INFACT
Joint Action on Health Information

Road map and case study report

Deliverable 9.1

Innovative use of data sources: A Cross-sectional study of Data Linkage Practices across European Countries

February 12, 2020

Santé publique France

HIGIENOS INSTITUTAS

Submission Date: February 12, 2020

WP9 Lead: Department of Non-Communicable Disease and Injury, Santé Publique France, France
WP9 Co-Lead: Health Information Centre and Institute of Hygiene, Lithuania
Joint Action Coordination:
Sciensano | Rue Juliette Wytsmanstraat 14 |
1050 Brussels | Belgium| e-mail: infact.coordination@sciensano.be|
Website: www.inf-act.eu|Twitter: @JA_InfAct

This project is co-funded by the Health Programme of the European Union

Table of Contents

Executive summary

Background: The availability of data generated from different sources is increasing with the possibility to link these data sources together. However, linked administrative data can be complex to use and may require advanced expertise and skills in statistical analysis. The main objectives of this study were to describe the current use of data linkage at the individual level and the artificial intelligence (AI) in routine public health activities and to identify the related health outcome and intervention indicators and determinants of health for non-communicable diseases.

Method: We performed a survey across European countries to explore the current practices applied by national institutes of public health and health information and statistics for innovative use of data sources (i.e., the use of data linkage and/or the AI).

Results: The use of data linkage and the AI at the national institutes of public health and health information and statistics in Europe varies. The majority of European countries use data linkage in routine by applying a deterministic method or a combination of two types of linkages (i.e., deterministic & probabilistic) for public health surveillance and research purposes. The use of AI to estimate health indicators is not frequent at national institutes of public health and health information and statistics. Using linked data, 46 health outcome indicators related to seven health conditions, 34 indicators related to determinants and 23 to health interventions were estimated in routine. Complex data regulation laws, lack of human resources, skills, and problems with data governance, were reported by European countries as obstacles to linking different data sources in routine for surveillance and research.

Conclusions: Our results highlight that the majority of European countries have integrated data linkage in routine public health activities but a few use AI. A sustainable national health information system and a robust data governance framework allowing to link different data sources are essential to support evidence-informed health policy development process. Building analytical capacity and awareness of the added value of data linkage in national institutes is necessary for improving the utilization of linked data in order to improve the monitoring of public health activities.

Keywords: Innovation, linked data, artificial intelligence, machine learning technique, health status monitoring, public health surveillance, health information, health indicators.

Key points

- The use of data linkage and the AI at the national institutes of public health and health information and statistics in Europe varies.

- Complex data regulation laws, lack of human resources, skills, and problems with data governance, were reported by European countries as obstacles to linking different data sources in routine public health surveillance and research purposes.

- Building analytical capacity and awareness of the added value of data linkage at national institutes is necessary for improving the utilization of linked data in order to improve the monitoring of public health activities.

I. Background

The availability of administrative data generated from different sources is increasing and the possibility to link these data sources with other databases offers unique opportunities to answer those research questions which require a large sample size or detailed data on hard-to-reach populations [1]. Data linkage can generate evidence with a high level of external validity and applicability for policymaking [1]. Over an extended period, these population data (i.e., linked administrative data) can ensure high statistical power, reducing methodological issues relating to attrition, recall bias and lost-of-follow up [2], allowing more detailed stratified analyses of subgroups according to age, or specific geographical regions, and provide rapid access to data collected in a standardized format [3-5].

The value of any surveillance system ultimately depends on timely and reliable information [6]. There are several data sources that are used for public health surveillance, for example, health interviews and examination surveys, diseases-specific registries, epidemiological cohort studies, hospital discharge data, health insurance claims, mortality database, etc. Traditional data sources (e.g., health interview and examination surveys, disease-specific registries, etc.) and administrative data sources (e.g., hospital discharge, health insurance claims, causes of mortality data, etc.) complement each other and can increase the completeness and comprehensiveness of health information by taking into account various dimensions of health and risk factors influencing health status directly and indirectly.

Linking various data sources improves completeness and comprehensiveness of information to guide the health policy process, effective patient care and health services management [7]. Data linkage is an important technique that connects detailed individual-level information from different data sources. This methodology potentiates the capacity to study disease burden and progression, risk factors, care pathways and long-term outcomes for public health research and health surveillance [1]. However, linked administrative data can be complex to use and may require advanced expertise and skills in statistical analysis [8]. Generating efficiently comparable and timely health information across the European Union (EU), European Economic Area (EEA) and other European countries require to perform data linkage and apply AI to estimate health indicators. Many countries have already invested in data linkage to improve their health information system [9], but there are wide differences in capacity across European countries to perform data linkage in routine.

We explored the differential use of data linkage in routine health monitoring based on the latest developments in new methods and analysis across European countries. This study was carried out under the InfAct (Information for Action) [10] which is a joint action of Member States (MSs) aiming to develop a more sustainable EU health information system through improving the availability of comparable, robust and policy-relevant health status data and health system performance information. InfAct gathers 40 national health authorities from 28 Member States. This study is part of a work package (WP9) focused on innovation in the health information system (i.e., using data linkages and/or AI) to improve public health surveillance and health system performance for the health policy development process.

II. Objectives

The main objectives of this study were 1. to describe the current use of data linkage at the individual level and the AI techniques applied in routine public health activities and 2. to identify the relevant health outcome and intervention indicators estimated and determinants of health for non-communicable diseases.

III. Methodology

We performed the following steps to achieve the objectives of this study:

1. Literature search

We reviewed the existing literature published on the use of data linkage and the AI (i.e., one technique of AI is machine learning technique) for health status monitoring using PubMed on Dec. 1, 2018. We included in our search peer-reviewed articles, systematic reviews and published reports published in the English language. The search strategies used are reported in *additional file 1*. Based on the review, we identified different data sources used for data linkage, the use of artificial intelligence [AI]), health outcome and intervention indicators and determinants of health (A*dditional file2).* This was not an exhaustive search and was performed only to identify any existing questionnaire or relevant information to be used to develop a questionnaire to identify the current practices in innovative use of data sources across European countries.

2. Definition of innovative use of data sources

We developed the definition of "innovative use of data sources" in the context of public health and health information system and defined as:

- *The linkage of different data sources* (health surveys and/or disease-specific or population-based registries and/or national cohort and/or clinical research datasets and/or administrative data and/or electronic health records and/or X-data sources i.e., information on determinants of health and can include data on various exposures *[Additional file 2]*) with each other using linkage technology and/or
- *The use of AI* either to linked data or to an individual data set,

allowing a better understanding of what determines population health or to promote the efficiency of the health system and guide decision making at different geographical levels or at other categorization parameter level.

3. Development of web-based survey

We developed a questionnaire and requested information on data sources used for linkage, general characteristics of data linkage, the use of AI to estimate health indicators, related health outcomes and intervention indicators estimated and determinants of health of non-communicable diseases *(Additional file 3)*. We reported these results according to the status of using data linkage or the AI in routine public health activities across European countries into the following three categories: *1. Advanced* (i.e., those who use data linkage or the AI in routine to estimate health indicators), *2. In Progress* (i.e., those for whom the deployment of these innovative techniques [i.e., data linkage or the AI] is still underway and expect to integrate these techniques in the next 5 years), and *3. Not yet* (i.e., those for which the use of these techniques are not foreseen yet).

Survey participants were asked to report at least three health indicators related to priority medical conditions in their country. We adopted the Euro-REACH Framework (It is an international collaboration to improve access to health care data through cross-country comparisons) [11] to classify the identified health outcome indicators, determinants of health and intervention indicators under the following categories: health outcome indicators (1. Health characteristics, 2. Mortality, 3. Human function and quality of life and 4. Life expectancy and well-being), determinants of health (1. Physical environment, 2. Socioeconomic and environment, 3. Health behavior and lifestyle and 4. Biological /metabolic parameters) and intervention indicators (1. Prevention, 2. Promotion and 3. Others) (*Fig.1, additional file* 2).

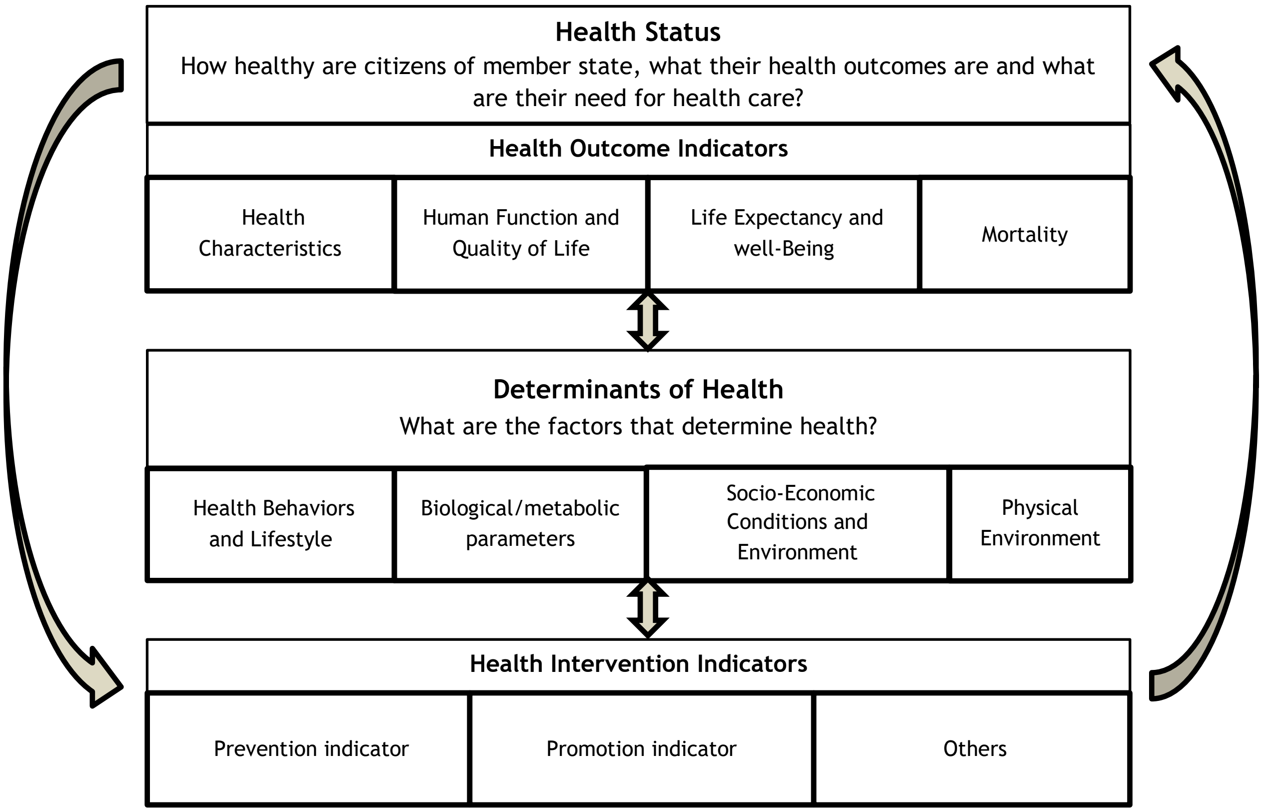


Fig. 1: EuroREACH Framework for health status monitoring

We also asked specific information about objective of health indicators estimation (i.e., for public health monitoring, scientific research [clinical, epidemiology, public health], both), status of their use (i.e., was used, currently in use or could be produced in future) and level of estimation (i.e., national, sub-national, metropolitan, at all levels). If the same health indicator was reported more than once either as being estimated currently or to be estimated in the future by different countries, we counted those health indicators once. The web-based questionnaire was developed using the Lime Survey tool by the Data lab of Santé Publique France. The questionnaire includes both closed and open-ended questions (i.e., 20 questions) *(Additional file 3)*. This questionnaire was reviewed by a group of experts in health information systems in their country and revised according to their feedback before the launch of the survey. The web-based version of the questionnaire was pretested by the co-authors (SME, RH and RG) from respective national public health institutes of Austria, France and Lithuania to check the visibility of the questions and contents.
The survey participants were the partners of the InfAct project and/or national representatives or experts or advisors for health information in their countries and could be either employed by the government, national institutes of public health and health information and statistics or research departments of the universities.
The invitation email with an electronic link to the questionnaire was sent on April 1, 2019, to the identified representatives in 31 European countries to complete the survey in four weeks (i.e., April 30, 2019) *(Additional file 4)*. The first reminder through email was sent after one month of survey launch, on May 3, 2019, and the second reminder after two weeks of the first reminder, on May 23, 2019. The abbreviations of the member countries and the names of the survey respondents are reported in *additional file 5*.

