



A framework for user centred privacy and security in the cloud

“eHealth_hclinic_clarus” dataset

Author(s)	FCRB, KUL, URV
Document ID	CLARUS-DatasetEHealth-v1.0
Version	1.0
Date of Issue	27/05/2016
Document Distribution	Public
Abstract	<p>This dataset simulates a subset of a Passive Medical Record Database which contains the whole medical history of the patients of a big hospital. These patients have become “passive” due to the lack of encounters over a specific period of time or due to the death or change of residence of the patient, etc. Specifically, The “eHealth_hclinic_clarus” dataset represents coherent clinical data obtained from the discharge reports over one year (2014).</p> <p>The dataset has been synthetically generated from the real data of the “Hospital Clínic de Barcelona”. All the actual data are artificial and random, even though they preserve some statistical properties of the original data.</p> <p>Since all the data are synthetic, this dataset should not be used in any kind of medical research. Its only purpose is to provide an artificial but realistic medical dataset (with regard to data structure and distribution of variables) that can be used as input for the design and evaluation of privacy-preserving mechanisms.</p>

Table of Contents

1.	INTRODUCTION	3
2.	DATABASE SCHEMA	4
2.1.	RELATIONAL DIAGRAM.....	4
2.2.	DESCRIPTION OF THE TABLES	4
2.2.1.	<i>Patient</i>	4
2.2.2.	<i>Episode</i>	5
2.2.3.	<i>Discharge report</i>	6
2.2.4.	<i>Diagnose CIE9MC</i>	8
2.2.5.	<i>Document diagnose</i>	8
2.2.6.	<i>Lab results</i>	8
2.2.7.	<i>Medical service LOINC</i>	9
2.2.8.	<i>Document MS (Medical Service)</i>	9
3.	QUERIES.....	11
4.	HISTOGRAMS	12
4.1.	DISCHARGE REPORTS	12
4.2.	GENDER.....	12
4.3.	ZIP CODES	13
4.4.	AGE	13
4.5.	ADMISSION TYPE.....	15
4.6.	FACILITY DESTINATION	16
4.7.	DIAGNOSES	17
5.	SYNTHETIC REGENERATION OF DATA	21
6.	ATTRIBUTE CHARACTERIZATION AND PRIVACY REQUIREMENTS	22

1. Introduction

The “eHealth_hclinic_clarus” dataset represents coherent clinical data obtained from the discharge reports over one year (2014), in which ten of the more relevant medical services in HCB (Cardiology, Endocrinology, Gastroenterology, Gynecology, Hematology, Hepatology, Nephrology, Neurology, Ophthalmology and Urology) participated.

The data have been synthetically generated from the real data of the “Hospital Clínic de Barcelona”. All the actual data are artificial and random, even though they preserve some statistical properties of the original data.

Since all the data are synthetic, this dataset should not be used in any kind of medical research. Its only purpose is to provide an artificial but realistic medical dataset (with regard to data structure and distribution of variables) that can be used as input for the design and evaluation of privacy-preserving mechanisms.

The provided SQL script creates the tables of a simulated Medical Record Database and inserts the data. A second SQL script with the definition of four useful views is provided.

