & a_{42} & a_{43} & 1.00 \end{bmatrix}$$

Given (fixed)

$$\begin{bmatrix} 1.00 & \alpha_{12} & \alpha_{13} & \alpha_{14} \\ \alpha_{21} & 1.00 & \alpha_{23} & \alpha_{24} \\ \alpha_{31} & \alpha_{32} & 1.00 & \alpha_{34} \\ \alpha_{41} & \alpha_{42} & \alpha_{43} & 1.00 \end{bmatrix}$$

**Unspecified
(unstructured)**

Generalized Estimating Equations

Working Correlation structures

$$\begin{bmatrix} 1.00 & \alpha & \alpha^2 & 0 \\ \alpha & 1.00 & \alpha & \alpha^2 \\ \alpha^2 & \alpha & 1.00 & \alpha \\ 0 & \alpha^2 & \alpha & 1.00 \end{bmatrix} \quad \textbf{Stationary M-dependent (m=2)}$$

$$R_{jk} = \begin{cases} \alpha^{|t_j - t_k|}, & |t_j - t_k| \leq m \\ 0, & |t_j - t_k| > m \end{cases}$$

$$\begin{bmatrix} 1.00 & \alpha_{12} & \alpha_{13} & 0 \\ \alpha_{21} & 1.00 & \alpha_{23} & \alpha_{24} \\ \alpha_{31} & \alpha_{32} & 1.00 & \alpha_{34} \\ 0 & \alpha_{42} & \alpha_{43} & 1.00 \end{bmatrix} \quad \textbf{Non-stationary M-dependent (m=2)}$$

$$R_{jk} = \begin{cases} \alpha_{jk}, & |t_j - t_k| \leq m \\ 0, & |t_j - t_k| > m \end{cases}$$

$$\begin{bmatrix} 1.00 & \alpha & \alpha^2 & \alpha^3 \\ \alpha & 1.00 & \alpha & \alpha^2 \\ \alpha^2 & \alpha & 1.00 & \alpha \\ \alpha^3 & \alpha^2 & \alpha & 1.00 \end{bmatrix} \quad \textbf{AR-m (m) (m=1)}$$

$$R_{jk} = \alpha^{|t_j - t_k|}$$

Data patterns

- Different patterns, including sparse matrices, may be accepted as input
BUT consequently the range of working correlation structures that could be used may depend upon the selected patterns.

Cluster	Observation Points		
Id	1	2	3
1	X	X	X
2	X	X	X
3	X	X	X

A Clusters have the same size, observations are taken at a common set of observation points.
All forms of correlation are possible.

Cluster	Observation Points		
Id	1	2	3
1	X	X	X
2	X	X	X
3	X	X	

B Clusters have different size, observations are taken at a common set of observation points.
Stationary forms of correlation are NOT allowed.

Cluster	Equally spaced Observation Points		
Id	1	2	3
1		X	X
2	X	X	
3		X	

C Clusters have different size, observations are taken at a common set of *equally spaced* observation points. No missing value *interrupts* the equal spacing.
Given, non stationary and unspecified forms of correlation are NOT allowed.

Cluster	Equally spaced Observation Points		
Id	1	2	3
1		X	X
2	X		X
3		X	

D As the previous form, but some missing value interrupts the equal spacing.
Given, non stationary and unspecified forms of correlation are NOT allowed.
Need a sufficient number of adjacent observations.

Cluster	Observation Points		
Id	1	2	3
1			X
2		X	
3	X	X	X

E There is no structure in the observation points.
An uncorrelated (identity matrix) or exchangeable form of correlation are the only choices available.

Lower extremity amputation rates in people with diabetes as an indicator of health systems performance. A critical appraisal of the data collection 2000–2011 by the Organization for Economic Cooperation and Development (OECD)

F. Carinci¹ · M. Massi Benedetti² · N. S. Klazinga^{3,4} · L. Uccioli⁵

```
modglm<-glm(value~primary_tax+Year,data=retros_final,family=gaussian)

modgee<-geeglm(value~primary_tax+Year,data=retros_final,id=Country,
                  family=gaussian,corstr="exchangeable") # Model 1
summary(modgee)

modgee<-geeglm(value~Year,data=retros_final[retros_final$primary_tax==0,],
                  id=Country,family=gaussian,corstr="exchangeable") # Model 2
summary(modgee)

modgee<-geeglm(value~Year,data=retros_final[retros_final$primary_tax==1,],
                  id=Country,family=gaussian,corstr="exchangeable") # Model 3

summary(modgee)
```

Lower extremity amputation rates in people with diabetes as an indicator of health systems performance. A critical appraisal of the data collection 2000–2011 by the Organization for Economic Cooperation and Development (OECD)

F. Carinci¹ · M. Massi Benedetti² · N. S. Klazinga^{3,4} · L. Ucciali⁵

Initial GEE model output (GLM model)

Call:

```
glm(formula = value ~ primary_tax + Year, family = gaussian,  
     data = retros_final)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-12.272	-3.896	-0.278	3.398	24.471

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	642.292	340.021	1.89	0.061 .
primary_tax	-4.830	1.046	-4.62	8.2e-06 ***
Year	-0.314	0.169	-1.85	0.066 .

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 41.7)

Null deviance: 7376.4 on 153 degrees of freedom

Residual deviance: 6296.8 on 151 degrees of freedom

AIC: 1017

Lower extremity amputation rates in people with diabetes as an indicator of health systems performance. A critical appraisal of the data collection 2000–2011 by the Organization for Economic Cooperation and Development (OECD)

F. Carinci¹ · M. Massi Benedetti² · N. S. Klazinga^{3,4} · L. Uccioli⁵

Final estimates of the GEE Model

Call:

```
geeglm(formula = value ~ primary_tax + Year, family = gaussian,
        data = retros_final, id = Country, corstr = "exchangeable")
```

Coefficients:

	Estimate	Std. err	Wald	Pr(> w)
(Intercept)	573.911	226.025	6.45	0.011 *
primary_tax	-4.114	2.292	3.22	0.073 .
Year	-0.280	0.112	6.23	0.013 *

Signif. codes: 0 '****' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Estimated Scale Parameters:

	Estimate	Std. err
(Intercept)	42.4	12

Correlation: Structure = exchangeable Link = identity

Estimated Correlation Parameters:

	Estimate	Std. err
alpha	1.04	0.111

Number of clusters: 26 Maximum cluster size: 12

ORIGINAL ARTICLE

Lower extremity amputation rates in people with diabetes as an indicator of health systems performance. A critical appraisal of the data collection 2000–2011 by the Organization for Economic Cooperation and Development (OECD)

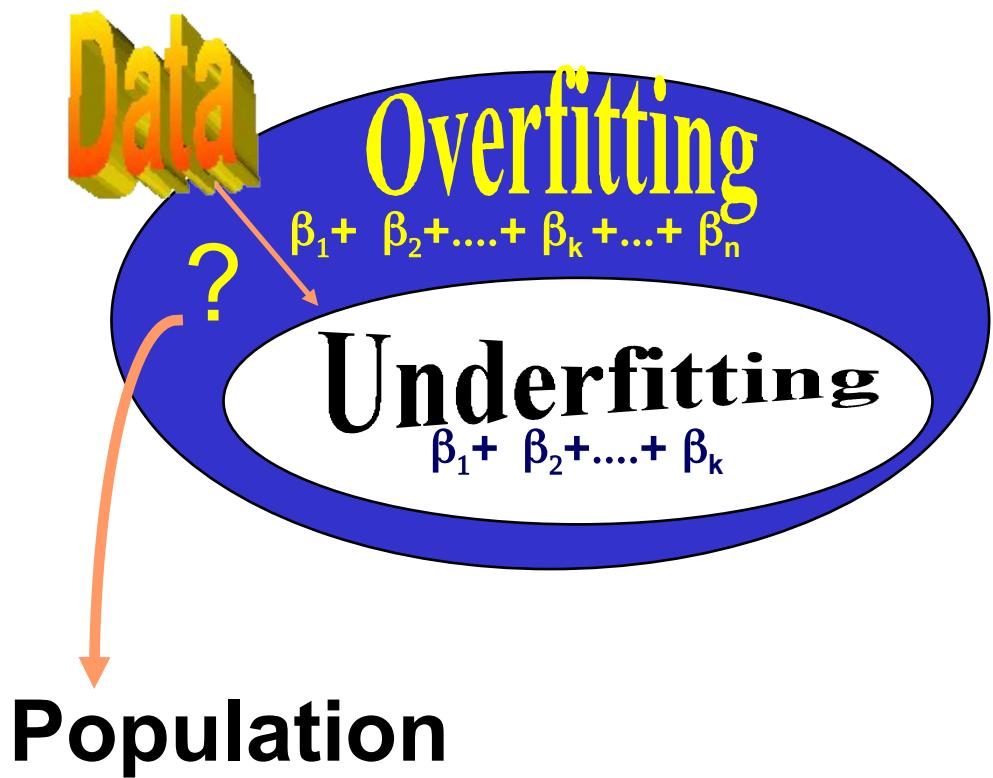
F. Carinci¹ · M. Massi Benedetti² · N. S. Klazinga^{3,4} · L. Uccioli⁵

Table 3 Results of multivariate linear regression (generalized estimating equations), OECD 2000–2011. *Source* OECD health system characteristics survey, 2012; health care quality indicators project (revised version, data collection 2013)

Model/Variable	Estimate	S.E.	95 %C.I.	P > Z
Model 1 [Complete dataset; N countries = 26]				
Tax-based system	-4.55	1.95	-8.38, -0.72	0.020
Use of registry	2.93	2.53	-2.03, 7.89	0.247
Non-ICD coding	-7.04	2.14	-11.24, -2.84	0.001
Average year change	-0.27	0.11	-0.50, -0.05	0.015
Model 2 [Financing: Tax-based; N countries = 12; Median LEARD: 7.55 (2000), 6.25 (2011)]				
Average Year Change	-0.16	0.09	-0.33, 0.01	0.064
Model 2 [Financing: Social insurance; N countries = 14; Median LEARD: 17.50 (2000), 8.15 (2011)]				
Average year change	-0.36	0.18	-0.71, -0.01	0.046

Model Validation

- A statistical model is valid when estimates are stable across multiple samples
- There is a thin line between underfitting and overfitting. The investigator is not always aware that adding more complexity, even with significant covariates, may hamper validity with respect to the whole population
- A frequent problem arise when more variables are added through multiple comparisons. Several solutions have been proposed to make inferential testing more conservative (e.g. Bonferroni method: dividing α by the number of hypotheses tested; or adopting a default $\alpha=0.01$)



How to select the best model among **nested** alternatives?

Well formulated hierarchical model

$$Y = \alpha + \beta E + \sum_{i=1}^{p_1} \gamma_i V_i + E \sum_{j=1}^{p_2} \delta_j W_j$$

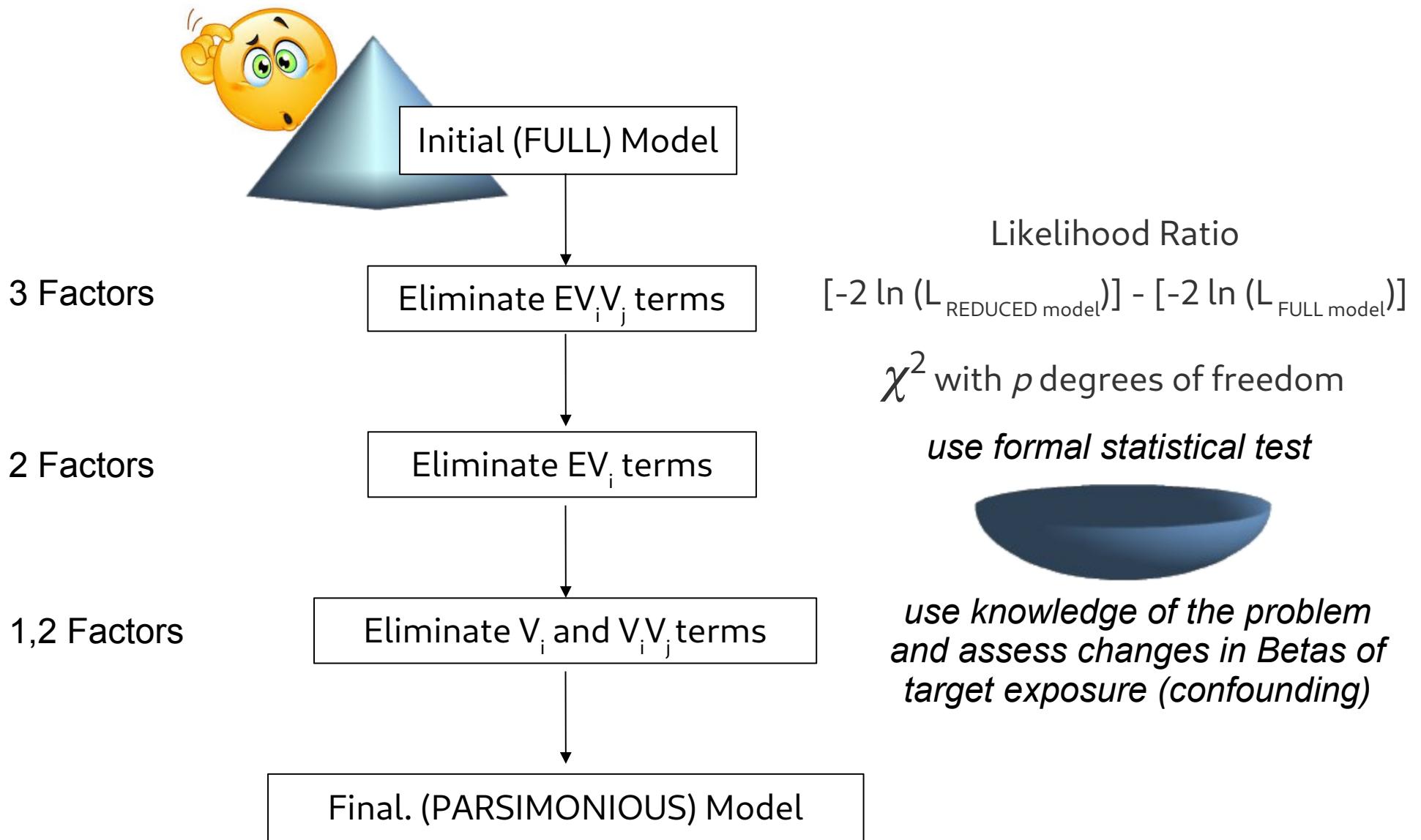
- Keep parameter(s) when the following turns significant

Likelihood Ratio

$$[-2 \ln (L_{\text{REDUCED model}})] - [-2 \ln (L_{\text{FULL model}})]$$

χ^2 with p degrees of freedom

Hierarchical Backward Elimination Approach



Model performance

Once we have a set of predictions, various metrics can be used to control for **overfitting/underfitting** and evaluate performance of the model

A “**confusion matrix**” is a cross-tabulation of the observed and predicted classes. R functions for confusion matrices are in the **caret** package (`confusionMatrix`).

ROC curve functions are found in the **ROCR** package (`performance`) and the **pROC** package (`roc`).

For **Logistic Regression** we can use:

Sensitivity: *given that a result is truly an **event**, what is the probability that the model will predict an event results?*

Specificity: *given that a result is truly a **non event**, what is the probability that the model will predict a negative results?*

These conditional probabilities are directly related to the false positive and false negative rate of a method. Unconditional probabilities (the positive-predictive values and negative-predictive values) require an estimate of what the overall event rate is in the population of interest (prevalence)

Confusion matrix

Logistic Model	True state of OUTCOME	
	Yes	No
Predicted Outcome=Y	Correct	FALSE POSITIVE
Predicted Outcome=N	FALSE NEGATIVE	Correct

Analogue to diagnostic testing

Logistic Model	True state of OUTCOME	
	Yes	No
Predicted Outcome=Y	Correct	FALSE POSITIVE
Predicted Outcome=N	FALSE NEGATIVE	Correct

		DISEASE	
		Present	Absent
TEST	Positive	True Positive	False Positive a
	Negative	False Negative c	True Negative d

Sensitivity

	Outcome Y	Outcome N
Predicted Y	True Positives	False Positives
Predicted N	False Negatives	True Negatives



Proportion of subjects with the Outcome Y
who have been Predicted Y:

$$P(T^+|D^+) = \text{TP} / (\text{TP} + \text{FN})$$

Specificity

	Outcome Y	Outcome N
Predicted Y	True Positives	False Positives
Predicted N	False Negatives	True Negatives



Proportion of subjects with the Outcome N who have been Predicted N:
$$P(T-|D-) = \text{TN} / (\text{TN} + \text{FP})$$

Accuracy

	Outcome Y	Outcome N
Predicted Y	True Positives	False Positives
Predicted N	False Negatives	True Negatives



Proportion of correct Predictions:
 $P([D+|T+] + [D-|T-]) = (TP + TN) / N$

Confusion Matrix from a 50/50 split