Study outcomes

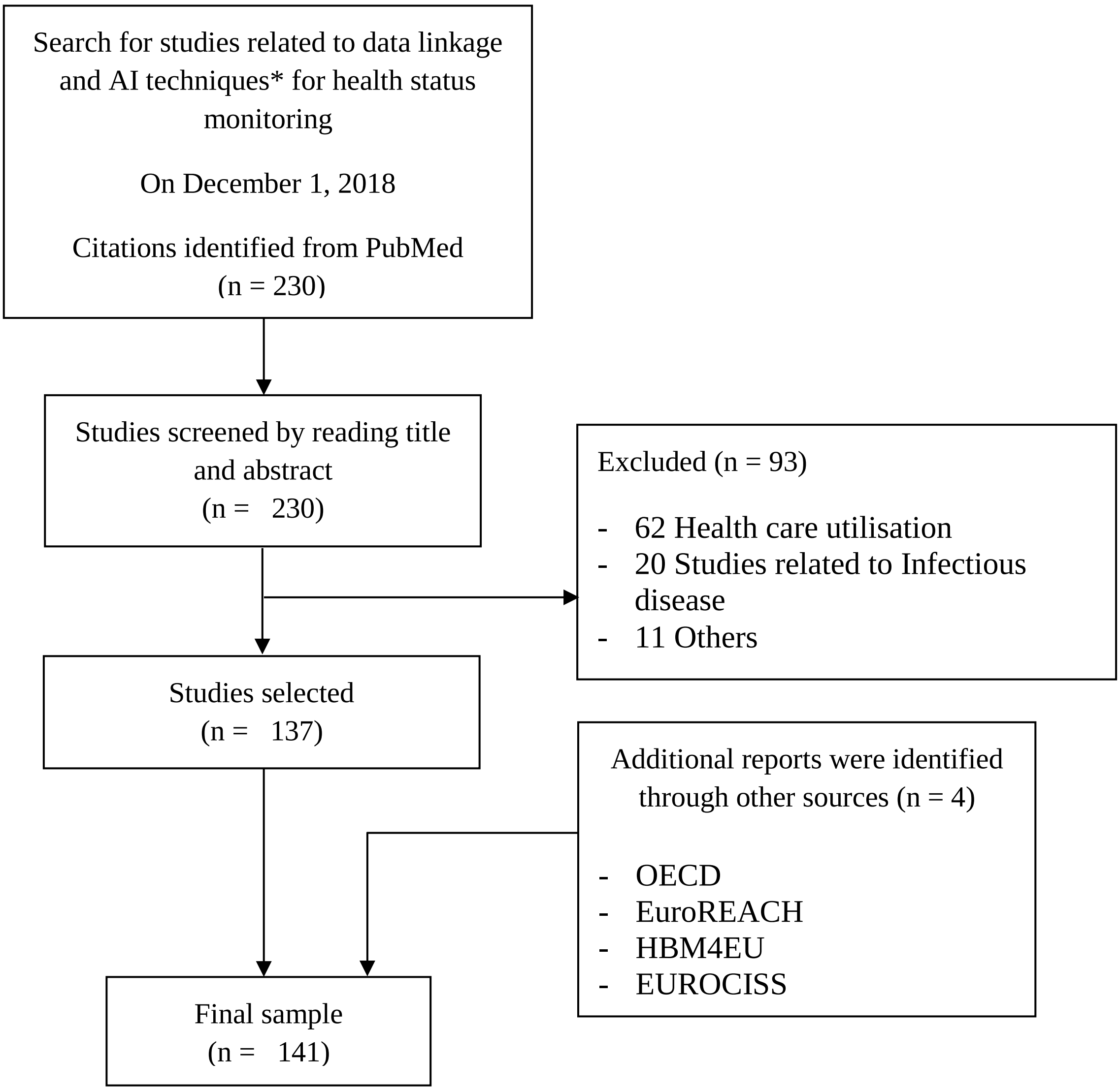
The main outcomes of this study were the current practices in data linkage and the AI and related health indicators estimated in routine public health activities across European countries. A descriptive analysis of the web-based questionnaire results has been performed using Microsoft Excel.

IV. Results

A. Literature search

We reviewed 137 citations from PubMed and four reports from the following organizations: OECD (Organization for Economic Co-operation and Development) [12], EuroREACH [13], HBM4EU (Human Biomonitoring for Europe) [14], EUROCISS (European Cardiovascular Indicators Surveillance Set) [15], to develop this questionnaire (Fig. 2).

Fig. 2: Flow diagram of studies using linked data and artificial intelligence for health status monitoring to develop a questionnaire identifying various practices of data linkages across European countries in 2019



Thirty-one European countries [28 EU-MSs + 2 EEA (Iceland and Norway) + others (Serbia)] were invited to participate in the WP9 survey and twenty-nine countries (i.e., EU MSs 27 + EEA 1 [Norway] + Others 1 [Serbia]) participated with a response rate of 94% (29/31). Hungary, Iceland and Northern Ireland did not participate. For the UK, data were provided separately by the three countries England, Scotland and Wales but was counted as one member state. The results have been validated by all survey respondents.

B. Use of data linkage in routine public health activities

Our survey results highlighted that 24 European countries perform data linkage in their routine public health activities. These countries link administrative data such as EHRs, mortality data, disease-specific registries whereas six of them (Cyprus, Italy, Poland, Portugal, Spain and Slovakia) are also developing this technique further to link with different other data sources (i.e., demographic data, domestic/leisure accidents data, congenital anomalies registry). Ireland and Latvia have ongoing initiatives of data linkage *(Table 1.1)*.

Table1.1: Current status of European countries using data linkage in routine public health activities in 2019 for innovative use of data sources

Use of Data Linkage			
	Advanced **N = 24**	**In progress*** **N = 8**	**Not yet** **N = 3**
European Countries	AT, BE, BG, CY, CZ, DE, DK, EE, ES, FI, FR, HR, IT, LT, MT, NL, NO, PL, PT, SL, SK, SRB, SW, UK (ENG, SC, WL)	CY, ES, IE, IT, PL, PT, SK, LV	GR, LU, RO

** 6 countries (CY, ES, IT, PL, PT & SK) use data linkage in routine (i.e., advanced) but also developing further this technology to link different other data sources (i.e., in progress).*

Three countries (Greece, Luxembourg, and Romania) have not yet planned any perspectives to integrate data linkage in routine public health activities. Following reasons were mentioned by some countries for not having institutionalized data linkage in their country: lack of a public health institution which should collect and govern the health-related data, data linkage is not part of the health agenda, lack of commitment from the ministry of health, lack of resources to establish a national health information system, and the institutional complexity of the Ministry of Health and strict laws and regulations which hinder data linkage with different data sources.

Objectives of data linkage: Data linkage is performed in routine for different objectives such as for health status monitoring, health system performance, health policy or for scientific research (i.e., public health, epidemiology or clinical) purposes. Our results showed that data linkage was performed for health status monitoring in 20 countries (BE, CY, CZ, DE, DK, EE, ES, FI, FR, HR, IT, LT, MT, NL, PT, SI, SK, SRB, SW, UK (SC, WL), for health policy development in 13 (AT, BE, BG, DK, EE, FR, MT, NL, NO, PL, SK, SW, UK (SC, WL) and for scientific research (public health, epidemiological and clinical) purposes in 13 (BE, CZ, DE, DK, EE, ES, FI, FR, NL, PT, SI, SW, UK (ENG, SC, WL). Finland, Spain, Sweden and Scotland also perform data linkages to identify the risk factors. In Sweden, data linkage is also used to monitor compliance with national treatment guidelines to improve health care quality.

Data sources used for linkage: Our results showed that 24 European countries who perform data linkage in routine, used most frequently five following data sources: health-related administrative data sources, non-health related administrative data sources, disease-specific registries, national health surveys, population-based epidemiological cohort, and clinical trials. *(Table 1.2)*.

Table 1.2: Data sources used for linkage across European countries in 2019 for innovative use of data sources

S/No	Data sources used for linkage		European countries		
			Advanced N = 24		In progress N = 2
1	*Health-related administrative data sources (i.e., Electronic Health Records) ʘ*	Primary care visits, emergency care, referral records, hospital discharge, prescribed medications, health insurance claims, diagnostics procedures, laboratory tests, biobank	21	AT, BE, CY, DE, DK, EE, ES, FI, FR, HR, IT, LT, MT, NL, NO, PT, SI, SK, SRB, SW, UK[ENG, SC, WL]	LV
2	*Non-health related administrative data sources ǂ*	Birth and mortality database, education level, income tax, GIS, occupation, housing conditions, criminal statistics, land and housing, socioeconomic, census (demographic), house of handicap persons, environmental, road and transport, air pollution, UV light exposure	22	BE, CY, CZ, DE, DK, EE, ES, FI, FR, HR, IT, LT, MT, NL, NO, PL, PT, SI, SK, SRB, SW, UK[ENG, SC, WL]	IE, LV
3	*Disease-specific registries*	Cancer, diabetes, cardiovascular, congenital malformation, tuberculosis, HIV/AIDS, inflammatory bowel disease, renal, reproductive health, dementia, organ transplantation, traffic accidents/trauma or injury, hospital registry of domestic and leisure accidents	22	BE, BG, CY, CZ, DE, DK, EE, ES, FI, FR, HR, IE, LV, MT, NL, NO, PL, PT, SK, SRB, SW, UK [ENG, SC, WL]	LV
4	*National health surveys**	National health examination and interview surveys	15	BE, CZ, DK, DE, EE, ES, FI, FR, IT, NL, NO, PT, SI, SW, UK [ENG, SC, WL]	PL
5	*Population-based epidemiological cohort/National cohorts*	DANCOS, IDEFICS, CONSTANCE, ELFE, Growing up in Scotland, HealthWise Wales cohort, Millennium cohort, Caerphilly cohort study	7	DK, EE, FI, FR, NO, PL, UK [ENG, SC, WL]	
6	*Clinical trials data*	FINGER, PRISMATIC	3	DK, FI, UK [ENG, WL]	

ʘ Latvia is developing data linkage techniques to link EHRs with other data sources.
ǂ In Ireland, income database is linked with EHRs of prescribing medicine at a local level. Latvia is developing data linkage techniques to link birth and mortality databases either with EHRs or with disease-specific registries.
** Poland is planning to link this national health survey data with other health data sources in near future. In Ireland, this is done for specific surveys such as housing and health conditions at local scale.*

These data sources are linked with each other in different combinations and some examples of various combinations used across member countries, are reported in *table 1.3*. These countries perform data linkage by using one of the following information: social security number, patient unique identification number, person unique pseudonymous identifier, encrypted personal identification number, citizen or national identification number. In Ireland, the lack of a unique patient identifier number limits the potential to link with different data sources.