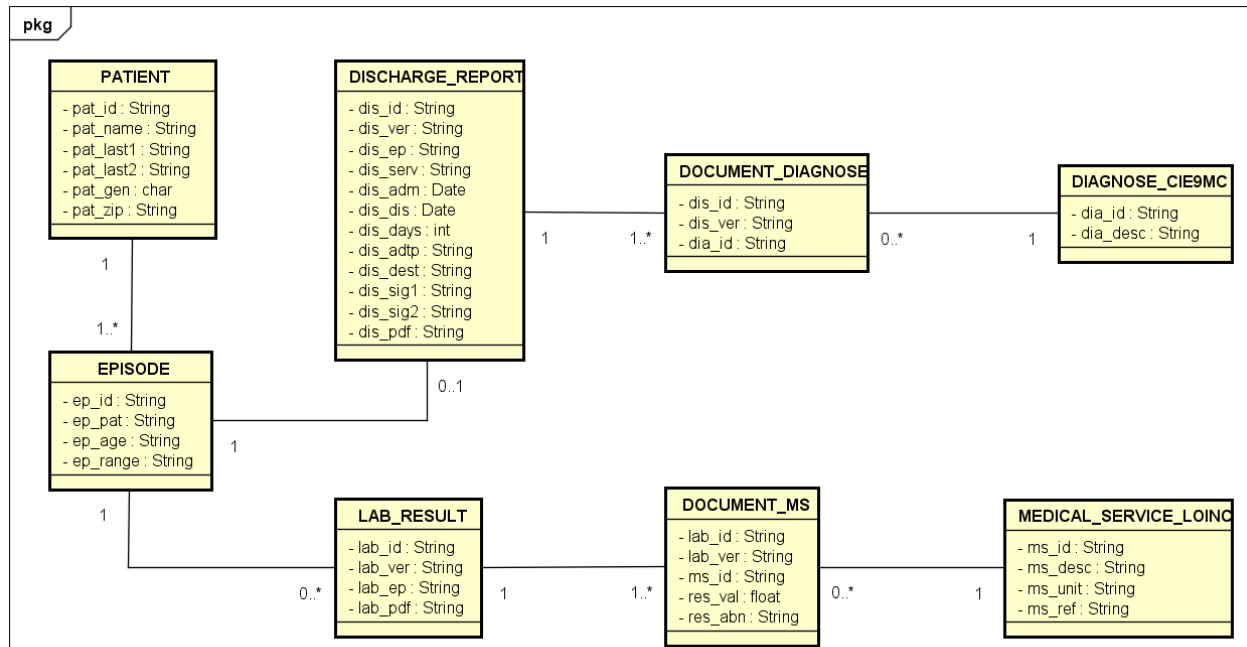
The content of the document is structured as follows: section 2 presents the structure of the database and details the content of every table. Section 3 sketches some examples of queries that can be carried out with this dataset. Section 4 shows the distribution of the different data attributes. Section 5 explains how the synthetic data have been generated. In Section 6 the different data attributes are classified according to their sensitiveness. Finally, we attach an annex with the description of the ICD-9-CM/CIE9MC¹ and LOINC² codes used.

¹ <http://www.cdc.gov/nchs/icd/icd9cm.htm>.

² <http://loinc.org/>

2. Database schema

2.1. Relational Diagram



powered by Astah

2.2. Description of the tables

2.2.1. Patient

Patient identification number

- Field name: pat_id
- Description: Unique eight-digit identifier assigned to each patient.
- Variable type: varchar (8)

Patient name

- Field name: pat_name
- Description: Name of the patient
- Variable type: text

Patient last name

- Field name: pat_last1, pat_last2
- Description: Last name of the patient split in two parts, just like the Spanish naming convention. Usually, the first part corresponds to the first part of the last name of the patient's father and the second part corresponds to the first part of the last name of the patient's mother.
- Variable type: text

Patient gender

- Field name: pat_gen

- Description: Patient gender codification character.
 - o M = Male
 - o F = Female
 - o U= Unknown
- Variable type: varchar (1)

Patient ZIP code

- Field name: pat_zip
- Description: Zip code of the patient's residence
- Variable type: varchar (5)

2.2.2. Episode

Episode identification number

- Field name: ep_id
- Description: Unique ten-digit identifier assigned to each episode.
- Variable type: varchar (10)

Patient identifier

- Field name: ep_pat
- Description: Patient assigned to a specific episode. A patient can be assigned to more than one episode, but an episode can only be assigned to one patient.
- Variable type: varchar (8)

Episode age of admission

- Field name: ep_age
- Description: Age of the patient on a specific episode.
- Variable type: int

Episode age range of admission

- Field name: ep_range
- Description: Age range of the patient, on admission time, from a specific episode.
 - o 01 = Less than 1 year
 - o 02 = 1-4 years
 - o 03 = 5-9 years
 - o 04 = 1-14 years
 - o 05 = 15-19 years
 - o 06 = 20-24 years
 - o 07 = 25-29 years
 - o 08 = 30-34 years
 - o 09 = 35-39 years
 - o 10 = 40-44 years
 - o 11 = 45-49 years
 - o 12 = 50-54 years
 - o 13 = 55-59 years
 - o 14 = 60-64 years