```
split=0.50
training_set<-createDataPartition(y=amidata$dead,p=split,list=FALSE)

amidata_train <-amidata[training_set,]

Logit_full_train<-glm(dead~cl_age+males+surgical,family = binomial("logit"),data=amidata_train)
amidata_train$predictions<-predict(Logit_full_train,newdata=amidata_train,type="response")
optimal_cut<cutpointr(data=amidata_train,x=predictions,class=dead,na.rm=TRUE,
method=maximize_metric,metric=youden,use_midpoint=TRUE)
confusionMatrix(as.factor(amidata_train$predictions.value),as.factor(amidata_train$dead))
```

Confusion Matrix and Statistics

Reference		
Prediction	0	1
0	9194	253
1	7437	631
Accuracy : 0.5609		
95% CI : (0.5536, 0.5683)		
No Information Rate : 0.9495		
P-Value [Acc > NIR] : 1		
Kappa : 0.055		
McNemar's Test P-Value : <0.0000000000000002		
		Sensitivity : 0.55282
		Specificity : 0.71380
		Pos Pred Value : 0.97322
		Neg Pred Value : 0.07821
		Prevalence : 0.94953
		Detection Rate : 0.52492
		Detection Prevalence : 0.53937
Balanced Accuracy : 0.63331		
'Positive' Class : 0		

Key messages

- Risk adjustment is required to make fair comparisons among practices and units, whose crude health indicators are confounded by different population characteristics. Complex regression models can be involved in these procedures.
- Complex models can arise from complex data structures and research questions that need to be resolved within frameworks that violate the basic assumptions e.g. correlated responses and time dependent covariates in the GEE model.
- To select the best model, nested regression models can be compared by testing the difference of -2LogL between the reduced and full model. Non nested regression models can use parameters e.g. the AIC.

Materials

- Kleinbaum, Logistic Regression. A self learning text. 3rd Edition.
Logistic Regression for Correlated Data: GEE, p.521-52
- Carinci et al. Lower extremity amputation rates, Acta Diabetologica
2016



ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Second cycle degree programme (LM) in Statistical Sciences

85278 – Methods and tools for Health Statistics

**85277 – Methods and tools for Official Statistics: Population
and Health Statistics**

a.a. 2022/2023

Stream 2. Risk stratification and standardization

Topic 2.2

Model validation: AIC, ROC Analysis, Cross-validation and Bootstrapping

Fabrizio Carinci

fabrizio.carinci@unibo.it

Monday, 6th March 2023

How to select the best model among **non nested** alternatives?

Akaike Information Criterion:

- *for a model with k parameters:*

$$2k - 2\log L$$

calculate AIC value for each model with the same data set, and the “best” model is the one with minimum AIC value

- **Information loss** when fitted model is used instead than the best approximating model:

$$\Delta_i = AIC_i - AIC_{min}$$

Likelihood ratios

- In general, the **Likelihood Ratio** is the probability of a Prediction Y/N among people with Outcome Y, **divided** by the probability of Prediction Y/N among people with Outcome N: $LR = P(T_i|D^+) / P(T_i|D^-)$.
- Two measures are needed to describe a dichotomous prediction: likelihood ratio of a positive prediction and the likelihood ratio of a negative prediction
- The LR+ expresses the change in odds favouring outcome given a positive prediction:

$$LR^+ = \frac{\text{Sensitivity}}{(1 - \text{Specificity})} = \\ \text{True Positive Fraction (TPF) / False Positive Fraction (FPF)}$$

- The LR- expresses the change in odds favouring outcome given a negative prediction:

$$LR^- = \frac{(1 - \text{Sensitivity})}{(\text{Specificity})} = \\ \text{False Negative Fraction (FNF) / True Negative Fraction (TNF)}$$

ROC Curve

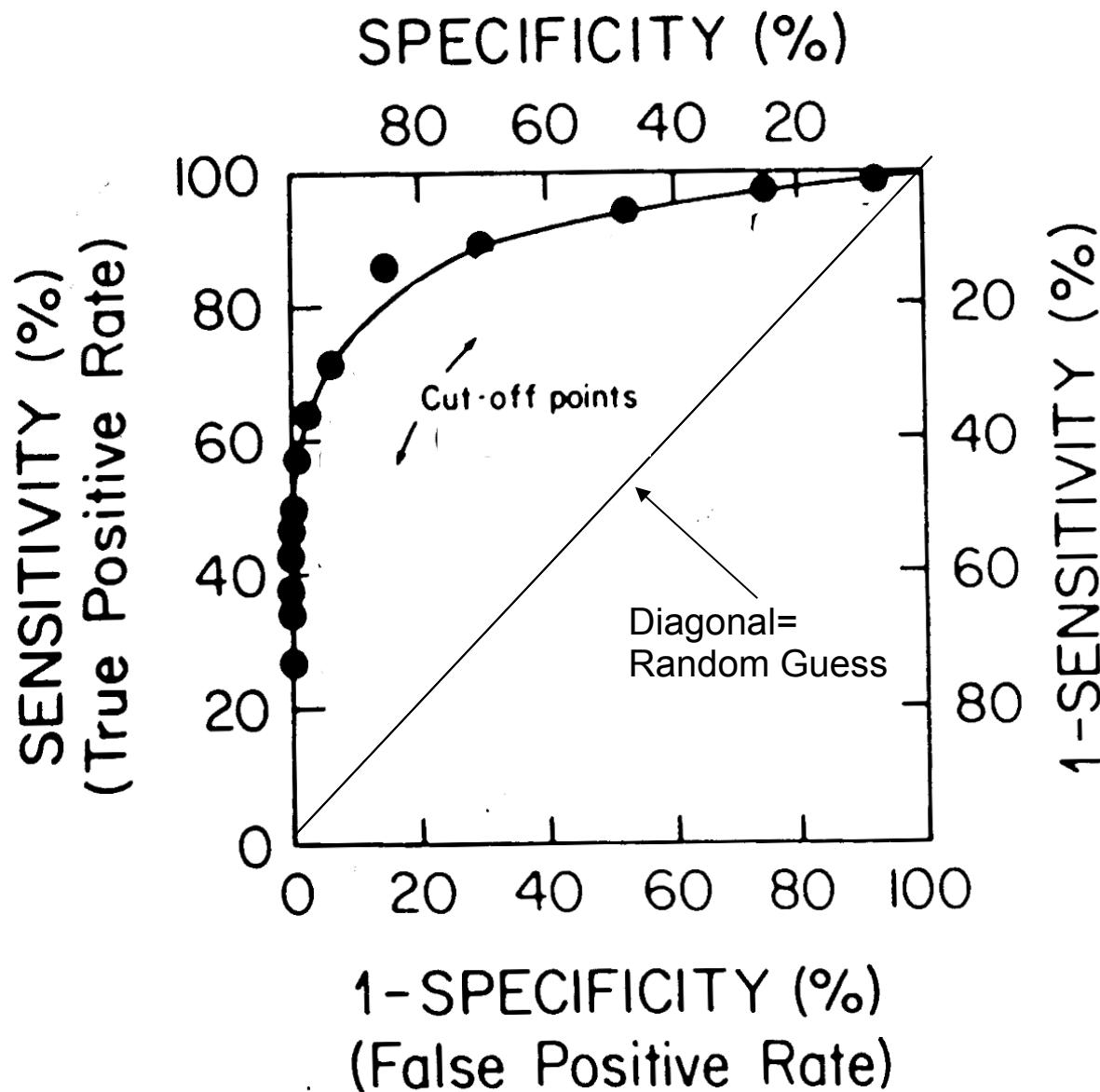
With two classes the **Receiver Operating Characteristic (ROC) curve** can be used to estimate performance using a combination of sensitivity and specificity.

Given the probability of an event, many alternative cutoffs can be evaluated (instead of just a 50% cutoff). **For each cutoff**, we can calculate the sensitivity and specificity.

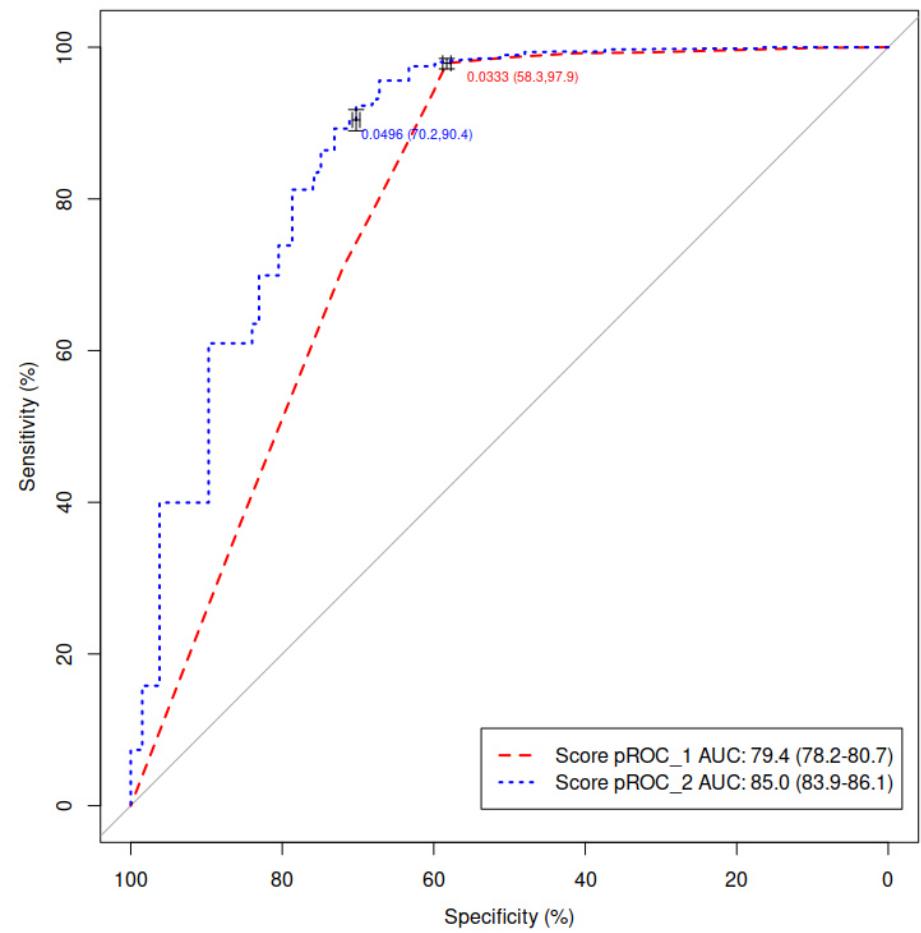
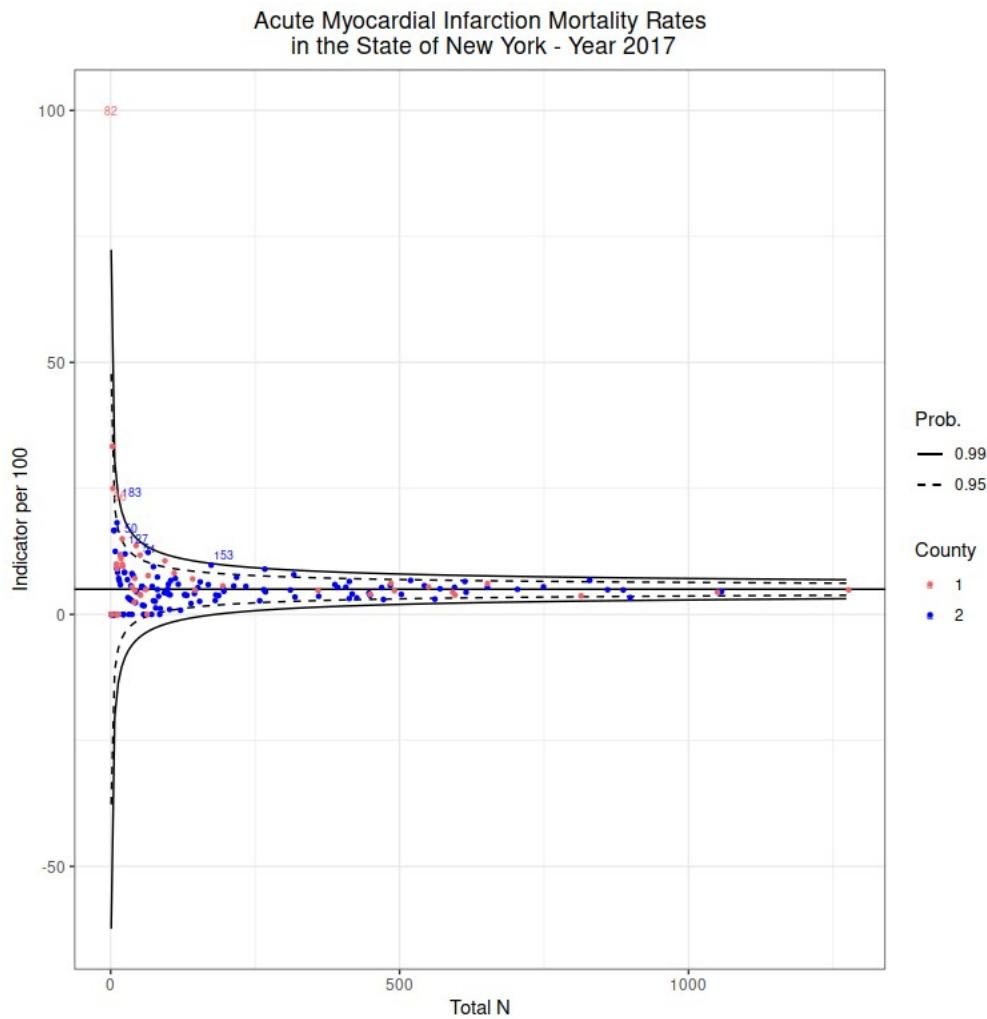
The ROC curve plots the sensitivity (eg. true positive rate) by one minus specificity (eg. the false positive rate).

The **area under the ROC curve (AUC)** is a common metric of performance (equivalent to the **C statistic**).

ROC Curve



Case Study. Hospital Outcome Logistic Model



Example from the NY Hospital Dataset (AMI)