Table 1.3: Examples of different combinations of data linkages across European countries in 2019 for innovative use of data sources

S/No	European countries	Different combinations of data linkages (N = ~ 85)
1	Austria	Hospital discharge with outpatient visit (primary care visit)
2	Belgium	Hospital discharge with health insurance claim
		Educational attainment with mortality database
		Census with mortality database
		Health interview survey with mortality (cause-specific mortality, StatBEL)/use of care (IMA)/prescribed medication (INAMI)
		Disease-specific registries with mortality database
3	Bulgaria	Registry of rare diseases with oncology registry
4	Croatia	Primary health care visits with hospital discharge/health insurance claim/mortality (cause-specific mortality)
		Cancer registry with geospatial registry
5	Cyprus	Hospital discharge with mortality database
		Cancer, diabetes, HIV/AIDS registries with mortality database
6	Czech Republic	Health insurance claims with mortality database
		Registry of hospitalizations is linked with mortality database
		Disease specific registries: Cancer, cardiovascular surgery and intervention, reproductive health, TBC registry, registry of injuries with mortality database/health insurance claims/registry of hospitalizations
7	Denmark	National patient health register is linked with education, income, housing, transfer payments, socioeconomic status, criminal statistics, etc.
8	Estonia	Hospital stay, primary and special ambulatory care linked with health insurance claims
		Health insurance claims and prescriptions are linked with causes-specific mortality
		Birth register linked with causes-specific mortality
		Cancer, tuberculosis and myocardial infarction linked with causes-specific mortality
		Chernobyl Cleanup workers (cancer, causes of death) families and children data with birth register (on irregular basis)
	(In progress)	Genomic database linked with EHRs
	(In progress)	Estonia health insurance database linked with prescription and diagnostics procedures
9	Finland	KANTA and KANSA health register linked with
		Finis birth cohort 87 and 97 linked with
		National HES from 1972-2017 linked with
		National HIS since 1978 onwards linked with
10	France	Hospital discharge linked with health insurance claims and mortality database (national health database: SNDS)
		Population-based epidemiology cohorts (CONSTANCES & ELFE) linked with national health database
		Cancer, congenital malformation, cardiovascular, inflammatory bowel disease and traffic accidents registries linked with national health database

		National health surveys (Esteban examination/interview) linked with national health database
	(In progress)	UV light and air pollution exposure linked with national health database
	(In progress)	House of handicap person s' health and social assistance linked with national health database
11	Germany	National health examination survey in adults linked with mortality database
		National health examination survey in adults linked with health insurance claims
		Cancer registry operated by the public health institute and included in health reporting
		National health surveys use national and sub-national data for weighting
		National health examination surveys use inter-metropolitan socioeconomic data for improvement of field work (in progress)
		Use of socioeconomic data at the metropolitan level for small area estimation (in progress)
		Use of real-time emergency room data for surveillance of infectious diseases (in progress in a local project)
		Linkage of data from national health surveys, health insurance data, cancer registry and other data sources for national burden-of-disease calculation (in progress)
12	Greece	No
13	Ireland (in progress)	Cancer registry linked with Hospital admission linked and mortality database
		Census data linked with mortality database (one off)
		Prescribed medication data Medical eligibility and claims data linked with income level (one off)
14	Italy	Hospital discharge linked with mortality database and national health examination survey
15	Latvia (in progress)	Hospital discharge, primary health care, emergency care records linked with birth and mortality database
		Patient register with specific diseases linked with mortality database
1.6	Lithuania	Compulsory health insurance information system (inpatient, outpatient specialized, primary care, emergency care) linked with causes-specific mortality database
17	Luxembourg	No
18	Malta	Health insurance claims, prescribed drugs, surgical operations, laboratory information system, radiology information system, patient administration system, outpatients attendance, patient discharge summaries linked with birth and mortality database
		Congenital anomalies, injuries, cancer, dementia, organ transplants registries linked with mortality database
19	The Netherlands	Health examination and interview surveys linked with mortality database
		Health insurance claims with perinatal data
		Cancer registry data with mortality database
20	Norway	Linkage between almost all sources by means of unique personal identification. Both within health and care services, and across other governmental areas. Big data solution in use for accessibility modulation using national health registries linked with land and housing, road and transport, and GIS databases.
21	Poland	Cancer and tuberculosis registry databases linked with mortality, demographic and GIS databases
	(in progress)	National health surveys linked with electronic health records
22	Portugal	Hospital discharge, primary care and medical records linked with hospital registry of domestic and leisure accidents, e-death certification
		Cancer, tuberculosis, HIV and congenital anomalies registries linked with e-death certification and hospital discharge data
23	Romania	No
24	Slovakia	National registry of EHRs (Hospital discharge, general practitioner record, referrals, prescribed medications, laboratory results, diagnostic procedures medical consultations) linked with national disease-specific registries
		National registry of EHRs linked with national registry of health care workers and heath care providers

25	Slovenia	Hospital discharge, drug prescription and perinatal health linked with mortality database
		Hospital discharge, drug prescription and perinatal health linked with census data on education and socioeconomic variables (inequality analysis)
		Hospital discharge, drug prescription and perinatal health linked with European Health Interview Survey
26	Serbia	Hospital discharge linked with cancer registry
		Mortality database linked with cancer registry
27	Spain	National health interview survey linked with mortality database
		Primary care data linked with drugs prescription and laboratory tests
		National Health survey linked with cause-specific mortality data
		All cohort studies can link their data with cause-specific mortality information through an agreement with the National Institute of statistics
28	Sweden	National Patients register linked with causes-specific mortality database
		National Patients register linked with birth database
		National Patients register linked with dental health database
		National Patients register linked with vaccination database
		National Patients register linked with education, income tax, occupation, country of origin and population based register
		National health surveys (ULF/SILC), environmental health survey (MHE) and European health interview survey (EHIS) linked national health care quality registries (each deals with a disease-specific condition)
29	UK-England	UK Cancer Registry is linked with Hospital and Mortality Records
	UK-Scotland	EHRs linked with each other: General and Psychiatric Hospital Stays/ day cases (including intensive care/high dependency stays), Outpatient attendances, Emergency department attendances, Maternity, birth records and Neonatal Care
		EHRs linked with mortality database and census (demographic) database
		Cancer and diabetes registries linked with hospital and mortality records
		Scottish health interview survey linked with hospital and mortality records
	UK-Wales	EHRs linked with each other: Primary care general practice datasets linked with hospital inpatient, emergency department visits, outpatient attendances, child health dataset, congenital anomalies, maternity records, population register and laboratory results in the Secure Anonymized Information Linkage (SAIL) database www.saildatabank.com
		EHRs linked with mortality database, GIS and census (demographic) database
		Cancer, trauma and renal registries linked with all of the above
		Welsh Health Survey and National Survey for Wales (interviews) linked with all of the above
		Healthwise Wales Cohort, Millennium Cohort, Caerphilly Cohort study and UK-Biobank linked to SAIL and all of the above
		EHRs linked Education Attainment records and GIS derived metrics e.g. pollution, housing quality, urban design, alcohol outlets etc.

General characteristics of a linked dataset: Our results showed that among 24 European countries who perform data linkage in routine, 17 do linkage at the national level *(Table 1.4)*. France, Portugal and Scotland do data linkage both at national and sub-national levels. Denmark, Germany, Norway and Sweden do data linkage at all levels. 23 countries either use the deterministic type of linkage (12 countries) or a combination of deterministic and probabilistic linkage (11 countries). Among 16/24 countries, linked data is available and is used in a routine manner. Among 12/24 countries, the registry owner (i.e., who governs the data register) provides the approval to access linked data. Among 15/24 countries, the accessibility to linked data is in routine or permanent whereas, in 13 countries, the accessibility could be ad-hoc or at intermittent basis depending on the project. Among 15/24 countries, linked data do not operate in real-time (i.e., integrate the updated information with minimum delay in time). Among 19/24 countries, linked data are flexible to integrate new variables.

Table 1.4: General characteristics of linked datasets in European countries in 2019 for innovative use of data sources

S/No	General characteristics of linked datasets		European countries		
			Advanced N = 24		In progress N = 2
1	Level of data linkage use/implementation				
		National level	17	AT, BE, BG, CY, CZ, EE, ES, FI, HR, LT, MT, NL, NO, SI, SK, SRB, UK (ENG, WL)	IE, LV
		Sub-national level	1	IT	
		Both (National and Sub-national) levels	3	FR, PT, UK-SC	
		Metropolitan level	4	MT, PL, SI, UK-WL	
		All of the above	4	DE, DK, NO, SW	
2	Type of linkage				
		Deterministic	12	AT, CY, HR, FI, LT, MT, NL, NO, SI, SK, SRB, SW	IE, LV
		Probabilistic	1	UK-SC	
		Combination of both (i.e., deterministic and probabilistic)	11	BE, CZ, DE, DK, EE, ES, FR, IT, PL, PT, UK (ENG, WL)	
		None of the above	1	BG	
3	Current status of linked data usage				
		Available and is used in routine	16	AT, BE, CY, CZ, DK, EE, FI, FR, LT, MT, NL, NO, PL, SI, SW, UK (ENG, SC, WL)	
		In progress of development	4	BG, ES, HR, PT	IE, LV
		Partial in use & partial in progress of development	2	DE, SK	
		Available but not in use	2	IT, SRB	
4	Type of approval to access				
		By law	5	AT, CZ, MT, NO, SW	
		By ethical committee	7	BE, ES, FR, IT, NO, PT, UK (ENG, SC, WL)	
		By register owner	13	BG, CY, ES, HR, FI, FR, IT, NO, PL, PT, SI, SK, SRB	
		Others (i.e., depend on linkage/data protection inspector/under conditions/not applicable (data linkage in safe environment)/by statistical authority	7	CZ, DE, DK, EE, ES, LT, NL	LV
5	Type of accessibility				
		In routine/permanent	15	BE, BG, CZ, DK, EE, FI, FR, NL, NO, PL, PT, SI, SK, SW, UK (ENG, SC, WL)	
		Ad-hoc/Intermittent	13	BE, CY, DE, EE, ES, HR, IT, FR, MT, NO, PT, SK, SRB	
		Under conditions (i.e., restricted to certain projects for a limited period)	6	AT, EE, ES, FR, LT, NO	LV
6	Operate in real-time				
		Yes	10	DK, EE, FI, FR, LT, NO, SI, SK, SW, UK-SC	
		No	15	AT, BE, BG, CY, CZ, DE, ES, HR, IT, MT, NL, PL, PT, SRB, UK (ENG, WL)	IE, LV
7	Flexible to integrate new variables				

		Yes	19	AT, BG, CY, CZ, DE, DK, ES, FI, FR, HR, MT, NL, NO, PL, PT, SI, SK, SW, UK (ENG, SC, WL)	
		No	5	BE, EE, IT, LT, SRB	IE, LV

Perspectives and opportunities of data linkages

Some projects on data linkage are in progress (i.e., in the next five years) to integrate this technology in their routine public health activities in following member states: Austria, Cyprus, Czech Republic, Ireland, Italy, Latvia, Norway, Poland, Portugal, and Spain.

Austria: Since 2015 the Austrian Diagnosis Related Group (DRG) system, in German "Leistungsorientierte Krankenanstaltenfinanzierung" – LKF ("Procedure-oriented Hospital Financing") has been extended to ambulatory care sector named as "Cross-sector documentation". Visits of patients are linked by a pseudonym identification and using these pseudonym identifiers, patient pathways can be analyzed. However, there are limitations, for example, no diagnosis is recorded in the ambulatory sector.
Cyprus: Currently, the linkage is performed manually, as there is no electronic linkage available among databases (registries).
Czech Republic: Data linkage is possible by law within the National Health Information System (NHIS).
Ireland: The Health Research Board (HRB) is funding a pilot project to design and develop the infrastructure needed to share and link health data securely. This two-year project will design and build a prototype technical infrastructure to demonstrate how secure, controlled access for researchers to health and social care datasets can be implemented in a safe environment for new types of data analyses that have not been possible in the past. The goal is to lay the foundation for a national infrastructure for data access, sharing, storage and linkage of sensitive health and social care data in line with legal and ethical requirements and provide guidelines for the upscaling of the model [16].
Italy: A pilot project was performed during the period of 2016-2017 to link current health and administrative data from the Lombardia region with longitudinal epidemiological cohort studies carried out at the National Health Institute (Istituto Superiore di Sanità) in Rome.
Latvia: There is an ongoing development of the public health monitoring system to develop a transparent and efficient healthcare quality evaluation system. The database contains pseudonymous patients' data and identifiable service providers (institutions and physicians) to link different data sources.
Norway: Registries for linkage contain some of the major health information resources. The program "Health Data Platform" was launched in 2018, and aims to collect data

from more than 50 national health registries and medical quality registries in one advanced analytical platform, available for government, research and industry.
Poland: Ministry of Health Analyses and Strategy Department link national health fund data (payer) with different sources as demographics data, social insurance data, etc. by ourselves. Key elements to link data are discussed first and checked with data owners for verification of accuracy.
Portugal: A probabilistic linkage is in a testing phase for two following projects: first is the linkage between congenital anomalies registry data and hospital admissions database and second is the linkage between domestic/leisure accidents' system and hospital admissions database. Some initiatives to link Cancer Registries with different databases such as mortality, hospitalization and population data have been taken.
Spain: At the national level, primary health care data i.e., BDCAP (Base de Datos Clínicos de Atención Primaria) is linked with laboratory & image test results, as well as pharmacological prescription, and it is being already used to provide data on several cardiovascular risk factors. An ongoing case-study is about linking non-health related databases with geo-located health indicators. At the regional level, some Autonomous Communities (i.e. Comunidad Valenciana, Madrid or Cataluña) have linked health care data from primary and specialized care for their inhabitants, and these combined databases that in many cases operate in real-time are being used for surveillance and research.