```
Logit_reduced<-glm(dead~cl_age+males,family = binomial("logit"),data=amidata)
```

	Beta	2.5 %	97.5 %	P	OR	2.5 %	97.5 %
(Intercept)	-7.1389	-7.5999	-6.6882	0.0000	0.0008	0.0005	0.0012
cl_age	0.9414	0.8479	1.0367	0.0000	2.5635	2.3347	2.8199
Males	-0.1177	-0.2162	-0.0190	0.0192	0.8889	0.8056	0.9812

Accuracy of the REDUCED model: 0.949985726520126

-2 LogLik REDUCED: 13408.98109676

```
Logit_full<-glm(dead~cl_age+males+surgical,family = binomial("logit"),data=amidata)
```

	Beta	2.5 %	97.5 %	P	OR	2.5 %	97.5 %
(Intercept)	-6.5908	-7.0605	-6.1314	0.0000	0.0014	0.0009	0.0022
cl_age	0.8625	0.7683	0.9585	0.0000	2.3691	2.1561	2.6078
males	-0.0520	-0.1514	0.0476	0.3056	0.9493	0.8595	1.0488
surgical	-0.5049	-0.6079	-0.4027	0.0000	0.6036	0.5445	0.6685

Accuracy of the FULL model: 0.949985726520126

-2 LogLik FULL: 13313.34752524

Likelihood Ratio: 95.63357152 ; P(chi-square)= 0.00000000 ; df= 1

Example from the NY Hospital Dataset (AMI)

Confusion Matrix and Statistics

Reference

Prediction	0	1
0	16631	884
1	0	0

Accuracy : 0.9495

95% CI : (0.9462, 0.9527)

No Information Rate : 0.9495

P-Value [Acc > NIR] : 0.5089

Kappa : 0

McNemar's Test P-Value : <0.0000000000000002

Sensitivity : 1.0000

Specificity : 0.0000

Pos Pred Value : 0.9495

Neg Pred Value : NaN

Prevalence : 0.9495

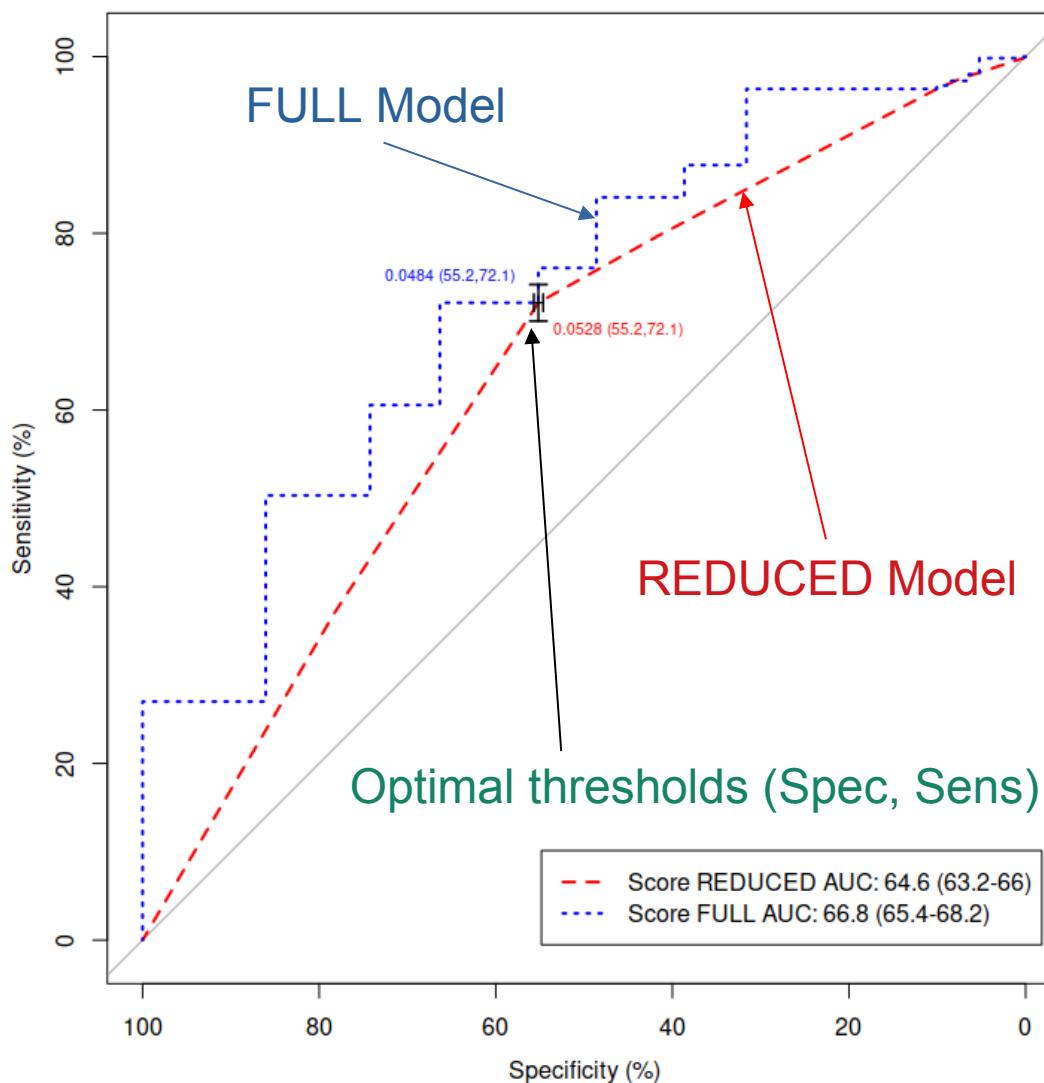
Detection Rate : 0.9495

Detection Prevalence : 1.0000

Balanced Accuracy : 0.5000

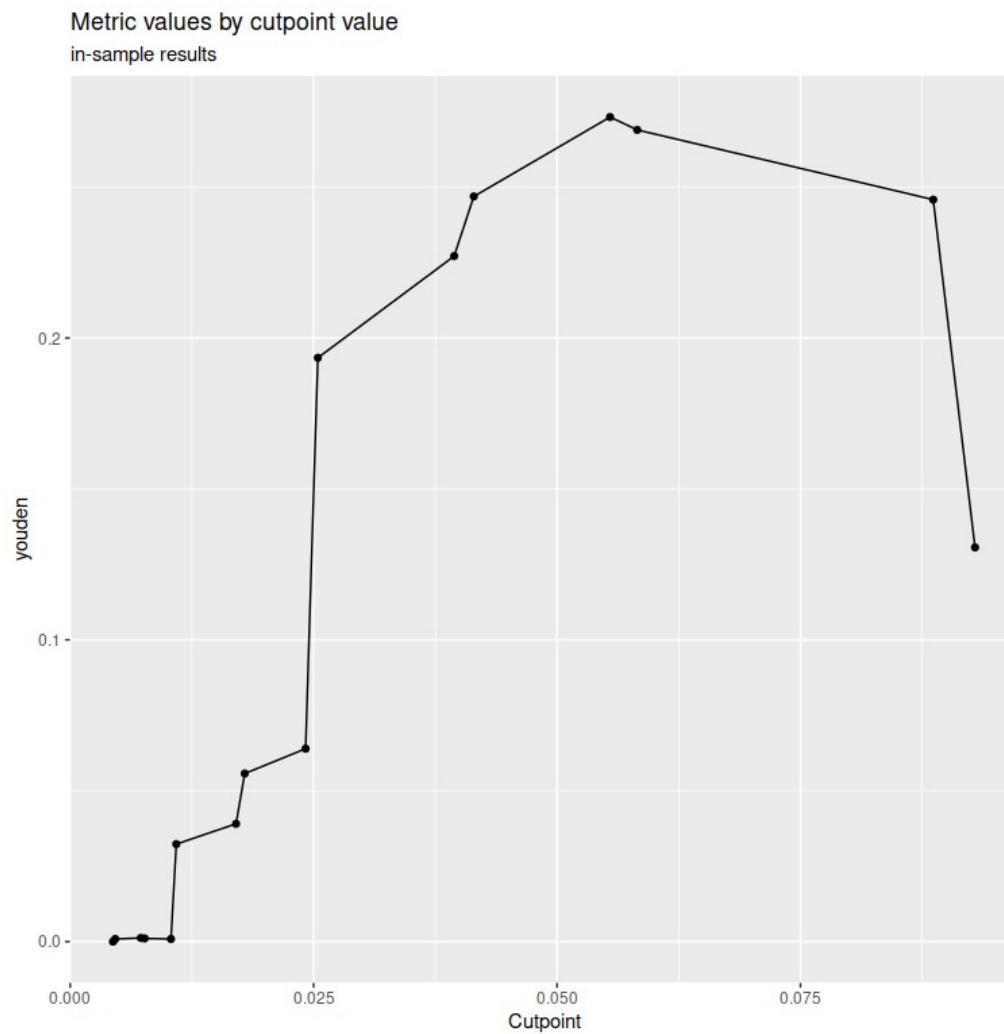
'Positive' Class : 0

Example from the NY Hospital Dataset (AMI)



Selection of optimal cutpoint

Best threshold based on
Youden Index:
Sensitivity+Specificity-1



Caution

- Most packages, including Caret, calculate “metrics” to cross validate model performance based on a fixed threshold set at p=0.5
- That would lead to heavily biased results for logistic regression models, when the outcome is rare (e.g. Acute Myocardial Infarction - AMI). In fact for the NY dataset the optimal threshold is approximately 0.05, ten times lower.
- In those cases, the accuracy would be highest with any model, just by assign each observation to the most frequent category.
- See what happens using 0.5 in AMI:

Confusion Matrix and Statistics

Reference					
Prediction	0	1			
0	16631	884	Sensitivity	:	1.0000
1	0	0	Specificity	:	0.0000
Accuracy :		0.9495	Pos Pred Value	:	0.9495
95% CI :		(0.9462, 0.9527)	Neg Pred Value	:	NaN
No Information Rate :		0.9495	Prevalence	:	0.9495
P-Value [Acc > NIR] :		0.5089	Detection Rate	:	0.9495
Kappa :		0	Detection Prevalence	:	1.0000
McNemar's Test P-Value :		<0.0000000000000002	Balanced Accuracy	:	0.5000
'Positive' Class :		0			

- Balanced accuracy (sens+spec)/2 signals that the accuracy is not relevant here!
- Since best cutpoints may vary across different samples, we will use the **Area under the Curve** (or Method=ROC in Caret) as a measure taking all threshold into account simultaneously.

Cross validation

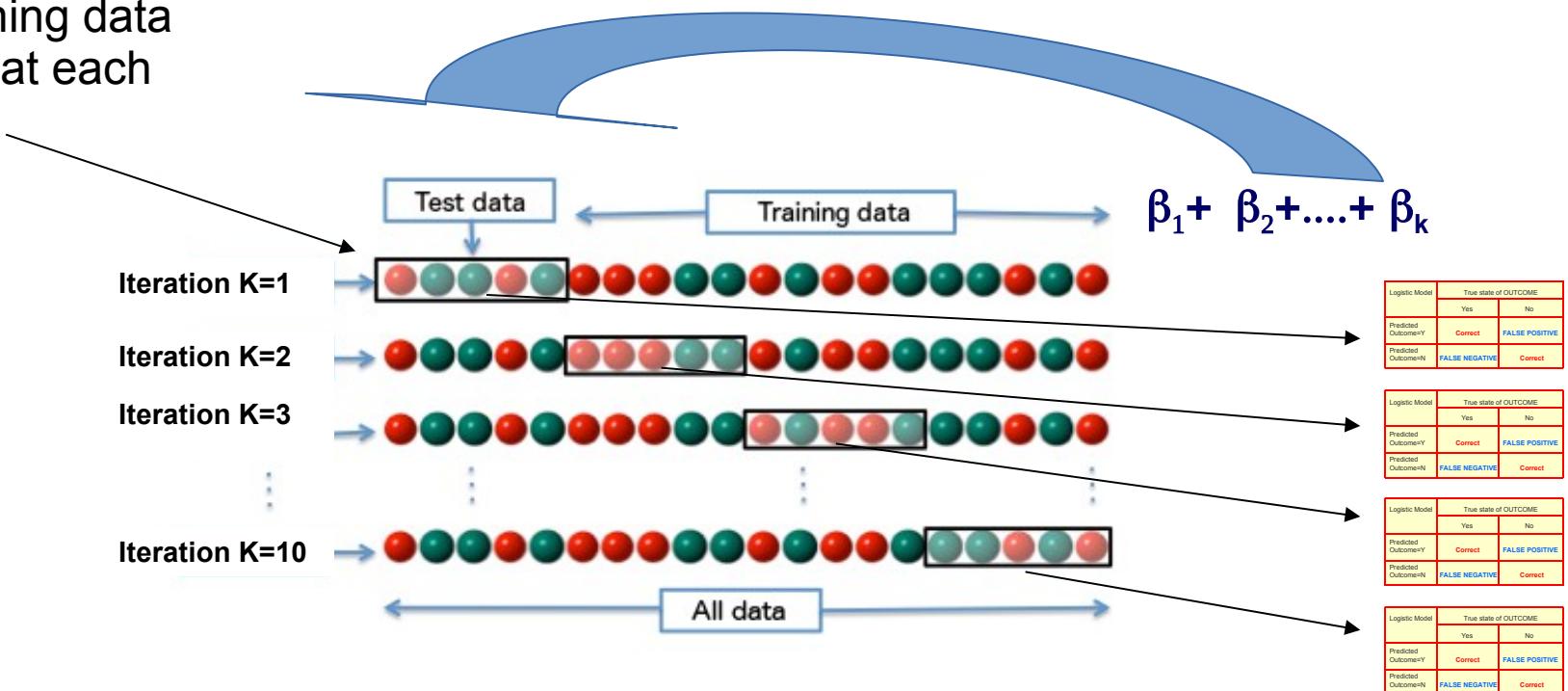
- By definition, any regression model overestimates its performance in the whole dataset, since those routines are designed to minimise prediction error.
- Cross validation is a data intensive technique that may provide useful insight on the stability of a candidate regression model.
- *A model is estimated on a “**training dataset**” and its accuracy is evaluated on a different “**testing set**” using a measure of choice (e.g. **Mean Squared Error** – MSE - of beta or **% of concordant predicted** results in logistic regression).*

Cross validation (2)

- Different validating schemes are possible:
 - *External*: measure average performance under real life conditions. In this way, the validating set is completely independent from the training set generating the model. Normally, there are very practical limitations making this impossible at model formulation. However, this can be assessed in further studies, particularly when prediction can have a high impact on resolving a problem
 - *Internal*: assess performance using subsets of data available:
 - *Leave-p-out*: train the model on $n-p$ observations and validate on p observations. Repeat the process using all possible combinations and use average performance for accuracy.
 - *K-fold*: split the sample in k parts, train on $(k-1)$ and validate on the k -th. Repeat the process k times and average performance over all tests

10-fold cross validation

Presence of 0/1 in
the test/training data
is balanced at each
iteration
(stratified
sampling)



Measures of accuracy can be computed at each iteration and then averaged for the whole process

Sampling

There are a few different ways to do the split: simple random sampling, stratified sampling based on the outcome, by date and methods that focus on the distribution of the predictors.

The base R function `sample` can be used to create a completely random sample of the data.

The `caret` package has a function `createDataPartition` that conducts data splits within groups of the data.

For **logistic regression**, this would mean balancing the sample within the two levels of the outcome to preserve the distribution of the outcome in the training and test sets

For regression, the function determines the quartiles of the outcome and/or predictor variables

Leave One Out Cross-Validation (LOOCV)

- This would require computing a number of models equal to the number of observations (computationally expensive)
- There are convenient properties that allow calculating the average accuracy through the mathematical properties of LOOCV e.g. in linear regression (prediction sum of squares statistic using the average observed value and the diagonal of the Hessian matrix)
- This allows calculating an exact or approximate performance of a model using the LOOCV method directly in one fit performed over the sample dataset

Bootstrap

- The original performance on the whole dataset is labelled as **apparent performance**
- The bootstrap replicates the process of sample generation from an underlying population by drawing **samples with replacement** from the original data set, **of the same size** as the original data set
- Models may be developed in bootstrap samples and tested in the original sample. The **performance in the bootstrap sample** represents an estimate of the *training performance*, and the **performance in the original sample** represents *test performance*

The difference between these performances is an estimate of the optimism in the apparent performance at each round

- This difference is averaged to obtain a stable estimate of the **Optimism= average (bootstrap performance – test performance)**
- **Estimated performance = Apparent performance – Optimism**

Comparing performance measurement methods



Journal of Clinical Epidemiology 54 (2001) 774–781

Journal of
Clinical
Epidemiology

Internal validation of predictive models: Efficiency of some procedures for logistic regression analysis

Ewout W. Steyerberg^{a,*}, Frank E. Harrell Jr^b, Gerard J. J. M. Borsboom^a,
M. J. C. (René) Eijkemans^a, Yvonne Vergouwe^a, J. Dik F. Habbema^a

^aCenter for Clinical Decision Sciences, Ee 2091, Department of Public Health, Erasmus University, P.O. Box 1738, 3000 DR, Rotterdam, The Netherlands

^bDivision of Biostatistics and Epidemiology, Department of Health Evaluation Sciences, University of Virginia, Charlottesville VA, USA

Received 28 June 2000; received in revised form 26 October 2000; accepted 20 December 2000

Abstract

The performance of a predictive model is overestimated when simply determined on the sample of subjects that was used to construct the model. Several internal validation methods are available that aim to provide a more accurate estimate of model performance in new subjects. We evaluated several variants of split-sample, cross-validation and bootstrapping methods with a logistic regression model that included eight predictors for 30-day mortality after an acute myocardial infarction. Random samples with a size between $n = 572$ and $n = 9165$ were drawn from a large data set (GUSTO-I; $n = 40,830$; 2851 deaths) to reflect modeling in data sets with between 5 and 80 events per variable. Independent performance was determined on the remaining subjects. Performance measures included discriminative ability, calibration and overall accuracy. We found that split-sample analyses gave overly pessimistic estimates of performance, with large variability. Cross-validation on 10% of the sample had low bias and low variability, but was not suitable for all performance measures. Internal validity could best be estimated with bootstrapping, which provided stable estimates with low bias. We conclude that split-sample validation is inefficient, and recommend bootstrapping for estimation of internal validity of a predictive logistic regression model.

Comparing performance measurement methods (2)



ELSEVIER

Journal of Clinical Epidemiology 54 (2001) 774–781

**Journal of
Clinical
Epidemiology**

Internal validation of predictive models: Efficiency of some procedures for logistic regression analysis

Ewout W. Steyerberg^{a,*}, Frank E. Harrell Jr^b, Gerard J. J. M. Borsboom^a,
M. J. C. (René) Eijkemans^a, Yvonne Vergouwe^a, J. Dik F. Habbema^a

Table 1

Procedures considered to estimate internal validity of a logistic regression model with eight predictors

Method	Training sample	Test sample	Estimated performance	Repetitions
Apparent	Original	Original	Original sample	1
Split-sample	50%	50% of original	Test	1
	33%	66.67% of original	Test	1
Cross-validation	50%	50% of original	Average(test)	2
	10%	90% of original	Average(test)	10
	10×10%	90% of original	Average(test)	100
Bootstrapping	Regular	Bootstrap	Apparent - average(bootstrap-test)	100 ^a
	.632	Bootstrap	0.368 × Apparent + 0.632 × average(test)	100 ^a
	.632+	Bootstrap	(1-w) × Apparent + w × average(test) ^b	100 ^a

^a100 bootstrap samples were drawn for EPV 5, 10 or 20, while 50 samples were used for EPV 40 or 80.

^bThe weight w was calculated as: $w = .632 / (1 - .368 \times R)$, with $R = (\text{test performance} - \text{apparent performance}) / (\text{"no information" performance} - \text{apparent performance})$ [8] (see text).

Validation tools (1)

```
# 10-fold Cross-validation

train(make.names(dead) ~ cl_age+males+surgical, data=amidata, method="glm", family="binomial",
  trControl=trainControl(method="repeatedcv", number=10, repeats=10,
  classProbs = TRUE, summaryFunction = twoClassSummary),
  metric="ROC", na.action=na.exclude)
```

Generalized Linear Model

```
35032 samples
 3 predictor
 2 classes: 'X0', 'X1'
```

```
No pre-processing
Resampling: Cross-Validated (10 fold, repeated 10 times)
Summary of sample sizes: 31527, 31527, 31527, 31527, 31527, 31527, ...
Resampling results:
```

ROC	Sens	Spec
0.6682245	1	0

Validation tools (2)

```
# Bootstrap
train(make.names(dead) ~ cl_age+males+surgical, data=amidata, method="glm", family="binomial",
      trControl=trainControl(method="boot", number=100, classProbs = TRUE,
      summaryFunction = twoClassSummary), metric="ROC", na.action=na.exclude)
```

```
Generalized Linear Model
35032 samples
 3 predictor
 2 classes: 'X0', 'X1'
```

```
Resampling: Bootstrapped (100 reps)
Resampling results:
```

ROC	Sens	Spec
0.6673585	1	0

```
# Bootstrap .632
train(make.names(dead) ~ cl_age+males+surgical, data=amidata, method="glm", family="binomial",
      trControl=trainControl(method="boot632", number=100, classProbs = TRUE,
      summaryFunction = twoClassSummary), metric="ROC", na.action=na.exclude)
```

```
Generalized Linear Model
35032 samples
 3 predictor
 2 classes: 'X0', 'X1'
```