C. Use of artificial intelligence (AI) in routine public health activities

The use of AI is not frequent across European countries *(Table 2)*. Only five countries have reported applying following techniques in routine public health activities: machine learning (Denmark, Finland, Sweden, and UK-Wales), natural language processing (Finland, Sweden, and UK-Wales), Markov decision process (Finland), support vector machine (Finland, UK-Wales), data mining (Finland) and TSP [Travelling Salesman Problem] modelling (Norway). Denmark can apply these techniques not only at a national level but also at a metropolitan level.
There are ongoing projects on the use of the AI (i.e., in the next five years) to integrate this technology in routine public health activities in the following countries: Croatia, Czech Republic, France, Germany, Norway, Portugal, and Spain. The objectives of these initiatives are for epidemiological research and surveillance of non-communicable and communicable diseases estimating the prevalence and prediction of incidences of certain health conditions at various geographical levels.
Two countries mentioned that due to lack of human resources (Lithuania) and capacities/skills (Republic of Serbia) within their public health institutes, AI techniques are not applied in routine public health activities.
Some European countries also mentioned the use of classical statistical techniques without the use of AI *(Table 2)*.

Table 2: Current status of European countries using artificial intelligence in routine public health activities in 2019

Use of Artificial Intelligence (AI)			
	Advanced **N = 5**	**In progress** **N = 9**	**Not yet** **N = 16**
European countries	DK, FI, NO, SW, UK-WL	AT, CZ, DE, ES, FR, HR, PL, PT, SK	BE, BG, CY, EE, GR, IE, IT, LT, LU, LV, MT, NL, RO, SL, SRB, UK (ENG, SC)
Level of application of AI			
National level	DK, FI, NO, SW, UK- WL		
Sub-national level			
Metropolitan level	DK, SW		
Use of classical statistics without the use of AI			
	Advanced **N = 19**	**In progress** **N = 5***	**Not yet** **N = 8**
European countries	BE, BG, CZ, DE, EE, ES, DK, FR, FI, IT, MT, NL, NO, PL, PT, SI, SK, SW, UK (ENG, SC, WL)	AT, CZ, ES, HR, SK	CY, GR, IE, LT, LU, LV, RO, SRB
Level of use of classical statistics without AI			
National level	BE, BG, CZ, DK, EE, FR, FI, IT, NL, NO, PL, PT, SK, SW, UK- WL		
Sub-national level	DE, ES, IT, PL, NO, SI, UK (ENG, SC)		
Metropolitan level	DK, MT, NO		
**Two countries (CZ & SK) use classical statistic in routine (i.e., advanced) but also developing further this technology (i.e., in progress)*			

Perspectives and opportunities of using advanced statistics

Some projects on the use of advanced statistics are in progress (i.e., in the next five years) to integrate this technology in routine public health activities in the following member states: Croatia, Czech Republic, France, Germany, Norway, Portugal, and Spain.

Croatia: Intermittent collaborations within small initiatives and projects such as proactive advising: a machine learning-driven approach was applied to predict vaccine hesitancy.

Czech Republic: Advanced statistical modelling is being used for prediction in epidemiology, further applications in epidemiology and health services research are expected in the near future.

France: Currently, some pilot studies are ongoing to use AI predicting health outcome indicators using individual-level data.

Germany: Several projects are underway to include machine learning and AI procedures in infectious disease surveillance, including a local project involving the use of real-time emergency room data, projects aimed at detecting novel pathogens, and a project on benchmarking machine-learning approaches for outbreak detection. Furthermore, machine learning and AI procedures to predict mental health indicators are being tested, e.g. employing data on media usage collected from electronic devices.

Norway: The Norwegian government decided to launch [a national strategy] for artificial intelligence during the autumn of 2019.
Portugal: Statistical models are in place and are used to compute daily and weekly heat and cold air temperature-related risk of death which are used for monitoring and early warning purposes.
Spain: In spatial epidemiology, advanced statistical modelling is being used to estimate the risk of death by disease at the municipal level (http://ariadna.cne.isciii.es/MapaM/).

D. Health indicators estimated using linked data

Using linked data, the majority of European countries estimate the following health indicators:

Health outcome indicators
Participants were asked to select at least three health conditions and to report the related health outcome indicators which are most important for public health in their country. Using linked data, 46 health outcome indicators related to following seven health conditions were reported from 22 countries: *cardiovascular* (14), neurodegenerative disease (6), maternal and perinatal health (6), diabetes (6), suicide/trauma/injury (7), cancer (6) and hepatic failure (1) *(Table 3.1)*. The main objectives to estimate these indicators were for public health monitoring and research purposes and the level of estimation was mainly at national and sub-national levels.

Table 3.1: Description of health outcome indicators estimated using linked data across European countries in 2019

S/No	Categories	Health outcomes indicators (N = 46)	European countries
1	Cardiovascular (N = 14)		
	Health characteristics	Incidence of stroke among less than and more than 65 years old	CZ, FR, LT, NL, NO, SW, UK-WL
		Prevalence of stroke among less than and more than 65 years old	CZ, LT, NL, NO, SW, UK-WL
		Incidence of myocardial infarction among less than and more than 65 years old	CZ, FR, LT, NL, NO, SW, UK-WL
		Prevalence of myocardial infarction among less than and more than 65 years old	CZ, LT, NL, NO, SW, UK-WL
	Mortality	Mortality due to stroke within 30-days of hospitalization	FR, IT, LV, MT, NO, SW
		Mortality due to myocardial infarction within 30-days of hospitalization	CZ, FR, NO, SW
		Risk of mortality due to myocardial infarction at the municipal level	SW
		Risk of mortality due to myocardial infarction at the municipal level (in future)	ES
		Neurorehabilitation and functional outcomes of patients after stroke	FR, SW

	Human function and quality of life	Neurorehabilitation and functional outcomes of patients after stroke (in progress)	NO
	Life expectancy and well-being	30- days survival following stroke	FR, NO, SW, UK-WL
		90-days survival following stroke	SW, UK-WL
		365-days survival following critical care for stroke	
		Hospital utilization in the 365 days following intensive care discharge	
		365-days survival following stroke	FR
	All of the above categories	DALYs, YLL, YLD	SW, UK-SC
2	**Neurodegenerative disease (N = 6)**		
	Health characteristics	Prevalence of Multiple Sclerosis	CZ, HR, FR, LT, NO, SW, UK-WL
		Prevalence of Alzheimer	CZ, FR, LT, SW, UK-WL
		Incidence of Alzheimer	FR, LT, SW, UK-WL
		Incidence and prevalence of Alzheimer (in progress)	NO
		Prior event rate ratio to estimate the influence of exposure to antipsychotic medication on acute cardiac events and hip fracture due to dementia	SW, UK-WL
	Mortality	Mortality due to Parkinson	FR, SW
		Mortality due to Dementia	FR
3	**Maternal and perinatal health/child health (N = 6)**		
	Health characteristics	Incidence of low birth weight	CY, CZ, PL, FR, LT, NO, SW
		Incidence of pre-term birth rate	SW
		Incidence of pre-term birth rate (in future)	CZ, FR, NO, UK-WL
		Incidence of gestational diabetes	SW
		Incidence of gestational diabetes (in future)	FR, NO
		Prevalence of congenital anomalies	CZ, PT, LT, NO, SW, UK-WL
		Emergency admissions for potentially preventable hospitalizations (PPH) between the age of 1 and 5 years	UK-WL
	Mortality	Stillbirth	FR, NL, LT, NO
4	**Diabetes (N = 6)**		
	Health characteristics	Incidence of diabetes	CZ, FR, LT, SW, UK-WL
		Prevalence of diabetes	
		Incidence and prevalence of diabetes (in progress)	NO
	Mortality	Mortality due to diabetes and related risk factors	BG, CZ, DE, FR, SW
	Human function and quality of life	Amputation rate (related complications)	BE, FR, MT, LT, SW
	Human function and quality of life	Number of patients with installed insulin pump during diabetes curation	FR, PL, SW
	All of above categories	DALYs, YLL, YLD	SW, UK (SC, WL)
5	**Suicide/Trauma/Injury (N = 7)**		
	Health characteristics	Use of health care services before suicide	LT, SW, UK-WL
		Prevalence of morbid conditions before suicide	
		Standardized prevalence of suicide	CZ, PL, FI, LT, NO, SW
		Incidence of suicide	
		Risk of road accident among users of prescribed medicines	FR
		Incidence of injuries	LT, NO, SW, UK-WL

	Mortality	Death rates in road accidents	EE, LT, NO, SW
6	**Cancer (N = 6)**		
	Health characteristics	Incidence of various types of cancer	CY, CZ, DE, FI, FR, PL, NL, NO, SW, UK-WL
		Prevalence of various types of cancer	
		Incidence rate by stage for colorectal cancer	CZ, MT, NO, SW, UK-WL
	Mortality	Mortality rates due to various types of cancer	CY, CZ, ES, FI, IT, PT, LT, NO, SW, UK-WL
	Life expectancy and well-being	5-years relative survival rates	CY, CZ, DE, EE, FI, IT, PT, NL, NO, SK, SW
	Human function and quality of life	Scale of return to work after cancer and determining factors	SW
		Scale of return to work after cancer and determining factors (in future)	BE
7	**Alcoholic liver disease and hepatic failure (N = 1)**		
	Mortality	Standardized mortality ratios at 60 days and 5-years following unscheduled admissions	SW, UK-WL

Health determinants

For the health determinants, participants were asked to report the corresponding determinants of the identified health conditions. 34 health determinants related to various health conditions were reported by 15 member states *(Table 3.2)*. These determinants are related to the physical environment (12), socioeconomic and environment (10), health behavior and lifestyle (6) and biological and metabolic parameters (3) *(Table 3.2)*. These determinants were used to measure the potential associations between these risk factors and health conditions for public health monitoring and research purposes. These determinants can be stratified by age, sex, socioeconomic status and by area of residence.

Table 3.2: Description of health determinants using linked data across European countries in 2019

S/No	Domain/Health condition	Health determinants (N = 34)	Variables can be stratified by	European countries
1	**Physical environment (N = 12)**			
	Emphysema	Air quality	Area of residence	BE, UK-WL
	Injury	Place of injury	Age, sex and area of residence	CY, NO, UK-WL
		Type of injury		
		After injury hospitalized or not		
	Parkinson	Exposure to pesticides (i.e., agricultural activities, in vineyards, metallurgy and solvents, in textile industry)	Area of residence	FR
	Breast cancer mortality	Industrial pollution	Area of residence	ES
	Adiposity	Proximity of fast food outlets from areas of residence	Area of residence	UK (ENG, WL)

	Various chronic health conditions	Alcohol outlet density	Area of residence	NO, SW, UK (SC, WL)
	Mental health	Presence to green-blue spaces		
		Access and visit to green-blue spaces		
		Visit to green-blue spaces		
		Housing quality		
2	**Socioeconomic and environment (N = 10)**			
	Multi-morbidity	Number of single households of older people	Age and living condition	AT, NO, SW
	Breast cancer mortality, injury, diabetes, cardiovascular, mental health	Sociodemographic status	Age, sex and area of residence	EE, ES, FR, NL, NO, PT, SI, SW, UK (SC, WL)
		Socioeconomic status		FR, NO
		Employment status		FR
		Level of education achieved		FR, NO
		Deprivation index		FR
	Pre-term birth	Maternal education to measure social disparities		FR, NO
	Injury	Time and distance between road accident and emergency room		PL
		Standardized absenteeism and attributable indirect costs		
	All types of cancer	Accessibility to linear accelerators for radiotherapy	Area of residence	
3	**Health behavior and life style (N = 6)**			
	Stroke, myocardial infarction, lung cancer, mental health, obesity, other chronic condition	Smoking rate	Age, sex, socioeconomic status and area of residence	BE, CY, FR, IT, MT, NL, NO, UK (SC, WL)
		Alcohol consumption	Age and sex	NO, UK-SC
		Physical activity		UK-SC
		Dietary consumption		UK-SC
		Drug use		NO, UK-SC
	Diabetes	Diabetes risk score	Age, sex and area of residence	CY, NO, SL
4	**Biological/metabolic parameters (N = 3)**			
	Obesity	Self-reported BMI	Age, sex and area of residence (i.e., in county, municipality)	SW, NO, UK-SC
	Diabetes	Blindness	Age, sex and area of residence	CY, FR
		Proteinuria		
5	**Others (N=3)**			
	Road accidents, neurodegenerative disease	Multi-morbidity	Age and sex	FR
	Chronic health conditions	Disability		FR
		Frailty		FR

Health intervention indicators

Participants were asked to report at least three health intervention indicators under three categories (i.e., prevention, promotion, others) corresponding to the given health conditions which are most important for public health in their country. Using linked data, 23 health intervention indicators related to following six health conditions were reported from 17 member states: maternal and perinatal health (7), cancer (6), diabetes (4), cardiovascular (2), neurodegenerative disease (2), suicide/trauma/injury (1) and lower/upper respiratory infections (1), *(Table 3.3)*. The main objectives to estimate these indicators were to guide the health policy process, public health monitoring, and research purposes. These intervention indicators are estimated mainly at national and sub-national levels and currently are in use.