```
Resampling: Bootstrapped (100 reps)
Resampling results:
```

ROC	Sens	Spec
0.6673213	1	0

Key messages

- By definition, any regression model overestimates its performance in the whole dataset, since those routines are designed to minimise prediction error. Techniques e.g. cross validation and the bootstrap can be used to provide a more realistic guess of the true accuracy
- Advanced modeling is required for risk adjustment and for the standardization of health care quality indicators
- Practical issues arise with the management of statistical routines to perform various types of techniques involving large datasets. Knowledge of statistical programming e.g. R specialised routines is essential to deliver the range of results expected by health statisticians.

Materials

- E.Steyerberg et al. Internal validation of predictive models: Efficiency of some procedures for logistic regression analysis, *Journal of Clinical Epidemiology* 54 (2001) 774–781



ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Second cycle degree programme (LM) in Statistical Sciences

85278 – Methods and tools for Health Statistics

**85277 – Methods and tools for Official Statistics: Population
and Health Statistics**

a.a. 2022/2023

Stream 3. Statistical methods and tools to plan and to evaluate health policies

Topic 3.1.1

Study design and theory of propensity scores

Fabrizio Carinci

fabrizio.carinci@unibo.it

Tuesday, 6th March 2023

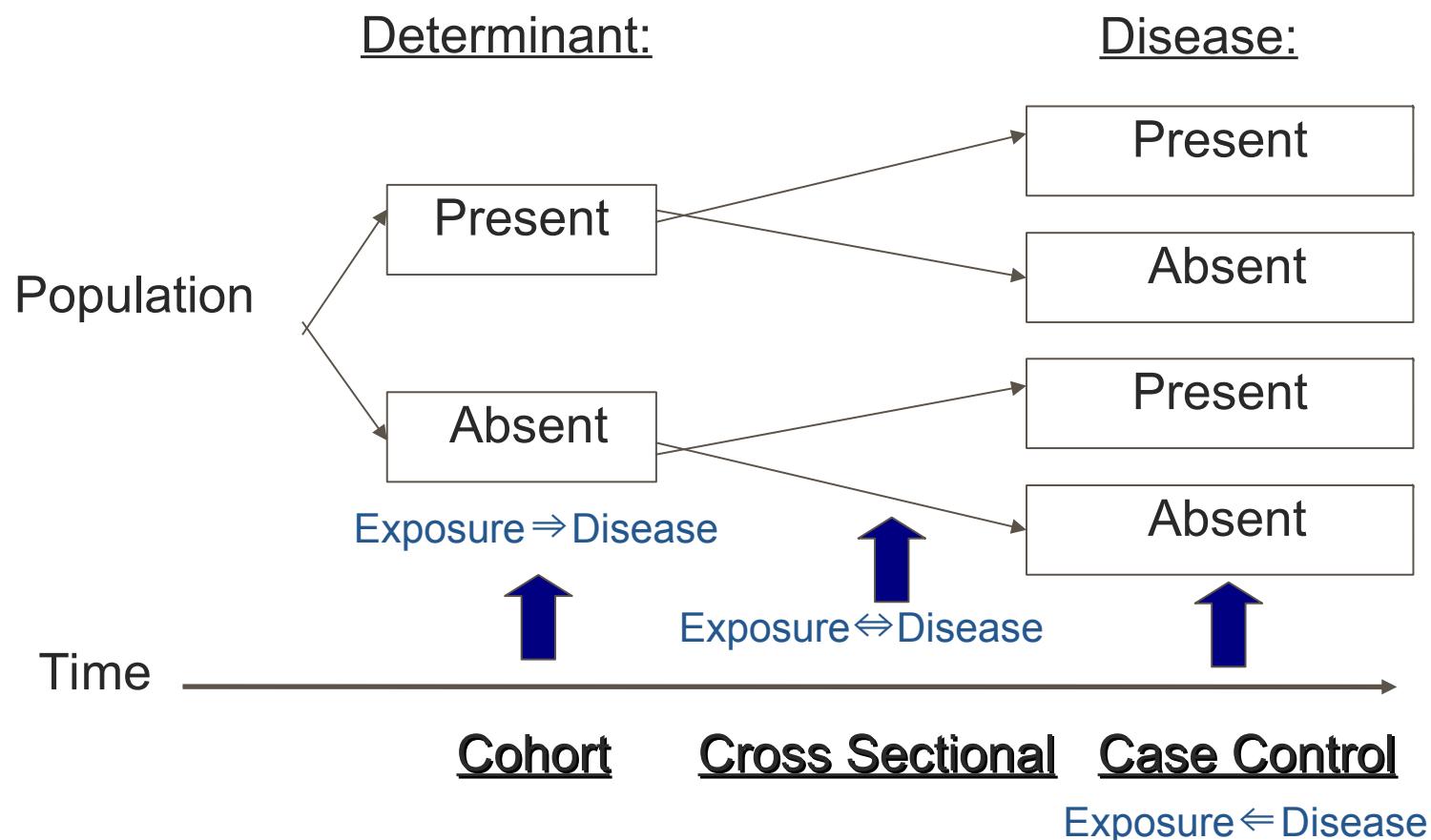
Sources of bias

- Two types of error may affect study designs for policy evaluation: random error and systematic error.
- Random error is strictly related to the statistical analysis and can be defined as the variability in the data that is not explained. It concerns the residual variability.
- Systematic error is also called **Bias** and can depend on different factors like the way in which the individual has been selected, the way in which the study variables have been measured, or some confounding factor that is not completely controlled.
- Random error will tend to 0 if the study size will increase infinitely. It is then related to the sample variability. On the other hand, systematic error does not decrease.

Types of studies

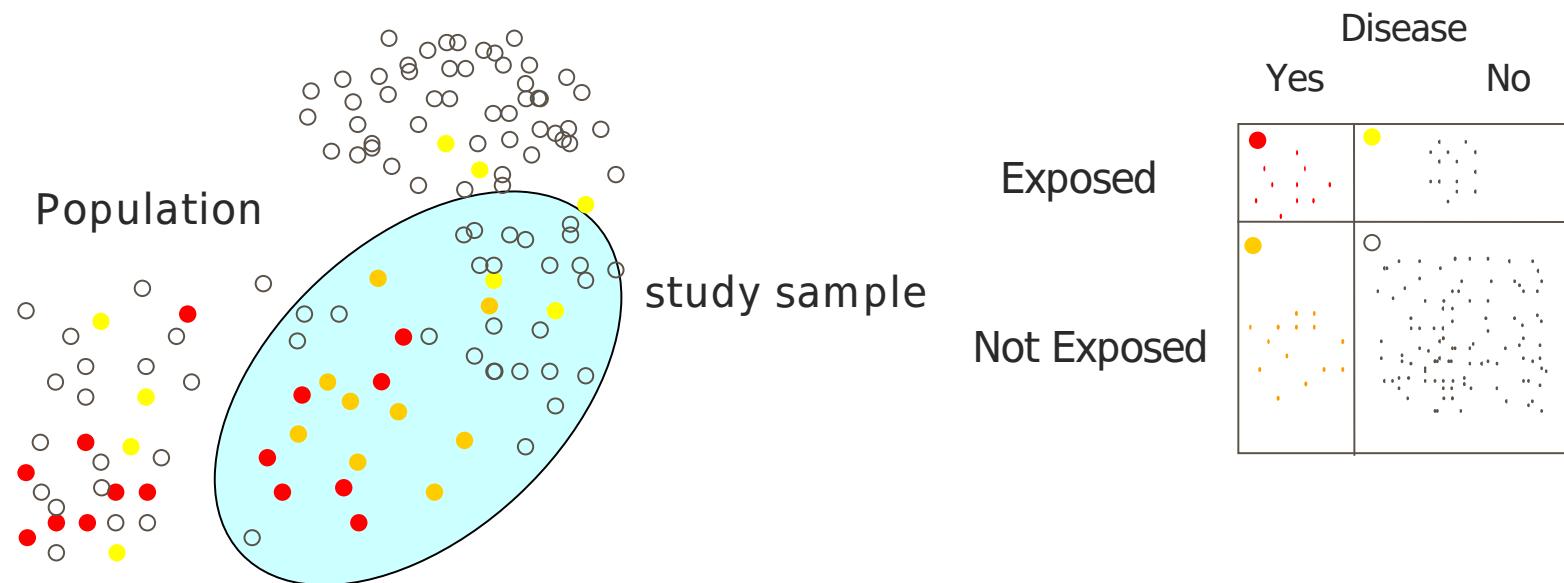
- **Observational:** Studies in which the researcher observes the trends of phenomena and draws conclusions
 - *Ecological (descriptive):* the unit of analysis is a population/group of subjects. Data collected from routine data
 - *Transversal (cross sectional):* used for measuring the prevalence of a disease/condition or association between a risk factor and the outcome in a population or representative sample
 - *Cohort:* groups of subjects exposed/unexposed to one or more risk factors are followed over time to evaluate the incidence of a condition
 - *Case-control:* evaluate the exposure to one or more risk factors in two different groups e.g. those affected by a disease (case) vs not affected (control).
- **Experimental:** field, community or controlled trials (preventive, therapeutic, active interventions, etc). In this studies the researcher directly intervenes on the study variables, by controlling the conditions in the design and applying specific strategies e.g. selected interventions.

Observational studies: the role of time



Cohort studies (prospective/retrospective)

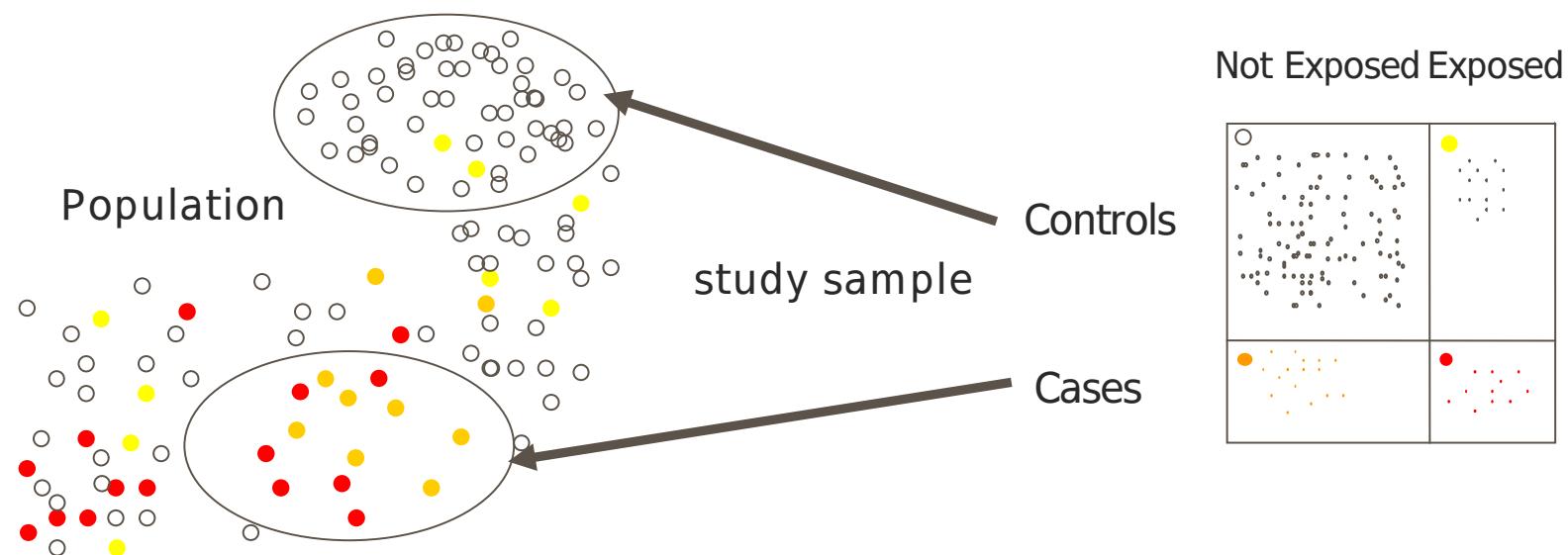
- A group of individuals sharing a common experience, who are disease-free, defined at a point in time, followed up over a period of time to identify the subsequent appearance of the disease (incidence) in those with or without the exposure



Case-control studies

The frequency of the determinant in a group of persons with a disease is compared to the expected frequency of the determinant in the population from which cases originate

- *Cases*: equal chance of being selected; sources: facility; registries; community
 - *Controls*: representative sample of the population from which cases arise; sources: living same locality or same work place as cases, population registries, hospitals (as cases); community (random dialling; relatives)



Experimental

A study in which a population is selected for a planned trial of an intervention whose effects are measured by comparing the outcome in the experimental group with the outcome of a conventional intervention or placebo (no active intervention) in a control group

		Treatment Success	
		Yes	No
Determinant (Intervention)	Experimental		
	Standard		

RCTs vs Observational studies

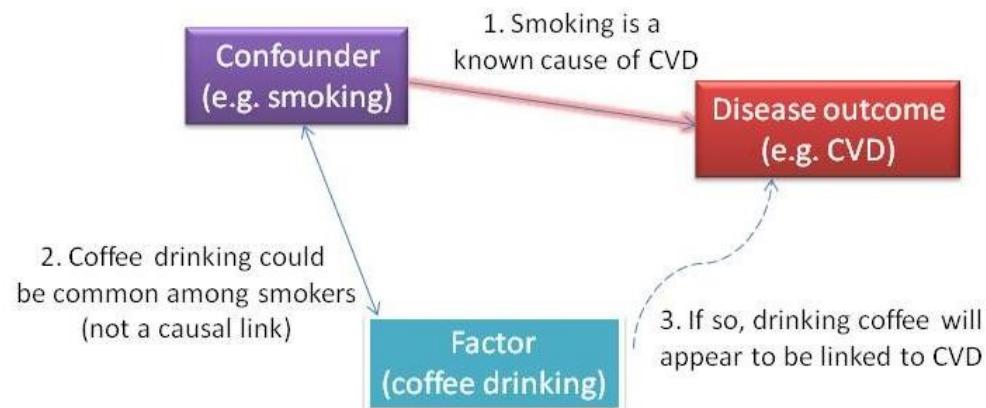
- In randomized controlled trials (RCTs), **random treatment allocation** ensures that treatment status will not be confounded with either measured or unmeasured baseline characteristics
- In observational studies, treatment selection is often influenced by subject characteristics (**potential confounders**)
- Analysts must account for systematic differences between treated and untreated subjects when estimating the effect of treatment on outcomes

Confounding

CONFUSION OF EFFECTS. Confounding occurs when a potential risk factor is associated both with the disease and the factor under study. Such a “confounding” role can make an exposure appear as a potential risk, while it **is not**.

Example: can coffee cause acute myocardial infarction? If drinking coffee is strongly associated with smoking, the latter can be a confounding variable that can make coffee falsely appear as a risk factor for AMI.

In this case we could say that “...the relation between coffee and acute myocardial infarction is confounded by smoking” or “...smoking is a confounder of the relation between coffee and AMI”



Conditions for confounding

For a variable to be a confounder, it must have three necessary (but not sufficient) characteristics:

In order for a factor to be a confounder,

1. *A confounding factor must be associated with the disease (either as a cause or as a proxy for a cause)*
2. *A confounding factor must be associated with the exposure under study in the population at risk*
3. *A confounding factor must not be an effect of the exposure (Example: smoking=> hypertension=>AMI; hypertension cannot be a confounder for smoking, is an **intermediate factor**)*

Example of confounding

Do seatbelts reduce crash injuries?

The *counterfactual* model says that we should compare injury rates among people wearing seatbelts to injury rates for the same people at the same time but not wearing seatbelts. Instead we compare injury rates for people who are wearing seatbelts to people who are not wearing seatbelts.

But might people who wear seatbelts have other characteristics that affect injury rates?



For example, do these people drive safer cars? Are they less likely to drive after drinking alcohol? Are they more likely to obey speed limits?

Evaluating a policy effect

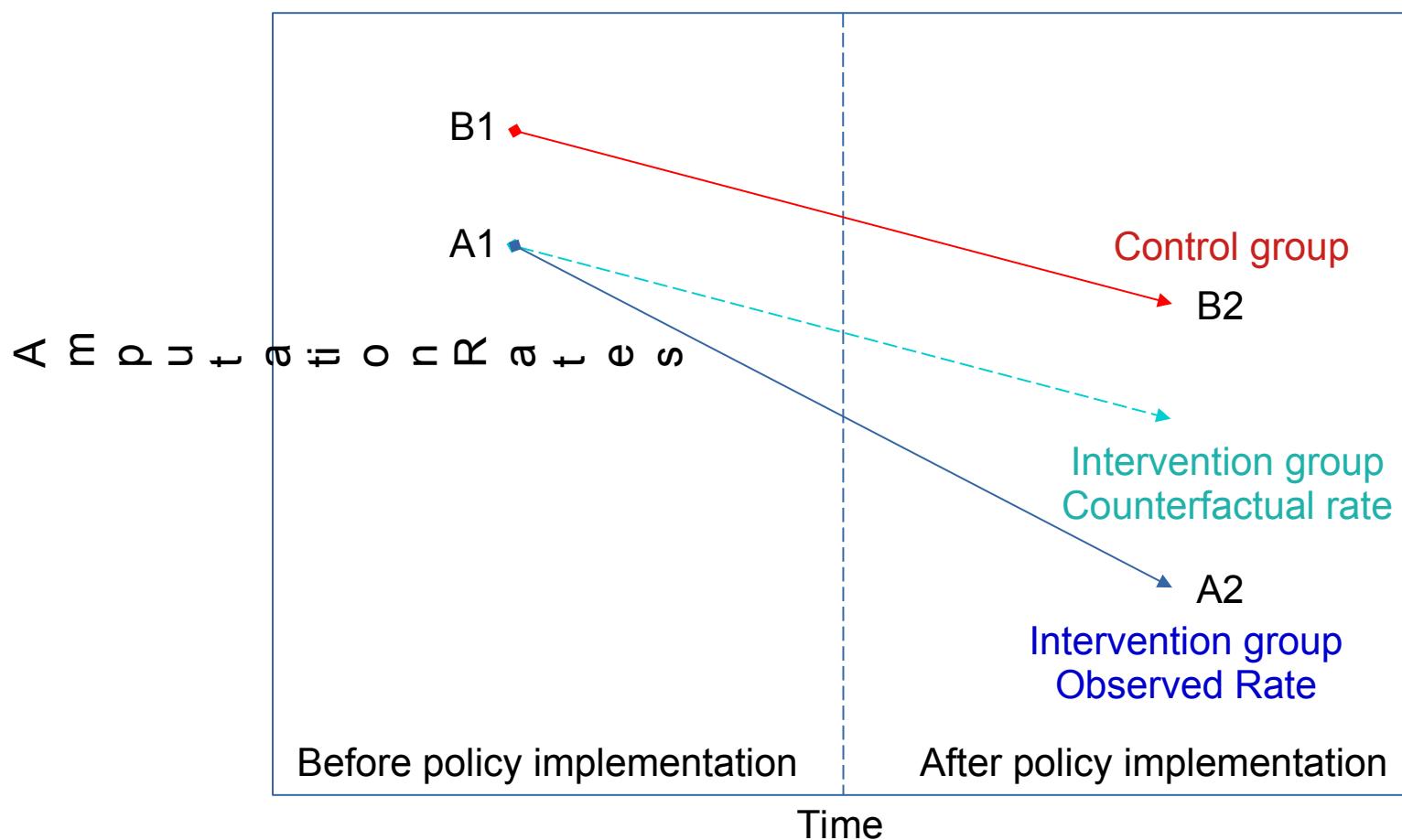
- A particular problem arises when the aim is **to distinguish a policy effect from time trends or differences in the communities.**
- For instance, data show that amputation rates overall have significantly declined in OECD countries. However, they seem to have declined more in certain countries, where certain policies have been implemented. How could we demonstrate that the policies have been effective?
- There may be changes in health outcomes over time that have occurred regardless of the policy, with trends that could be linear or nonlinear, which should be considered in addition to the policy effect.
- Under these circumstances, a simple **pre-post analysis** is not enough to distinguish changes due to the policy effect.

Difference-in-differences

- A study design that is increasingly used in these cases **difference-in-differences (DiD) approach**, which compares the change in health outcomes in the policy-exposed population to the simultaneous change in the health outcome in a comparable population unexposed to the policy.
- For example, **country A** passing a reform for public taxation in a reference year, after a certain period of time can be compared to **country B**, where the system remained always financed by health insurance. Because country B did not implement such legislation during this time, the time trend in country B is assumed to project what would have occurred over time in the policy-affected country A, had the policy not passed.
- In the **DiD analysis**, the pre-post time point differences in amputation rates in country A would be compared with the pre-post time point differences in amputation rates in country B.

DiD analysis

- The difference in pre-post time point differences in amputation rates between the two states [the quantity $(A_2 - A_1) - (B_2 - B_1)$] would be attributed to the effect of the policy.



DiD regression

- The DiD approach, while accounting for a time trend, considers unobserved confounders that might simultaneously affect both communities (e.g. countries) similarly (e.g., national economic changes that could affect amputation rates).
- This step can be carried out by estimating the **DiD coefficient** through a standard regression model:

$$Y_i = \beta_1 \times \text{affected}_i + \beta_2 \times \text{time} + \beta_3 \times (\text{affected}_i \times \text{time})$$

- where for individual i , Y is the outcome of interest ($Y=1$ amputated vs $Y=0$), “affected” is a dummy variable for residence in country A vs B, and “time” is a dummy variable for the policy period (0=pre-reform, 1=post-reform).
 - β_1 captures possible unobserved confounding differences between the populations in the two states prior to the policy
 - β_2 accounts for time trend in amputations, even without policy
 - β_3 is the **DiD coefficient of interest**—the causal effect of reform on amputations, equivalent to the quantity $[(A_2 - A_1) - (B_2 - B_1)]$, or the decrease in amputation rates in country A vs country B.
- **NB: There are many assumptions and limitations in the DiD analysis**

Potential outcomes framework

- Two potential interventions and an outcome
- Given a sample of subjects and a treatment, each subject has a pair of **potential** outcomes:

$Y_i(0)$ *outcome under the CONTROL intervention*

$Y_i(1)$ *outcome under the ACTIVE intervention*

Z indicator variable denoting the intervention received (0/1)

Only the outcome **under the actual intervention received**
is observed for each subject:

$$Y_i \quad (Y_i = Z_i Y_i(1) + (1-Z_i) Y_i(0))$$

Average treatment effects

- At subject level, the effect of the treatment is defined as:

$$Y_i(1) - Y_i(0)$$

- The **Average Treatment Effect (ATE)** is defined as:

$$E[Y_i(1) - Y_i(0)]$$

Average Treatment Effect, at population level, of moving an entire population from untreated to treated

- The **Average Treatment Effect for the Treated (ATT)**:

$$E[Y_i(1) - Y_i(0) | Z=1]$$

Average Treatment Effect on those subjects that ultimately receive the treatment

In an RCT these two measures coincide because, due to randomization, **the treated population will not, on average, differ systematically from the overall population**

ATE or ATT?

- Applied researchers should decide whether the ATE or ATT is of greater utility or interest.
- It depends from what is **RELEVANT** and **ACHIEVABLE**
- *Effectiveness of intensive, structured smoking cessation program => High barriers to participation and completion of the program=> Unrealistic to estimate the effect if it applied to all smokers=> Target current smokers wishing to enroll => ATT*
- *Effect of an information brochure given by family physicians to patients who are current smokers => Cost of distribution relatively low => minimal barriers to a patient receiving the brochure => ATE*

ATE in RCTs

- As a consequence of randomization, an unbiased estimate of the ATE can be directly computed from the study data
- Unbiased estimate of the ATE

$$E[Y_i(1) - Y_i(0)] = E[Y_i(1)] - E[Y_i(0)]$$

- Depending from the specific problem under investigation, it can be defined in terms of **means** (continuous outcomes), **difference in proportions**, **absolute risk reduction**, **relative risk**, **odds ratios** (dichotomous outcomes)
- For binary outcomes, the **Number Needed to Treat (NNT)** , or reciprocal of the absolute risk reduction, is a ***useful indicative summary measure of effect:***

$$1/ |(R_1 - R_0)| = 1/ARR$$

Observational studies

- Aim to investigate the cause-effect relationship in situations in which it is not possible to run RCTs
- In observational studies, **the treated subjects often differ systematically from untreated subjects**
- In general, it holds:

$$E[Y(1) | Z=1] \neq E[Y(1)]$$

- In general, an unbiased estimate of the average treatment effect cannot be obtained by directly comparing outcomes between the two treatment groups

Propensity Score

- The **propensity score** is the probability of treatment assignment conditional on observed baseline covariates:
$$e_i = Pr(Z_i=1 | X_i)$$
- The **propensity score is a balancing score: conditional on the propensity score, the distribution of measured baseline covariates is similar between treated and untreated subjects**
- In a set of subjects all of whom have the same propensity score, the distribution of observed baseline covariates will be the same between the treated and untreated subjects
- In RCTs the **propensity score is known and is defined by the study design**. In observational studies, the true propensity score is not, in general, known *but can be estimated from the study data*

Propensity Score estimation

- In practice, the propensity score is estimated using a **logistic regression model, in which treatment status is regressed on observed baseline characteristics**
- The estimated propensity score is the predicted probability of treatment derived from the fitted logistic regression model

$$P = \frac{e^{a+bX}}{1 + e^{a+bX}}$$

Covariates
predicting the
Treatment

- Other methods include bagging, boosting, regression trees, random forests and neural networks

Example: effects of PTCA on AMI Mortality

NY Hospital dataset