Table 3.3: Description of health intervention indicators estimated using linked data across European countries in 2019

S/No	Categories	Domain/Health condition	Health intervention indicators (N = 23)	Member States
1	**Maternal and perinatal health (N = 7)**			
	Prevention	Low birth weight	Prevalence of thyroid gland examination during pregnancy	CZ
	Prevention		Frequency of admission to intensive care unit	CY
	Prevention		Prevalence of maternal smoking and quitting smoking during the pregnancy	FI, FR, NO, SW
	Prevention	Pre-term birth	Percent of births in level III maternity units	FR, NO, SW
	Prevention	Perinatal mortality	Pregnant women with adequate prenatal care (number of visit/timing of initiation)	FR, NO, SW
	Prevention	Prenatal care	Screening programs as preventive check-ups during pregnancy	NO, SI, SW
	Prevention	Neural tube defect	Folic acid supplementation	PT, NO, UK-WL
2	**Cancer (N = 6)**			
	Prevention	Breast, cervical, colorectal and bowl cancer	Screening participation rates, effectiveness and evaluation	BE, EE, CZ, ES, FI, HR, IT, NO, SI, SW, UK-WL
	Prevention	Colorectal cancer	Frequency of surgery	CY, NO, SW
	Prevention	Colorectal cancer	Colonoscopy compliance rate	CZ, SI, NO
	Prevention	Breast cancer	Genetic screening among families (in future)	ES
	Prevention	Breast cancer	Stage distribution of detected cancer	CZ, SI, NO
	Others	All types of cancers	Re-integration in work	BE
3	**Diabetes (N = 4)**			
	Prevention	Diabetes related complications	Foot care	BE, FR, SW
	Prevention		Proportion of diabetics counselled by nurse to avoid complications	SW
	Prevention		Proportion of diabetics counselled by nurse to avoid complications (in future)	CY
	Prevention		Amputation rate	BE, FR, SW
	Prevention	Diabetes	Percentage of diabetics with latest HbA1c above 7.0	FR, MT, SW
4	**Cardiovascular diseases (N =2)**			
	Prevention	Stroke, myocardial infarction	Absolute global CVD risk assessment in primary prevention	IT
	Prevention	Stroke	Aortic aneurysm screening	SW, UK-WL
5	**Neurodegenerative disease (N = 2)**			
	Prevention	Multiple sclerosis	% of patients qualified for pharmacotherapy	PL
	Prevention	Dementia	% of patients using neuroleptic drugs	FR
6	**Trauma/Injury/Suicide (N = 1)**			

	Prevention	Injury/Trauma	Visit to primary care physicians before suicide	SI, LT
7	Lower/ Upper respiratory infections (N = 1)			
	Prevention	COPD (Chronic Obstructive Pulmonary Disease)	% of patients with non-invasive ventilations	PL, SW

V. Discussion

A. Main results

The results of this study showed variability in the use of data linkage and the AI at the national institutes of public health and health information and statistics across European countries. The majority of countries use data linkage in routine by applying either deterministic or a combination of two types of linkages (i.e., deterministic & probabilistic) for public health surveillance and research purposes. The use of a universal unique identifier, social security number or unique pseudonymous identifier is common to applying deterministic linkage technique among European countries. The use of AI is not frequent to estimate health indictors at national public health institutes. Across European countries, using data linkage, 46 health outcome indicators related to seven health conditions, 34 related determinants and 23 health intervention indicators were reported. Some initiatives are ongoing as pilot projects to apply these techniques to improve health surveillance and to guide the health policy development process.

A systematic review has shown some practices applied for data linkage in the field of perinatal health across Europe for health surveillance and research purposes [9]. Several other studies have explored various dynamics of population health such as social care, psychotic disorders, multi-morbidity, diabetes, obesity, mental health, cardiovascular, antibiotic use and Alzheimer using data linkage with different types of administrative data sources (both related to health and non-health) [7, 17-29]. For the surveillance of cancer, data linkage not only provides the opportunity to improve population-based screening [30] but also helps in detecting different types of cancer recurrence [31] and evaluation of the socio-economic status of patients with cancer (e.g., return to work) [32]. Linked data also allows evaluating the interventions at various levels of the population [33]. The diversity and the volume of health information have been increasing rapidly and push to discover new parameters to improve population health with innovative approaches. In that context, some initiatives have been launched at the national levels to create health data hub/platform to be used for research and to guide the policy development process [34, 35].

There are some studies available which have discussed the advantages of using AI in early detection and diagnosis of certain conditions, treatment, as well as outcome prediction and prognosis evaluation with high precision [36, 37] but their use in population health to estimate and predict health indicators remain limited [38].

Our results highlighted that a few member countries have achieved the most advanced level in data linkage by linking health information (i.e., clinical, biobanks/laboratory tests, genetics) with education, occupation, housing quality, air pollution, criminal statistics and transport/road accidents, etc. This offers exceptional opportunities to enrich information and to perform epidemiological research, health surveillance and consequently, guiding health policies to improve population health.

B. Main obstacles and recommendations

However, the majority of European countries have not reached that level in data linkage and the use of AI underlined following four main obstacles associated with the implementation and the use of data linkage and advanced statistics: 1. The complex laws and data protection regulations which block linkage between different data sources with a deterministic approach *(legal)*, 2. Lack of human resources and capacities/skills within national institutes of public health and health information statistics *(technical)*, 3. Lack of governance of health information *(data governance)* and 4. Limited resources to support the health information infrastructure *(organization and structural)*.
To address these gaps, we propose the following *recommendations: A. Legal aspects:* 1. more flexible data governance frameworks to support data linkage of different data sources should be encouraged [39], 2. Specific mandates to ensure data availability/access/capture and safe storage should be an integral part of the health information system, 3. Differences in implementation and the interpretation of the EU GDPR (General Data Protection Regulations) and additional national regulations should be mapped and if possible harmonized across EU-MSs [40]; *B. Technical aspects:* 4. more collaborations and partnerships should be encouraged to build up capacities for use of new technology, to share new methods, skills, experiences and data for comparative research studies among EU national institutes of public health and health information statistics; *C. Governance*, 5. Initiatives to strengthen national health information infrastructure should be encouraged; *D. Organizational and structural aspects*, 6. ministries of health and research in a member country should provide their support to develop national health data hubs/data platform to strengthen the national health information infrastructure.

C. Limitations

There are a few limitations in this study. *First,* current practices of data linkages at national institutes of public health and health information and statistics we surveyed may differ from other research institutes in that country which we did not cover and might influence the results of this study. However, this survey provides the latest overview of current practices in data linkage and highlights the related obstacles in performing data linkage. *Second,* we limited the response burden of health indicators to three priority health conditions. Therefore, our results do not constitute an exhaustive list of health indicators that are used in the country to inform policy and practice. It may limit the number of health indicators being estimated using linked data and advanced statistics.

VI. Conclusions

To our knowledge, this is the first study that provides information about the current practices of data linkage and the AI at the national institutes of public health and health information and statistics across European countries. Our results highlight that the majority of the countries have integrated data linkage in routine public health activities but few use the AI. The European countries who are advanced in using both techniques data linkage and the AI could guide others by an exchange of their experiences and examples of good practices. A sustainable national health information system and data governance framework to link different data sources are essential to support evidence-based practices for the health policy development process. Building analytical capacity in national institutes of public health and health information and statistics is necessary for improving the utilization of linked data in order to improve the monitoring of public health activities. These results ultimately contribute to strengthen the national health information system and would facilitate moving towards establishing an integrated EU-Health Information System.

VII. References

1. Harron K, Dibben C, Boyd J, Hjern A, Azimaee M, Barreto ML, Goldstein H: Challenges in administrative data linkage for research. *Big Data Soc* 2017, 4(2):2053951717745678-2053951717745678.
2. Ferrante A: The Use of Data-Linkage Methods in Criminal Justice Research: http://www5.austlii.edu.au/au/journals/ClCrimJust/2009/3.html. *Criminal Justice* 2009.
3. GF R: Administrative and claims records as sources of health care cost data. *Med Care* 2009, 47(7 Suppl 1).
4. Charlton RA, Neville AJ, Jordan S, Pierini A, Damase-Michel C, Klungsøyr K, Andersen A-MN, Hansen AV, Gini R, Bos JHJ *et al*: Healthcare databases in Europe for studying medicine use and safety during pregnancy. *Pharmacoepidemiology and Drug Safety* 2014, 23(6):586-594.
5. Thygesen LC, Ersbøll AK: When the entire population is the sample: strengths and limitations in register-based epidemiology. *European Journal of Epidemiology* 2014, 29(8):551-558.
6. WHO: Public Health Surveillance: https://www.who.int/topics/public_health_surveillance/en/.
7. Lloyd K, McGregor J, John A, Craddock N, Walters JT, Linden D, Jones I, Bentall R, Lyons RA, Ford DV *et al*: A national population-based e-cohort of people with psychosis (PsyCymru) linking prospectively ascertained phenotypically rich and genetic data to routinely collected records: Overview, recruitment and linkage. *Schizophrenia Research* 2015, 166(1):131-136.
8. Bradley CJ, Penberthy L, Devers KJ, Holden DJ: Health Services Research and Data Linkages: Issues, Methods, and Directions for the Future. *Health Services Research* 2010, 45(5p2):1468-1488.
9. Delnord M, Szamotulska K, Hindori-Mohangoo AD, Blondel B, Macfarlane AJ, Dattani N, Barona C, Berrut S, Zile I, Wood R *et al*: Linking databases on perinatal health: a review of the literature and current practices in Europe. *European Journal of Public Health* 2016, 26(3):422-430.
10. Joint Action on Health Information: https://www.inf-act.eu/. 2018.
11. EuroREACH: EuroREACH Framework: http://hdn.euhs-i.eu/performance/frameworks/euroreach-framework. 2013.
12. OECD: Health at Glance (OECD Indicators): https://www.health.gov.il/PublicationsFiles/HealthataGlance2017.pdf. 2007.
13. Navigator HD: EuroREACH Framework: http://hdn.euhs-i.eu/performance/frameworks/euroreach-framework. 2013.
14. HBM4EU: Linking HBM, health surveys and registers: https://www.hbm4eu.eu/deliverables/. 2018.
15. Eurociss: Cardiovascular Indicators Surveillance Set: https://ec.europa.eu/health/ph_projects/2000/monitoring/fp_monitoring_2000_frep_10_en.pdf. 2000.
16. Health Rsearch Board: https://www.hrb.ie/fileadmin/publications_files/Proposals_for_an_Enabling_Data_Environment_for_Health_and_Related_Research_in_Ireland.pdf. 2019.
17. Lyons RA, Jones KH, John G, Brooks CJ, Verplancke J-P, Ford DV, Brown G, Leake K: The SAIL databank: linking multiple health and social care datasets. *BMC Medical Informatics and Decision Making* 2009, 9(1):3.
18. Tuppin P, Rudant J, Constantinou P, Gastaldi-Menager C, Rachas A, de Roquefeuil L, Maura G, Caillol H, Tajahmady A, Coste J *et al*: Value of a national administrative database to guide public decisions: From the systeme national d'information interregimes

de l'Assurance Maladie (SNIIRAM) to the systeme national des donnees de sante (SNDS) in France. 2017(0398-7620 (Print)).