```
load(file="../data/ny_hospdata_2017.Rda")
load(file="../data/amidata_2017.Rda")
names(amidata_2017)

amidata<-amidata_2017

amidata$ptca<-NA
amidata$ptca<-ifelse(amidata$procedure=="185",1,amidata$ptca)
amidata$ptca<-ifelse(amidata$procedure!="185" & !is.na((amidata$procedure)),0,amidata$ptca)

glm(ptca~cl_age+males+risky+severe,family = binomial("logit"),data=amidata)

          Beta    2.5 % 97.5 % P    OR      2.5 % 97.5 %
(Intercept) 1.4416  1.2818  1.6015 0 4.2274 3.6033 4.9605
cl_age       -0.3808 -0.4163 -0.3453 0 0.6833 0.6595 0.7080
males        0.3984  0.3502  0.4466 0 1.4894 1.4193 1.5630
risky        -0.4153 -0.4462 -0.3844 0 0.6602 0.6401 0.6808
severe       -0.0860 -0.1152 -0.0568 0 0.9176 0.8912 0.9448

# Propensity Score
amidata$p<- predict(PropLogit,newdata=amidata,type="response")
```

Ignorable Treatment assignment

- Treatment assignment is strongly ignorable when:
Independent of the potential outcomes conditional on the observed baseline covariates => **“no unmeasured confounders”** => must use regression techniques to adjust for those observed
Every subject has a *nonzero probability* to receive either treatment
- The above conditions justify using propensity score methods to estimate treatment effects in observational studies

Key messages

- *The propensity score is a “balancing score”*: **conditional on the propensity score**, baseline covariates will be balanced between treatment groups
- Propensity score methods allow separating the design of an observational study from its analysis
- Similarly to RCTs, propensity score methods allow estimating marginal (or population-average) treatment effects

Materials

- P.Austin, An introduction to propensity score methods for reducing the effects of confounding in observational studies, *Multivariate behavioural research*, 46:399-424, 2011



ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Second cycle degree programme (LM) in Statistical Sciences

85278 – Methods and tools for Health Statistics

**85277 – Methods and tools for Official Statistics: Population
and Health Statistics**

a.a. 2022/2023

Stream 3. Statistical methods and tools to plan and to evaluate health policies

Topic 3.1.2

*Applied methods for propensity scores
Introduction to the Synthetic Control Method*

Fabrizio Carinci

fabrizio.carinci@unibo.it

Monday, 13th March 2022

Propensity Score Methods

- Four main propensity score methods:
 - **I. Propensity Score Matching**
 - **II. Stratification on the Propensity Score**
 - **III. Inverse Probability of Treatment Weighting**
 - **IV. Covariate adjustment using the Propensity Score**

Method 1. Propensity Score Matching

- A new dataset is formed by matched sets of treated and untreated subjects who share a similar value of the propensity score
- Most common is *one-to-one* matching: pairs of treated and untreated subjects are formed, so that matched subjects have similar values of the propensity score
- After matching, any test or technique for matched pairs would be appropriate to directly compare outcomes between treated and untreated subjects **as for any RCTs**: *paired t-test for matched pairs (continuous outcomes)*, *Mc Nemar Test for difference in proportions (binary outcomes)*
- Residual differences between covariates can be adjusted using regression models (using unconditional, conditional or GEE)

Techniques for Matching

- Decisions to be made:
 - Matching without vs with replacement
 - Greedy vs Optimal matching
 - Nearest neighbor without or with caliper

Matching without vs with replacement

- *Without replacement*: once an untreated subject has been selected to be matched to a given treated subject, that untreated subject is no longer available as a potential match for other treated subjects
- *With replacement*: allows a given untreated subject to be included in more than one matched set. The sampling scheme shall be considered when estimating variance.

Greedy vs optimal matching

- *Greedy*: a treated subject is selected at random. The untreated subject whose propensity score is closest to that of this randomly selected subject is chosen for matching. Process is repeated until untreated subjects have been matched to all treated subjects for whom a matched untreated subject can be found
- *Optimal*: matches are formed so as to minimise the total within-pair difference of the propensity score

Distance criteria to perform matching

- *Nearest neighbor:* select for matching to a given treated subject that untreated subject whose propensity score is closest to that of the treated subject. If multiple are equally close, select at random. *No restrictions on maximum acceptable difference*
- *Within a specific caliper distance:* the absolute difference in the propensity scores of matched subjects must be below some prespecified threshold (caliper distance). Unmatched treated subjects would be excluded from the resultant matched sample
 - *Caliper choice:* Theory supports the choice of 0.2 of the pooled standard deviation of the logit of the propensity score: $\log(p/(1-p))$

Example: effects of PTCA on AMI Mortality

NY Hospital dataset

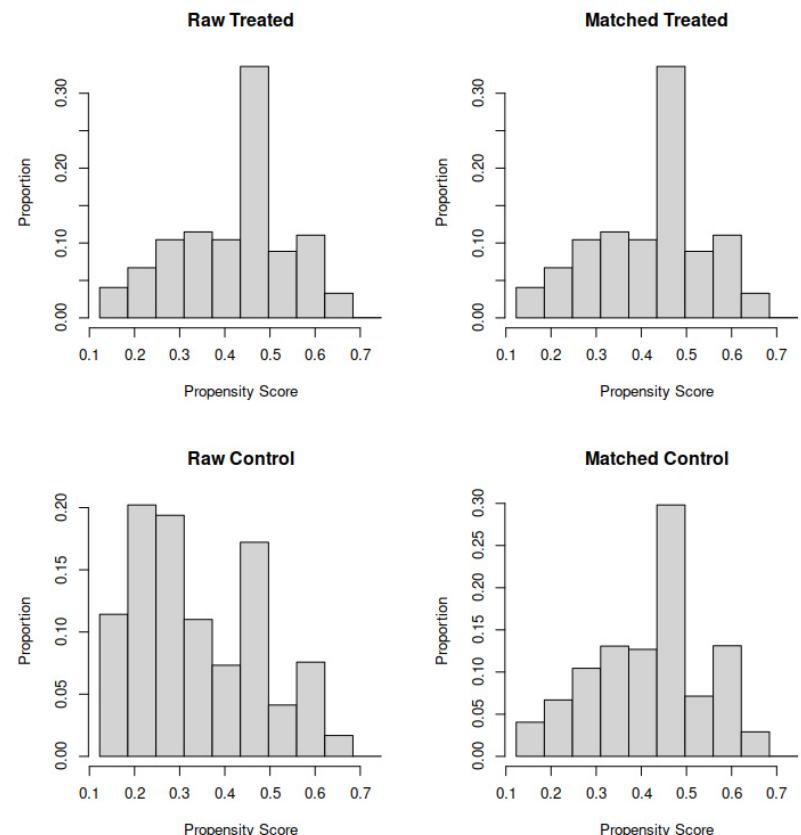
```
library(MatchIt)

p.match<-matchit(formula = ptca~cl_age+males+risky+severe,
                  data      = amidata,    # Cannot have missing values. Use complete dataset
                  method    = "nearest", # greedy match
                  distance = "logit",   # Distance defined by logistic model
                  ratio    = 1)          # 1:1 match is the default

plot(p.match,type = "jitter",interactive=F)
plot(p.match,type = "hist",interactive=F)

glm(dead~ptca,data=dat.match,
    family=binomial(link = "logit"))

          OR      2.5 % 97.5 %
(Intercept) 0.0321  0.0290  0.0354
ptca        0.3959  0.3284  0.4752
```



Method 2. Stratification on the propensity score

- Subjects are stratified into mutually excluding subsets based on previously defined thresholds of the estimated propensity score
- A common approach is to divide subjects into five equally sized groups using the quintiles of the estimated propensity score
- Similar to a meta-analysis of (quasi)RCTs. Within each stratum, the effect of treatment on outcomes can be estimated by comparing outcomes directly between treated and untreated subjects.
- The stratum-specific estimates of effect can then be pooled across strata to estimate an overall treatment effect.
- Mantel Haenszel estimator may represent a simple method. Stratified regression can be an alternative. A specific method of meta-analysis that would fit the type of outcome can be selected according to the problem under investigation

Example: effects of PTCA on AMI Mortality

NY Hospital dataset

```
amidata$p_group<-cut(amidata$p, quantile(amidata$p, probs=seq(0,1,0.20)),  
                      labels=FALSE, include.lowest=TRUE, right=TRUE)  
  
for (i in 1:5) {  
  DLogit<-glm(dead~ptca, family=binomial("logit"), data=amidata[amidata$p_group==i,])  
  ...  
  rownames(curmodel)<-paste("ptca STRATUM", i)  
  if (i>1) {alld<-rbind(alld, curmodel)} else {alld<-curmodel}  
}  
print(alld)  
          or  lcl_or ucl_or  
ptca STRATUM 1 0.4910 0.3938 0.6058  
ptca STRATUM 2 0.3099 0.2238 0.4185  
ptca STRATUM 3 0.1849 0.0897 0.3405  
ptca STRATUM 4 0.3820 0.1034 1.1741  
ptca STRATUM 5 0.1105 0.0060 0.5884  
  
library(Epi)  
clogistic(dead~cl_age+males+risky+severe+ptca, strata=p_group, data=amidata[amidata$p_group>2,])  
  
          Beta      2.5 %     97.5 %      OR      2.5 %     97.5 %  
cl_age   0.6819495 -0.1158768  1.4797759 1.9777297 0.8905849 4.3919613  
males    0.1586876 -0.9272987  1.2446739 1.1719718 0.3956210 3.4718026  
risky    0.6751293 -0.1703488  1.5206073 1.9642869 0.8433706 4.5750029  
severe   0.8606644  0.5763587  1.1449700 2.3647312 1.7795467 3.1423472  
ptca    -1.3921605 -1.9465509 -0.8377701 0.2485378 0.1427656 0.4326743
```

Method 3. Inverse Probability of Treatment Weighting

- Inverse probability of treatment weighting uses weights based on the propensity score to create a synthetic sample in which the distribution of measured baseline covariates is independent of treatment assignment
- Similar to **survey sampling weights** that are used to make samples representative of specific populations

Z_i , indicator variable of treatment in the i -th subject

e_i , propensity score for the i -th subject

Weights are defined as (*model-based standardization*):

$$w_i = \frac{Z_i}{e_i} + \frac{(1 - Z_i)}{1 - e_i}$$

Inverse Probability of Treatment Weighting

- Differently from matching, this method allows using **all observations in the sample** (in matching they are limited by the number of treated cases)
- A simple estimate of ATE with a set of weights would be the following:



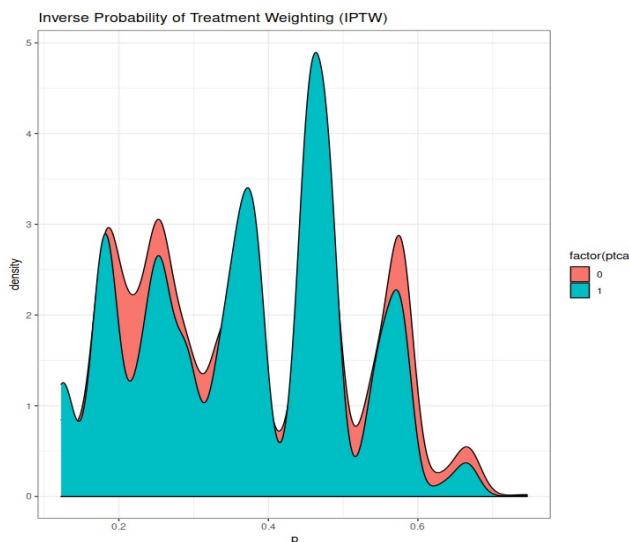
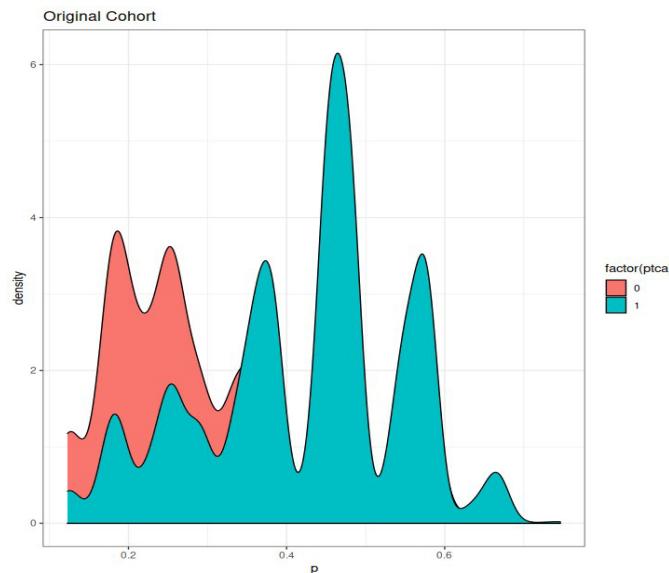
$$ATE = \frac{1}{n} \sum_{i=1}^n \frac{Z_i Y_i}{e_i} - \frac{1}{n} \sum_{i=1}^n \frac{(1 - Z_i) Y_i}{1 - e_i}$$

- When weights are derived using Z_i as the first term (instead of the inverse of the propensity score), a similar formula is obtained for ATT
- Regression models using weights are called marginal structural models and require robust variance estimation

Example: effects of PTCA on AMI Mortality

NY Hospital dataset

```
amidata$weight<-ifelse(amidata$ptca==1,1/amidata$p,1/(1-amidata$p) )
```



```
glm(dead~ptca,weights=weight,data=amidata,family=binomial(link="logit"))
```

	Beta	2.5 %	97.5 %	P	OR	2.5 %	97.5 %
(Intercept)	-2.809	-2.8544	-2.7641	0	0.0603	0.0576	0.0630
ptca	-0.863	-0.9447	-0.7821	0	0.4219	0.3888	0.4574

Method 4. Covariate Adjustment using the Propensity Score

- The outcome variable is regressed on an indicator variable denoting treatment status and the estimated propensity score
- The effect of treatment is determined using the estimated regression coefficient from the fitted regression model
- This method assumes that the relation between the treatment and propensity score and the outcome has been correctly modelled

Example: effects of PTCA on AMI Mortality

NY Hospital dataset