19. Chan Chee C, Chin F, Ha C, Beltzer N, Bonaldi C: Use of medical administrative data for the surveillance of psychotic disorders in France. *BMC Psychiatry* 2017, 17(1):386.
20. Rodgers Se Fau - Bailey R, Bailey R Fau - Johnson R, Johnson R Fau - Poortinga W, Poortinga W Fau - Smith R, Smith R Fau - Berridge D, Berridge D Fau - Anderson P, Anderson P Fau - Phillips C, Phillips C Fau - Lannon S, Lannon S Fau - Jones N, Jones N Fau - Dunstan FD *et al*: Health impact, and economic value, of meeting housing quality standards: a retrospective longitudinal data linkage study *Public Health Research* 2018.
21. Violán C, Foguet-Boreu Q, Hermosilla-Pérez E, Valderas JM, Bolíbar B, Fàbregas-Escurriola M, Brugulat-Guiteras P, Muñoz-Pérez MÁ: Comparison of the information provided by electronic health records data and a population health survey to estimate prevalence of selected health conditions and multimorbidity. *BMC Public Health* 2013, 13(1):251.
22. Fuentes S, Cosson E, Mandereau-Bruno L, Fagot-Campagna A, Bernillon P, Goldberg M, Fosse-Edorh S, Group C-D: Identifying diabetes cases in health administrative databases: a validation study based on a large French cohort. *International Journal of Public Health* 2019, 64(3):441-450.
23. Orriols L, Delorme B, Gadegbeku B, Tricotel A, Contrand B, Laumon B, Salmi L-R, Lagarde E, on behalf of the Crg: Prescription Medicines and the Risk of Road Traffic Crashes: A French Registry-Based Study. *PLOS Medicine* 2010, 7(11):e1000366.
24. Mason KE, Pearce N, Cummins S: Associations between fast food and physical activity environments and adiposity in mid-life: cross-sectional, observational evidence from UK Biobank. *The Lancet Public Health* 2018, 3(1):e24-e33.
25. Cleland B, Wallace J, Bond R, Black M, Mulvenna M, Rankin D, Tanney A: Insights into Antidepressant Prescribing Using Open Health Data. *Big Data Research* 2018, 12:41-48.
26. Gabet A, Danchin N, Puymirat E, Tuppin P, Olié V: Early and late case fatality after hospitalization for acute coronary syndrome in France, 2010–2015. *Archives of Cardiovascular Diseases* 2019, 112(12):754-764.
27. Williamson E DS, Morris S, Clarke CS, Thomas M, Evans H et al Risk of mortality and cardiovascular events following macrolide prescription in. *Rhinology* 2019, 57(4):252-260.
28. Hopkins C WE, Morris S, Clarke CS, Thomas M, Evans H, Little P et al Antibiotic usage in chronic rhinosinusitis: analysis of national primary care. *Rhinology* 2019, 6(10):136.
29. Ponjoan A G-OJ, Blanch J, Fages E, Alves-Cabratosa L, et al How well can electronic health records from primary care identify Alzheimer's. *Clin Epidemiol* 2019, 11:509-518.
30. Májek O, Anttila A, Arbyn M, van Veen E-B, Engesæter B, Lönnberg S: The legal framework for European cervical cancer screening programmes. *European Journal of Public Health* 2018, 29(2):345-350.
31. Hassett MJ, Uno H, Cronin AM, Carroll NM, Hornbrook MC, Ritzwoller D: Detecting Lung and Colorectal Cancer Recurrence Using Structured Clinical/Administrative Data to Enable Outcomes Research and Population Health Management. *Medical Care* 2017, 55(12):e88-e98.
32. Kiasuwa-Mbengi RL, Nyaga V, Otter R, de Brouwer C, Bouland C: The EMPCAN study: protocol of a population-based cohort study on the evolution of the socio-economic position of workers with cancer. *Archives of Public Health* 2019, 77(1):15.
33. Lyons RA, Turner S, Lyons J, Walters A, Snooks HA, Greenacre J, Humphreys C, Jones SJ: All Wales Injury Surveillance System revised: development of a population-based system to evaluate single-level and multilevel interventions. *Injury Prevention* 2016, 22(Suppl 1):i50-i55.
34. Health FMo: Health Data Hub: https://www.health-data-hub.fr/?lang=en. 2019.
35. Government W: Welsh Government takes innovative approach to policy making in Wales: https://www.swansea.ac.uk/press-office/news-archive/2019/welshgovernmenttakesinnovativeapproachtopolicymakinginwales.php. 2019.

36. Patel VL, Shortliffe EH, Stefanelli M, Szolovits P, Berthold MR, Bellazzi R, Abu-Hanna A: The coming of age of artificial intelligence in medicine. *Artificial Intelligence in Medicine* 2009, 46(1):5-17.
37. Jha S, Topol EJ: Adapting to Artificial Intelligence: Radiologists and Pathologists as Information Specialists. *JAMA* 2016, 316(22):2353-2354.
38. Flaxman AD, Vos T: Machine learning in population health: Opportunities and threats. *PLOS Medicine* 2018, 15(11):e1002702.
39. 2020 C: Data Protection and Artificial Intelligence: https://edps.europa.eu/data-protection/our-work/subjects/artificial-intelligence_en.
40. EPRS: How the General Data Protection Regulation changes the rules for scientific research: https://www.europarl.europa.eu/RegData/etudes/STUD/2019/634447/EPRS_STU(2019)634447_EN.pdf. 2019.

VIII. List of abbreviations

AI: Artificial Intelligence

EU: European Union

EEA: European Economic Area

MSs: Member States

InfAct: Information for Action i.e., a joint action of Member States to establish a sustainable European health information system.

WP: Work Package

Euro-REACH: It is an international collaboration to improve access to health care data through cross-country comparisons.

OECD: Organization for Economic Co-operation and Development

HBM4EU: Human Biomonitoring for Europe

EUROCISS: European Cardiovascular Indicators Surveillance Set

GDPR: General Data Protection Regulations

IX. Acknowledgments

This report is drafted by Romana HANEEF under the supervision of Anne GALLAY, Department of Non-Communicable Disease and Injuries, Santé Publique France (SpF), France.

We acknowledge the kind support of Jessica Gane (Data Department, Santé Publique France) for developing the web-based survey. We especially acknowledge Jean-Claude Desenclos and Yann Le Strat for their kind support in developing the survey questionnaire.

We especially acknowledge the kind support following partners for finalizing this report: Marie Delnord (Sciensano, Belgium), Emmanuelle Bauchet (SpF, France), Herman Van Oyen (Sciensano, Belgium), Zeynep Or (IRDES, France), Rita Gaidelyte (Health Information Center, Lithuania), Beatriz Pérez-Gómez (ISCIII, Spain), Luigi Palmieri (National Institute of Health, Italy), Peter Achterberg (RIVM, The Netherlands), Mariken Tijhuis (RIVM, The Netherlands), Metka Zaletel (NIJZ, Slovenia), Stefan Mathis-Edenhofer (The Austrian National Public Health Institute, Austria), Ondřej Májek (National Institute of Health information and statistics, Czech Republic), Michel Vernay (SpF, France), Håkon Haaheim (The Norwegian Directorate of Health, Norway), and Hanna Tolonen (THL, Finland).
We acknowledge following partners who contributed to complete this survey: Bulgaria: Raina Nikolova (Public Health Institute), Croatia: Ivan Pristaš (CIPH, Cyprus: Vasos Scoutellas (Public health institute), Czech Republic: Šárka Daňková, Jiří Jarkovský (Institute of health information and statistics of Czech Republic), Denmark: Mette Bjerrum Koch (Danish Health data authority), Maja Bæksgaard Jørgensen (Statens Institute for Folkesundhed), Estonia: Eleri Lapp (MoSH), Finland (Mike Gissler (THL), Germany: Angelika Schaffrath Rosario (RKI), Greece: Spyridon Goulas (Governmental Health Insurance Organization), France: Jennifer Zeitlin (INSERM U1153), Sandrine Fosse (Santé Publique France[SPF]), Laure Carcaillon-Bentata, (SPF), Valérie Olie (SPF), Louis-Marie Paget (SPF), Viet Nguyen-Thanh (SPF), Ireland: Sheona Gilsenan (DoH), Italy: Brigid Unim (ISS), Latvia: Janis Misins (CDPC), Luxembourg: Anne-Charlotte Lorcy (Health Directorate), Malta: Neville Calleja (MFH), Norway: Heidi Lyshol (Norwegian Institute of Public Health), Inger Johanne Bakken (Norwegian Directorate of Health, Poland: Jakub Adamski (MoH), Piotr Nowosielski (MoH), Portugal: Paulo Nogueira (DGS), Martin José (Public Health Institute), Carlos Dias (Public health Institute), Romania: Silviu Radulescu (Public Health Institute), Serbia: Maja Krstic (Batut), Slovakia: Jan Cap (National Health Information Centre), Sweden: Hanne Lobosco (MoH), Rosita Claesson Wigand (Public Health Agency of Sweden), Jenny Börlin (Public Health Agency of Sweden), England: Robert Aldridge (Institute of Health Informatics, University of London), Scotland: Ian Grant (National Statistics Department) and Wales: Ronan Lyons (University of Swansea).

X. Additional files

Additional file 1: It is a doc. word file. It describes the search strategies used to identify citations related to data linkage and machine learning technique used for health status monitoring.

Search strategy 1
((Health status monitoring [Title/Abstract] OR Surveillance [Title/Abstract]) AND Linked data [Title/Abstract]))

Search strategy 2
((Health status monitoring [Title/Abstract]) OR Surveillance [Title/Abstract]) AND Machine learning approach [Title/Abstract]))

Additional file 2: It is a doc. word file. It describes the definitions of different data sources used for health surveillance.

Here we describe the definitions of different types of data sources, artificial intelligence techniques applied, health outcome, determinants and intervention indicators.

Different types of data sources:

1. ***Health surveys/Population health surveys*** collect information of risk factors, health behaviors and non-health care determinants of health[1]. Health surveys are used to measure the prevalence of risk factors and healthy behavior, monitor the effects of interventions, measure community attitudes to health policy initiatives, as well as assess trends in health and disease outcomes. Health surveys could involve health interview surveys or health examination surveys. These health surveys will include those surveys which are performed either at national or sub-national levels.
2. ***Disease-specific or population-based registries:*** A registry is a collection of information about individuals, usually focused around a specific diagnosis or condition[2]. Many registries collect information about people who have a specific disease or condition, while others seek participants of varying health status who may be willing to participate in research about a particular disease. Individuals provide information about themselves to these registries on a voluntary basis. Registries could be disease-specific registries, screening registries, immunization registries, etc.
3. ***National cohorts*** are performed to investigate the causes of development of major chronic diseases, i.e. cardiovascular diseases, cancer, diabetes, neurodegenerative/-psychiatric diseases, musculoskeletal diseases, respiratory and infectious diseases, and their pre-clinical stages or functional health impairments at the national level.
4. ***Clinical trials data*** include data on safety and efficacy of interventions. This data may be available through national or international trial registries. For example, ClinicalTrials.gov, Cochrane Library, WHO International Clinical Trials Registry Platform (ICTRP), European Union Clinical Trials Database, etc.
5. ***Administrative data sources*** were initially developed for administrative use, not for public health surveillance and have a larger coverage of population. For example, birth certificates, death certificates, census, biobank data (i.e., it is a biorepository that accepts, processes, stores and distributes bio specimens [i.e., blood, urine, spinal fluid, etc.] and associated data for use in research and clinical care[3]), GIS (Geographical Information System/GPS/Geodata), socioeconomic data, and retirement/pension data, etc.
6. ***Electronic health/patients/medical records (EHRs)*** include a summary of administrative data, clinical data of patients and determinants of health indicators (i.e., various types of exposures). In scientific literature, electronic health records (EHRs) are often refer to patients' record or electronic medical records. These data sources are flexible to link with different types of data sources.
   "EHRs are described as a repository of patient data in digital form[4]and include the following information: active and past diagnosis; past medical history; physical examinations; laboratory test orders and results; current prescriptions; radiological images and reports; hospitalization information; consultant reports; details of emergency care; immunizations; pathology reports; social history; lifestyle; allergies; genetic information; health screening study results; physicians, nurse, social worker, physical therapy notes at admission and discharge[5]."

   EHRs may include the following data sources:

   1. *Hospital Record (HR)/In hospital medical record* includes information about a patient generated during a period of hospitalization with written accounts of consultants' opinions as well as nurses' observations and treatments[6].
   2. *Hospital Discharge Record is* a clinical report prepared by a physician or other health professional that summarizes the patient's chief complaint, the diagnostic findings, the therapy administered and the patients' response to it, and recommendations on discharge[7].
   3. *General Practitioner (GPs)/Primary care* include information on diagnoses and symptoms, laboratory test results, referrals to specialists and drug or healthcare product prescribed[8].
   4. *Specialist care* (i.e., cardiologists, neurologists, gynecologists, etc.) includes a highly skilled in a specific and restricted medical field[9].
   5. *Emergency care (i.e., emergency room and outpatient emergency)* include information on evaluation and initial treatment of medical conditions caused by trauma or sudden illness[10].

6. *Health insurance claim (i.e., healthcare reimbursement)* is a detailed invoice that a health care provider (such as doctor, clinic, or hospital) sends to the health insurer to reimburse the expenses spent on health services (i.e., drugs, diagnostic/laboratory tests, etc.) received by a patient[1].
7. *Drug prescription* include information on prescription date, type of drug, strength, dosage regimen, quantity, and route of administration[11].
8. *Genomic/DNA data* sources (i.e., information about functions of specific genes and to assess the association of gene mutations in certain diseases [2] such as for breast cancer BRCA 1/2).