```
DLogit<-glm(dead~ptca+p, family=binomial("logit") , data=amidata)  
...
```

	Beta	2.5 %	97.5 %	P	OR	2.5 %	97.5 %
(Intercept)	1.1485	0.9849	1.3143	0	3.1536	2.6775	3.7222
ptca	-0.9481	-1.1213	-0.7813	0	0.3875	0.3259	0.4578
p	-14.7743	-15.5486	-14.0203	0	0.0000	0.0000	0.0000

Comparisons of propensity score methods

- Propensity score matching eliminates a greater proportion of the systematic differences in baseline characteristics
- Weighting sometimes performed better than matching
- Stratification results in estimates of average treatment effects with greater bias than weighting
- Covariate adjustment using the propensity score is the only method that does not separate the study design from analysis. Under these conditions, the investigator can always force a model in that does perform better and has the “outcome in sight”, while eventually the problem may relate to the study design and cannot be resolved in the model

Balance diagnostics

- *How correctly has been the propensity score model specified?* This depends from the differences in baseline covariates between treated and untreated subjects, which should be minimal.
- A comparison of means and percentages represent a minimal check that should be always carried out. A threshold for standardized differences (e.g. 0.1) can be used to check for excessive deviations.
- The problem is more complex because the entire distribution of baseline covariates should be balanced. Various graphical methods may be required.
- If large deviations are found, this can be an indication that the propensity score model is not well specified and may require improvements e.g. additional covariates, interaction terms etc
- Check for statistical significance shall be discouraged due to sample size effects and derived nature of synthetic samples

Variable selection for the propensity score model

- The propensity score is defined to be the probability of treatment assignment:

$$E_i = \Pr(Z_i = 1 | X_i)$$

- The propensity score model should only include variables that are measured at baseline and not post-baseline covariates that may be influenced or modified by the treatment
- Theory encourages using only covariates that predict treatment
- However, in practical situations most subject-level baseline covariates likely affect both treatment assignment and the outcome. Therefore, in many settings, it is likely that one can safely include all measured baseline characteristics in the propensity score model

Conditional vs Marginal Estimates of Treatment Effect

- *Conditional treatment effect*: average effect of treatment on the individual
- *Marginal treatment effect*: average effect of treatment on the population
- In an RCTs, the results obtained from these two measures will coincide
- Propensity score methods allow for the estimation of the marginal treatment effect.
- In binary outcomes or time to event analyses, even when the propensity score model is well specified, there are no unmeasured confounders, and the relation with the outcome is known, it is not possible to obtain an unbiased estimate of the conditional treatment effect. The marginal and conditional treatment effects approximately coincide when using continuous outcomes and linear regression.

Regression adjustment vs Propensity Score methods

There are practical reasons for preferring the use of propensity score-based methods to regression-based methods when estimating treatment effects using observational data:

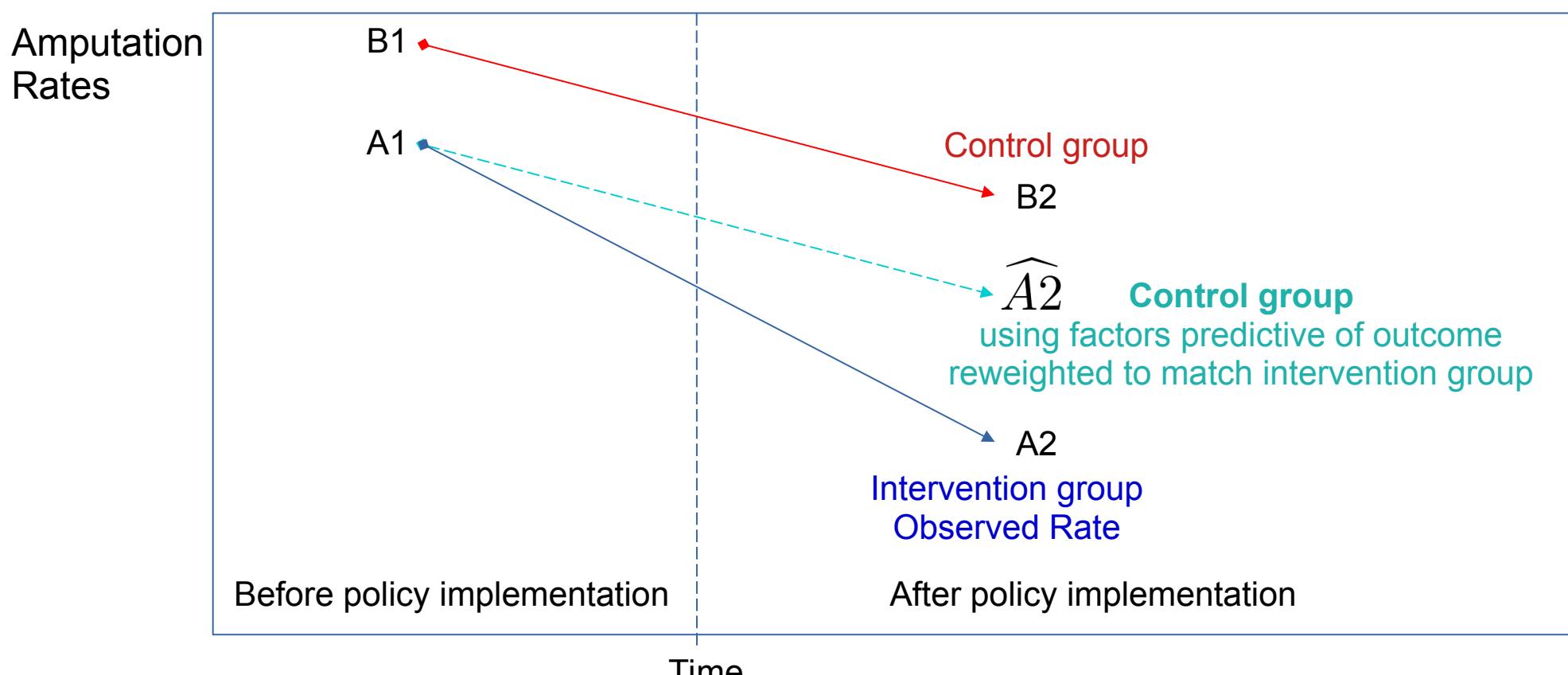
- 1) Simpler to determine whether the propensity score model has been adequately specified compared to assessing treatment and baseline characteristics together
- 2) These methods allow separating the design of the study from the analysis of the study
- 3) In situations of small number of events (e.g. rare disease) there may be more flexibility in modelling the propensity score, as the number of treated subjects is usually much higher than the number of outcomes
- 4) The degree of overlap in the distribution of baseline covariates between the two treatment groups can always be examined. In case of non overlap, it is possible to declare the analysis not plausible, while this can be missed by a naive analyst using regression models

Propensity scores in DiD analysis

- A concern with DiD models is that the program and intervention groups may differ in ways that are related to their trends over time, or their compositions may change over time.
- Propensity score methods may be used with DiD analysis. However, a particular complication is that there are no longer just two groups (treatment and comparison), but essentially four: treatment pre, treatment post, comparison pre, and comparison post.
- Propensity score methods can ensure the comparability of DiD analysis where the composition of each group may change over time, such as if the patient population served by physician practices changes systematically across time, or if the composition of physician groups changes differentially over time due to turnover.
- In practice, **the DiD analysis is run with propensity score by using a regression model with interaction term, where propensity score matching is used as a balancing factor between two groups**

Synthetic Control Method

- SCM is a method to evaluate interventions that occur at the aggregate level, in a distinct unit (e.g., a state, country, age group), and a clearly differentiated point of time e.g. policy in one country to reduce amputation rates
- The goal is always to find a proper control (counterfactual) from observational data. The control is a weighted set



Outline of the method

- Synthetic control uses optimization to determine the best set of weights for the controls given the available data.
- Unexposed units form a “donor pool” of *potential* controls:
 - Same type of unit as the one with intervention
 - Not exposed to the intervention
 - Not experiencing special events (“shocks”) not related to the outcome
- Outcomes should be approximately continuous e.g. mortality rates or counts, prevalence etc
- Use of covariates is optional. Covariates are pre-intervention characteristics that can affect the post-intervention outcomes that would have been realized in absence of the intervention

Set of weights

- The optimal weights w_i^* are determined by **minimizing the distance** between the synthetic control and the treated unit using a *variable importance* (v_k) - weighted mean squared error function

$$\sum_{k=1}^K v_k \left(X_{1k} - \sum_{i=2}^N X_{ik} w_i^* \right)^2$$

- k is the number of variables used by the algorithm
- X_{1k} represents the value of the variable k in the treated unit $i = 1$, and X_{ik} represents the values among controls that are used with weights to determine the distribution of the variables in the synthetic control unit.
- X can include the outcome variable at each pre-intervention time point, some combination of pre-intervention outcomes and other covariates
- v_k can be manually included or calculated using measure of correlation between the variables entered into the matching procedure and the pre-intervention outcomes **in the treated unit**

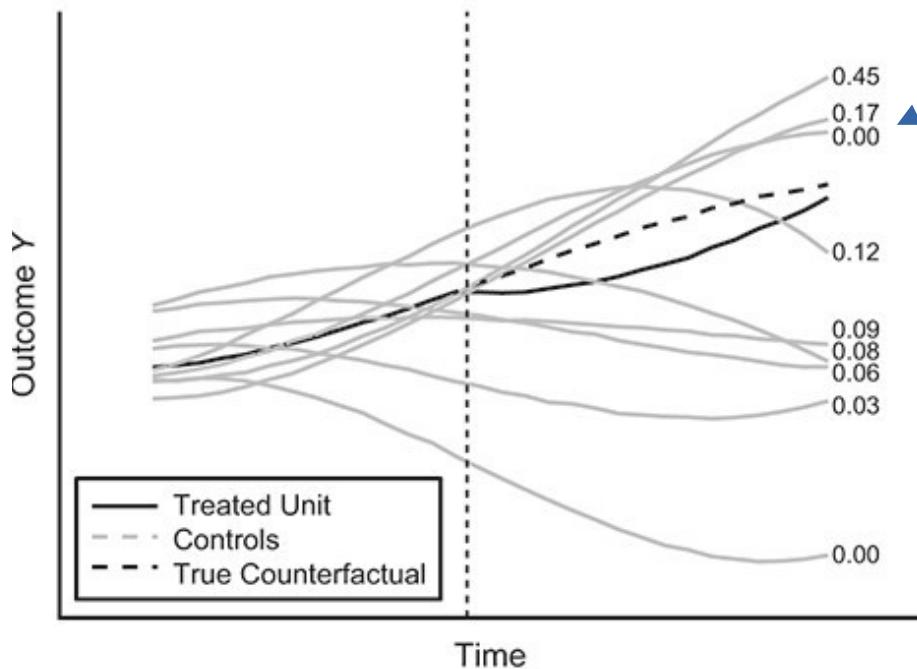
Estimated effect

- Calculate time-specific differences and plot the temporal evolution of the estimated effect, or take the difference or percentage change over the entire post-intervention period.
- Estimated counterfactual for each post-intervention time point t :

$$\widehat{Y_{1t}(0)} = \sum_{i=2}^N Y_{it} w_i^*$$

Y_{it} is the time-specific outcome in unit i ($i = 1$ is the treated unit, and the rest are controls), and w_i^* is the unit weight assigned to each control unit.

Synthetic Control Method



Unit weights are calculated in the pre-intervention period (before the dashed vertical line).

Each control unit is assigned a unit weight ranging from zero to 1 (the sum of the weights is always 1)

The synthetic control outcomes are given by the weighted sum of outcomes among controls, obtained by multiplying the time-specific outcomes in each unit with its respective unit weight, and then summing across all control units.

The time series of **post-intervention outcomes in the synthetic control** provide an estimate of the counterfactual outcomes in the treated unit, which is then **compared with the observed data** to estimate the **intervention effect**

Robustness checks

- What confidence should we place in the results? Confidence intervals have been difficult to calculate so far. 'Placebo effects' can be used to check the size of the effect for each unit separately. The aim is to show that the size of the effect is greater in the treated unit. If the gap for the 'real' treated unit is a true intervention effect then the gap should be large compared to the gap for all the placebo analyses
- The basis is the following: for each control unit i , shift the treated unit into the control group and run an equivalent synthetic control analysis to the one in the treated unit and store the estimated counterfactuals $\widehat{Y}_{it}(0)$
- Calculate the **squared prediction error** at each time point t by squaring the difference between the observed outcomes and the outcomes in the synthetic control in each unit (including the treated unit):

$$e_{it} = \left(Y_{it} - \widehat{Y}_{it}(0) \right)^2$$

Root mean squared prediction error

- The ratio between postintervention and pre-intervention root mean squared prediction error (RMSPE) is used to compare effects between units
- This is a measure of how big the gap between the real treated unit and synthetic control is after the intervention date, *as a function of the gap before the intervention*
- If the fit between the actual and synthetic treated units before the intervention is not good the ratio will be smaller.
- **The larger the ratio, the more convincing the evidence that the intervention has had an effect.**

Other methods

- Other recently emerging methods that have been applied to compare the effect of health policies for large-scale interventions include:
 - Regression discontinuity
 - Interrupted time series
 - Instrumental variables

Key messages

- *Propensity score methods allow estimating treatment effects in metrics that are similar to those reported by RCTs:* when outcomes are binary, one can report risk differences, NNTs or the relative risk, as opposed to odds ratio with logistic regression models
- Methods e.g. DiD, PSM and SCM can be used effectively to evaluate the health impact of large-scale population level interventions. These methods are applied on *aggregate data*
- Other methods are rapidly emerging in the scientific literature, and should be adequately considered in the immediate future

Materials

- P.Austin, An introduction to propensity score methods for reducing the effects of confounding in observational studies, *Multivariate behavioural research*, 46:399-424, 2011