7. *X-data sources)*: These type of data sources provide precise information on determinants of health and can include data on various exposures such as biological parameters, social behavior, life style, physical environment, nutrition, etc. These sources are considered as part of big data (i.e., voluminous amount of structured, semi structured and unstructured data that has the potential to be mined for information [12]).
We grouped these types of data sources as "X-data sources". Some of these data sources are enlisted below and more could be possible:

1. *m-Health (mobile-Health)* is the use of mobile phones, wireless health apps and wearable devices and measure a set of biological parameters which could be used for disease surveillance and health care services[13]. Data from these apps can also be used for disease surveillance, treatment support, epidemic outbreak tracking and chronic disease management[13].
2. *Social media* is an electronic communication through which users create online communities to share information, ideas, personal messages, and other content (such as videos) and the related data may reflect user's social behavior towards different aspects[14].
3. *Mobility mode data* (i.e., commuting for work by walking, using bicycle, public transport such as metro, train, bus, etc.) provide information about the mobility mode of general population whether walking or using different means of transport.
4. *Build physical environment* include data regarding green spaces, environmental exposure in terms of cleanliness, sound pollution, air quality, etc.
5. *Nutrition:* direct producers of seasonal/local fruits and vegetables, etc.
6. *Housing infrastructure* data source provide information about housing space and location.

Artificial intelligence (AI) techniques
The artificial intelligence techniques following techniques: machine learning, natural language processing, markov decision process, support vector machine, data mining, regression, etc., to analyze, estimate and predict the health indicators either from linked data or using an individual data set.

Artificial Intelligence					
Machine learning	Natural language processing	Markov decision process	Support vector machine	Data minin g	Others

Health outcome and intervention related indicators and determinants of health
We have selected the following non-communicable diseases based on burden of disease with higher incidence of mortality and morbidity across EU member states[15]: cardiovascular diseases, cancer, chronic obstructive pulmonary diseases, diabetes, neurodegenerative disease, mental health, accidents/trauma, maternal and perinatal health and any other disease.
Health outcomes indicators which are estimated from linked data and/or by using AI techniques to an individual data set, will be identified.

We have adopted EuroREACH Framework describing health status monitoring to classify the identified health outcome indicators, non-healthcare determinants of health and health intervention indicators under different categories[11] *(see figure 2)*.

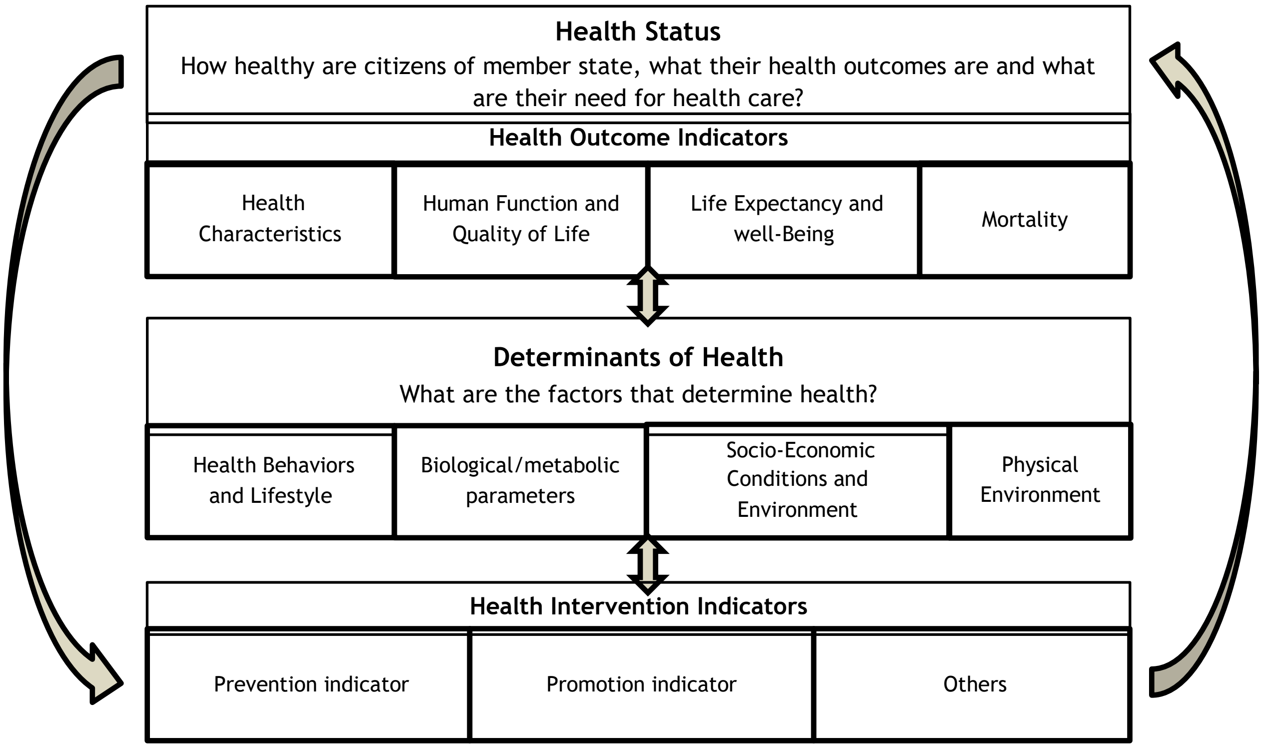


Figure 2: EuroREACH Framework for health status monitoring

Health outcome indicators which are estimated either from linked data or by applying AI techniques to an individual data set, describe the health status of a population in terms of health characteristics (e.g., prevalence of stroke among ≥ 65 years old), human function and quality of life (e.g., quality of life after stroke), life expectancy and well-being (e.g., survival of people with stroke) and mortality (e.g., causes of mortality).
These outcome indicators have the potential to improve health surveillance with more precise information. These health outcome indicators include prevalence, incidence, population attributable risk, population attributable fraction, relative risk, hazard ratio, etc.

Determinants of health which are identified either from linked data or from an individual data set, can provide more variables/parameters to better understand exposures factors related to health behavior and lifestyle (i.e., risk and/or protective behavior, response to health problems, etc.), biological/metabolic parameters (i.e., genetic, body structure and functioning, etc.), socio-economic conditions and environment (i.e., attitudes, social networks, education, employment, living standard, etc.) and physical environment (i.e., water quality, air quality, food safety, etc.). For example, use of public transport to commute for work. These determinants should be different from already compiled databases of OECD, WHO or Eurostat.

Health intervention indicators which are estimated either from linked data or by applying AI techniques to an individual data set, describe the effect of the interventions applied in terms of prevention (e.g., the use of genetic screening of BRCA1/2 genes among families with a family history of breast cancer) or promotion (i.e., integrated health programs at work place, schools, hospitals, policies and practices on healthy lifestyle, etc.).

References:

1. Madans JH. Health Surveys. In: Smelser NJ, Baltes PB, eds. *International Encyclopedia of the Social & Behavioral Sciences*. Oxford: Pergamon; 2001:6619-6627.
2. Health NIo. What is a registry? https://www.nih.gov/health-information/nih-clinical-research-trials-you/list-registries. 2018.
3. De Souza YG, Greenspan JS. Biobanking Past, Present and Future: Responsibilities and Benefits. *AIDS (London, England)*. 2013;27(3):303-312.
4. 20514 IT. Health Informatics-Electronic Health Record-Definition, Scope, and Context: https://www.iso.org/obp/ui/#iso:std:iso:tr:20514:ed-1:v1:en. 2005.
5. Häyrinen K, Saranto K, Nykänen P. Definition, structure, content, use and impacts of electronic health records: A review of the research literature. *International Journal of Medical Informatics*. 2008;77(5):291-304.
6. Katzenellenbogen JM BS, Somerford P, Anderson CS, Semmens JB, Codde JP, Vos T. Disability burden due to stroke in Western Australia: new insights from linked. *Int J Stroke*. 2010;5(4):269-277.
7. White C GM, Johnson B, Corbin T. Social inequalities in adult male mortality by the National Statistics. *Health Stat Q*. 2007;36:6-23.
8. Manuel DG SS. Health-related quality of life and health-adjusted life expectancy of people with. *Diabetes Care*. 2004;27(2):407-414.
9. Tuppin P, Rudant J, Constantinou P, et al. Value of a national administrative database to guide public decisions: From the systeme national d'information interregimes de l'Assurance Maladie (SNIIRAM) to the systeme national des donnees de sante (SNDS) in France. (0398-7620 (Print)).
10. Dunnell K BJ, Wood R, Babb P. Measuring aspects of women's life and work for the study of variations in health. *Am J Ind Med*. 1999;36(1):25-33.
11. EuroREACH. EuroREACH Framework: http://hdn.euhs-i.eu/performance/frameworks/euroreach-framework. 2013.
12. Big data: https://searchdatamanagement.techtarget.com/definition/big-data.
13. Lloyd K, McGregor J, John A, et al. A national population-based e-cohort of people with psychosis (PsyCymru) linking prospectively ascertained phenotypically rich and genetic data to routinely collected records: Overview, recruitment and linkage. *Schizophrenia Research*. 2015;166(1):131-136.
14. Atramont A, Bonnet-Zamponi D, Bourdel-Marchasson I, Tangre I, Fagot-Campagna A, Tuppin P. Health status and drug use 1 year before and 1 year after skilled nursing home admission during the first quarter of 2013 in France: a study based on the French National Health Insurance Information System. *European Journal of Clinical Pharmacology*. 2018;74(1):109-118.
15. WHO. Global status report on noncommunicable diseases: http://www.who.int/nmh/publications/ncd_report_full_en.pdf. 2010.

Additional file 3: It is a doc. word file including the survey questionnaire used to identify the data linkages practices across European countries.
To fill this questionnaire, it would require the assistance from the data department, statisticians, and epidemiologists to complete this questionnaire.

The questionnaire is divided into three parts:

PART I – GENERAL CHARACTERISTICS OF LINKED DATA
PART II – HEALTH OUTCOME AND INTERVENTION INDICATORS & DETERMINANTS OF HEALTH TO MONITOR NON- COMMUNICABLE DISEASES
PART III – EXPERIENCE WITH INNOVATIVE USE OF DATA IN PUBLIC HEALTH POLICY PROCESS

GENERAL CHARACTERISTICS OF RESPONDENTS:

Name of the country: ______________
Last name of the contact person: ______________________ First name: ___________________
Type of institute:
□ Public Health Institute (PHI) □ PHI affiliated with university □ PHI affiliated with research institute
□ Research Institute □ University □ National Statistics Department □ International Organization
□ Others, please specify: _________________
Name of current institute/agency: ____
Current position: __________________________
Work telephone number: ________________ E-mail: ____________________@_______________

***Please read the information detailed in *Additional file 2* before filling this questionnaire for a better understanding of the questions.**

PART I – GENERAL CHARACTERISTICS OF LINKED DATA

"Innovation in health information is described as:

- *Linkage of different data sources with each other using linkage technology and/or*
- *Applying artificial intelligence (AI) techniques either to linked data or to an individual data set, allowing a better understanding of what determines population health or the efficiency of the health system and decision making at different geographical levels or other categorization parameter level"*

1.1. Following the above definition, has your country been using linkage technology to link different data sources for health status monitoring in routine?
□ Yes
□ Development in progress
□ No
1.1.1. If yes, please indicate the level of data linkage *(and move to Q 1.3)*:
□ National level
□ Sub-national level
□ Metropolitan level

1.1.2. If development in progress, please give some details: _________________________
1.1.3. If no, is there any plan to develop that?
□ Yes
□ No
a. If yes, please mention the time horizon to implement that plan:
1. □ < 5 years □ 5 - 10 years □ > 10 years
b. If no, please describe some reasons why there is no project to develop approaches for innovative use of data sources: __________________________

1.2. Following the above definition, has your country been applying artificial intelligence techniques for health status monitoring in routine?
□ Yes
□ Development in progress
□ No
1.2.1. If yes, please indicate the level of application of AI techniques *(and move to Q 1.3)*:
□ National level
□ Sub-national level
□ Metropolitan level
1.2.1. If development in progress, please give some details: _________________________
1.2.2. If no, is there any plan to apply those techniques?
□ Yes
□ No
a. If yes, please mention the time horizon to adopt them:
□ < 5 years □ 5 - 10 years □ > 10 years
b. If no, please describe some reasons why there is no project to adopt those techniques:

- If the responses of above two questions (1.1 & 1.2) are "no", after describing the reasons, this questionnaire will end. A message will appear as *"Thank you very much for your participation to this survey."*

- If the responses of above two questions (1.1 & 1.2) are "yes or development/adoption in progress", *move to next question (i.e., Q 1.3).*

1.3. Please provide the following information about data sources (*i.e., either national or non-national data sources which are in use, available but not in use or development in progress)* , their linkage and general characteristics of linked data in your country as described in table 1:

Table 1: General characteristics of linked data sources/sets

Main objective of data linkage [e.g., to monitor health status, for risk factors, for for research (clinical and epidemiology, for health policy, others]	**Data sources used for linkages**							**Type of data linkage**	**Use of AI techniques applied**	**Estimated health indicators**	**General characteristics of linked datasets**					
	Health surveys	**Disease specific registries**	**National cohorts**	**Clinical trials data**	**Administrative data**	**EHRs**	**X-data sources**				**Current status in use of linked data**	**Level of use/implementation**	**Type of approval**	**Type of accessibility**	**Operate in real-time**	**Flexible to integrate new variables**

PART II: HEALTH OUTCOME AND INTERVENTION RELATED INDICATORS & DETERMINANTS OF HEALTH TO MONITOR NON-COMMUNICABLE DISEASE

In this section, please provide a set of those health outcomes and intervention related indicators which are estimated either from linked data and/or by applying AI techniques to individual level data. These health outcome and intervention related indicators could be estimated at national, Sub-national, metropolitan levels, or at all levels.
Please select *at least three or more medical conditions* and provide a set of related health outcome indicators, their category, the objective of estimation (i.e., for public health monitoring, research [clinical, epidemiology, public health], both), status of their use (i.e., was used, currently in use or could be produced in future) and level of estimation (i.e., national, Sub-national, metropolitan, at all levels) as described in the *table 2*:

1. *Cardiovascular diseases* (i.e., Stroke, myocardial infarction, heart failure, pulmonary embolism)
2. *Cancer* (i.e., lung cancer, breast cancer, colorectal cancer)
3. *Diabetes* (i.e., related complications)
4. *Chronic Obstructive Pulmonary Disease* (COPD i.e., emphysema, chronic bronchitis, refractory asthma)
5. *Neurodegenerative disease* (i.e., Alzheimer, Parkinson, dementia, multiple sclerosis)
6. *Mental health* (i.e., autism)
7. *Accidents/Trauma* (i.e., injuries)
8. *Maternal and perinatal health* (i.e., infections, severe bleeding, pre-eclampsia, gestational diabetes, pre-term birth, low birth weight)
9. *Any other medical condition (for example, arthritis, low back pain, joint pain, etc.)*

Table 2	Health outcome indicators					
	Medical condition	**Sub-category**	**Names of the indicators**	**Objective of estimation**	**Status of their use:**	**Level of estimation**
			1. Health characteristics			
	e.g., Cardiovascular disease	Stroke	e.g., prevalence of stroke among ≥ 65 years old	For public health monitoring	Currently in use	At all levels
			2. Human function and quality of life			
			e.g., quality of life after stroke	For public health monitoring	Currently in use	At all levels
			3. Life expectancy and well-being			
			e.g., survival of people with stroke	For public health monitoring	Currently in use	At all levels
			4. Mortality			
			e.g., causes of stroke mortality	For public health monitoring	Currently in use	At all levels

Please select *at least three or more medical conditions* and provide a set of related determinants of health, their category, the objective of estimation (i.e., for public health monitoring, research [clinical, epidemiology, public health], both), status of their use (i.e., was used, currently in use or could be produced in future) and level of estimation (i.e., national, Sub-national, metropolitan, at all levels) as described in the *table 3* below:

Table 3	**Determinants of health**						
	Medical condition	**Sub-category**	**Names of the determinants**	**Objective of estimation**	**Status of their use:**	**Data can be described by a set of following variables:**	**Level of estimation**
			1. Health behavior and life style (i.e., risk and/or protective behavior, response to health problems, etc.)				
			5. Biological/metabolic parameters (i.e., genetic, body structure and functioning, etc.)				
			6. Socioeconomic and environment (i.e., attitudes, social networks, education, employment, living standard, etc.)				
			7. Physical environment (i.e., water quality, air quality, food safety, etc.)				

Please select *at least three or more medical conditions* and provide a set of related health intervention indicators, their category, the objective of estimation (i.e., for public health monitoring, research [clinical, epidemiology, public health], both), status of their use (i.e., currently in use or could be produced in future) and level of estimation (i.e., national, Sub-national, metropolitan, at all levels) as described in the *table 4*:

Table 4	Health intervention indicators					
	Medical condition	**Sub-category**	**Names of the intervention indicators**	**Objective of estimation**	**Status of their use:**	**Level of estimation**
			1. Prevention (i.e., screening, vaccination, etc.)			
	e.g., cancer	Breast cancer	e.g., use of genetic screening of BRCA1/2 genes among families with a family history of breast cancer	For public health monitoring	Currently in use	At Sub-national levels
			2. Promotion (i.e., integrated health programs at work place, schools, hospitals, policies and practices on healthy lifestyle, etc.)			
			3. Others			

PART III: INNOVATIVE USE OF DATA SOURCES AND THEIR IMPLICATIONS IN PUBLIC HEALTH POLICY PROCESS

Based on your country experience (either previous experience or an ongoing study or an example which could be used in future) with the use of linked data and/or AI techniques, please answer the following questions:

1.1. Is there any <u>expected policy outcome</u> as a result of linked data in your country (*e.g., linking primary health care data and health insurance data, provides information on unequal access to health care, <u>expected policy outcome</u> is to provide better insight to reduce health inequalities*)?

□ Yes
□ In process of development
□ No

- *<u>If yes,</u>* please share *<u>an inspiring example of how the health outcome and health intervention related indicators</u>* (i.e., estimated from linked data or by using advanced statistical technique) are used
  - a) to *<u>support health policy decisions in real-time</u>* (if possible, please attach a report, case study or a methodological study describing that experience implying to a specific health condition and at certain geographical level [i.e., national, sub-national or metropolitan levels]).
  - b) *<u>to assess public health problems or to evaluate the impact of public health interventions or to plan public health programs</u>*
- *<u>If in process of development,</u>* please provide some details about ongoing experience.
- *<u>If no,</u>* please give some details on the reasons for not being integrated into the health policy process: ______________________________

(The inspiring examples will be used to develop the methodological guidelines to estimate/develop health indicators estimated either from linked data and/or using AI techniques and will be shared with other member states).

1.2. Additional comments (related to the use of linked data, AI techniques, health outcome or intervention related indicators, policy process, etc.): ____

1.3. Please attach a pdf copy or provide a link to the any existing guidelines or recommendations to develop the guidelines for health indicators in your country (if available).

<u>Additional file 4</u>: It is a doc. word file and describes the invitation email sent to the participants to complete the survey.

April 1, 2019

Dear colleagues,

We are contacting you on behalf of InfAct (Information for Action) Joint Action of MSs on Health Information Work Package (WP) 9.

WP 9 is about the “innovation of health information for public health policy development”. As a first step, we have developed a survey and kindly ask you to participate in this ***<u>national survey</u>***.

You are receiving this invitation because we have recognised you as an important member of the European public health information community. Your insight knowledge and experience would be a great support to develop a sustainable European research infrastructure on health information.

This survey explores the innovative use of data sources across the EU/EEA Member States in terms of data linkage, use of artificial intelligence techniques, related estimation of health indicators and their implications either in routine health status monitoring or in health policy process at any of the following levels: national, sub-national or metropolitan levels.

The information collected here will help us to develop a road map (i.e., current status, gaps, perspectives and opportunities) of innovative use of health information by comparing MSs.

Filling out this questionnaire, please access to the following survey link: https://casa.santepubliquefrance.fr/index.php/633298?lang=en

The deadline to complete this survey is <u>April 30, 2019.</u>

If you have any question or confusion or ambiguity about survey questions, please don’t hesitate to contact us on following emails:

- Romana.HANEEF@santepubliquefrance.fr
- Infact-FRANCE@santepubliquefrance.fr

Your kind support would be very much appreciated.

Best regards,

InfAct Work Package 9 Research Team
Santé Publique France (National Institute of Public Health)

Additional file 5: It is a doc. word file and include a table describing the names of survey respondents, their institutes and email addresses

S/No	Abbreviations	Country	First name	Last name	Institute	Email
1	AT	Austria	Stefan	Mathis-Edenhofer	Public Health Institute	stefan.mathis-edenhofer@goeg.at
2	BE	Belgium	Herman	Van Oyen	Sciensano	herman.vanoyen@sciensano.be
3	BG	Bulgaria	Raina	Nikolova	Public Health Institute	r.nikolova@ncpha.government.bg
4	HR	Croatia	Ivan	Pristas	CIPH (HZJZ)	ivan.pristas@hzjz.hr
5	CY	Cyprus	Vasilios	Scoutellas	Public Health Institute	VScoutellas@mphs.moh.gov.cy
6	CZ	Czech Republic	Sarka	Dankova	UZIS	sarka.dankova@uzis.cz
7	CZ	Czech Republic	Jiri	Jarkovsky	Institute of Health Information and Statistics	Jiri.Jarkovsky@uzis.cz
8	DK	Denmark	Mette Bjerrum	Koch	Danish Health Data Authority	mebk@sundhedsdata.dk
9	DK	Denmark	Maja Bæksgaard	Jørgensen	Statens Institut for Folkesundhed	mbha@sdu.dk
10	EE	Estonia	Eleri	Lapp	MoSA	eleri.lapp@sm.ee
11	FI	Finland	Mika	Gissler	THL	mika.gissler@thl.fi
12	DE	Germany	Angelika	Schaffrath Rosario	RKI	Schaffrath-RosarioA@rki.de
13	EL	Greece	Spyridon	Goulas	Governmental Health Insurance Organization	sgoulas@eopyy.gov.gr
14	FR	France	Anne	Gallay	Public Health Institute	Anne.GALLAY@santepubliquefrance.fr
15	FR	France	Jennifer	Zeitlin	INSERM U 1153 – Research Institute	jennifer.zeitlin@inserm.fr
16	IE	Ireland	Sheona	Gilsenan	DOH	sheona_gilsenan@health.gov.ie
17	IT	Italy	Luigi	Palmieri	ISS	luigi.palmieri@iss.it
18	IT	Italy	Brigid	Unim	ISS	brigid.unim@iss.it
19	LV	Latvia	Janis	Misins	CDPC	janis.misins@spkc.gov.lv
20	LT	Lithuania	Rita	Gaidelyte	HI	rita.gaidelyte@hi.lt
21	LU	Luxembourg	Anne-Charlotte	Lorcy	Health-Directorate	anne-charlotte.lorcy@ms.etat.lu
22	MT	Malta	Neville	Calleja	MFH	neville.calleja@gov.mt
23	NL	Netherlands	Peter	Achterberg	RIVM	peter.achterberg@rivm.nl
24	NO	Norway	Hakon	Haaheim	National Health Directorate	hakon.haaheim@helsedir.no
25	PL	Poland	Jakub	Adamski	MoH	j.adamski@mz.gov.pl
26	PL	Poland	Piotr	Nowosielski	MoH	p.nowosielski@mz.gov.pl
27	PT	Portugal	Paulo	Nogueira	DGS	paulo.nogueira@dgs.min-saude.pt
29	PT	Portugal	Martins	José	Public Health Institute	josemartins@dgs.min-saude.pt
30	PT	Portugal	Carlos	Dias	Public Health Institute	carlos.dias@insa.min-saude.pt
31	RO	Romania	Silviu	Radulescu	Public Health Institute	silviu.radulescu@insp.gov.ro

32	RS	Serbia	Maja	Krstic	Batut	maja_krstic@batut.org.rs
33	SK	Slovakia	Jan	Cap	National Health Information Centre	Jan.Cap@nczisk.sk
34	SI	Slovenia	Metka	Zaletel	NIJZ	metka.zaletel@nijz.si
35	ES	Spain	Beatriz	Perez-Gomez	ISCIII	bperez@isciii.es
36	SE	Sweden	Hanna	Lobosco	MoH	hanna.lobosco@folkhalsomyndigheten.se
	SE	Sweden	Rosita	Wigand	Public Health Agency of Sweden	rosita.wigand@folkhalsomyndigheten.se
	SE	Sweden	Jenny	Borlin	Public Health Agency of Sweden	jenny.borlin@folkhalsomyndigheten.se
37	UK-ENG	England	Robert	Aldrige	Welsh Government	r.aldridge@ucl.ac.uk
38	UK-SC	Scotland	Ian	Grant	National Statistics Department	ian.grant@nhs.net
39	UK-WL	Wales	Ronan	Lyons	University of Swansa	r.a.lyons@swansea.ac.uk