- 15 = 65-69 years
- 16 = 70-74 years
- 17 = 75-79 years
- 18 = 80-84 years
- 19 = 85-years or older
- 00 = Unknown age
- Variable type: varchar (2)

2.2.3. Discharge report

Discharge report identification number

- Field name: dis_id
- Description: Twenty-digit identification number assigned to each discharge report.
- Variable type: varchar (20)

Discharge report version number

- Field name: dis_ver
- Description: Two-digit number assigned to each discharge report to identify modifications over the same discharge report.
- Variable type: varchar (2)

Discharge report episode

- Field name: dis_ep
- Description: Episode assigned to a specific discharge report. An episode can only have one assigned episode at most.
- Variable type: varchar (10)

Discharge report medical service source

- Field name: dis_serv
- Description: Medical service source who signs the report.
 - CAR = CARDIOLOGIA
 - END = ENDOCRINOLOGIA
 - GAS= GASTROENTEROLOGIA
 - GIN = GINECOLOGIA
 - HEM = HEMATOLOGIA
 - HEP = HEPATOLOGIA
 - NEF = NEFROLOGIA
 - NRL = NEUROLOGIA
 - OFT = OFTALMOLOGIA
 - URO = UROLOGIA
- Variable type: varchar (3)

Discharge report admission date

- Field name: dis_adm
- Description: Admission date of the patient in YYYY-MM-DD format.
- Variable type: date

Discharge report discharge date

- Field name: dis_dis
- Description: Discharge date of the patient in YYYY-MM-DD format.
- Variable type: date

Discharge report days elapsed

- Field name: dis_days
- Description: Days elapsed from the admission date to the discharge date.
- Variable type: int

Discharge report admission type

- Field name: dis_adtp
- Description: Admission type from the discharge report.
 - o 00 = Unknown
 - o 01 = Diag/Tractament
 - o DP = Decision propia
 - o EN = Enfermedad
 - o OJ = O. Judicial
 - o OT = Otros
 - o PR = Programado
 - o UR = Urgente
- Variable type: varchar (2)

Discharge report facility service destination

- Field name: dis_dest
- Description: Destination of the patient when he/she is discharged.
- 00 = Unknown
 - o 01 = Dom. Familiar
 - o 02 = Dom + PADES
 - o 03 = Dom. + Rehabilit.
 - o 04 = Residencia
 - o 05 = CSS Convales.
 - o 06 = CSS Paliat.
 - o 07 = CSS L. Estada
 - o 08 = Hospital Agudos
 - o 09 = Defunc.
 - o 10 = Trasl. Int. CSC
 - o 11 = Hosp. Domiciliaria
 - o 12 = Custodia Policial
 - o 13 = Inst. Medicina Legal
 - o 14 = Ingr. Hospital
 - o 15 = Atenció Primària
- Variable type: varchar (2)

Discharge report doctor signature

- Field name: dis_sig1, dis_sig2
- Description: Six-digit identification number of the doctors who signed the discharge report.
- Variable type: varchar (6)

Discharge report PDF document

- Field name: dis_pdf
- Description: Discharge report PDF document in Base64 format.
- Variable type: text

2.2.4. Diagnose CIE9MC

Diagnose identification number

- Field name: dia_id
- Description: Unique six-digit identifier assigned to each diagnose which corresponds to the CIE9MC coding diagnoses system (See Annex)
- Variable type: varchar (6)

Diagnose description

- Field name: dia_desc
- Description: Description of the diagnose code
- Variable type: text

2.2.5. Document diagnose

Discharge report identification number

- Field name: dis_id
- Description: Twenty-digit identification number assigned to each discharge report.
- Variable type: varchar (20)

Discharge report version number

- Field name: dis_ver
- Description: Two-digit number assigned to each discharge report to identify modifications over the same discharge report.
- Variable type: varchar (2)

Diagnose identification number

- Field name: dia_id
- Description: Six-digit identifier assigned to each diagnose which corresponds to the CIE9MC coding diagnoses system.
- Variable type: varchar (6)

2.2.6. Lab results

Laboratory results identifier

- Field name: lab_id
- Description: Twenty-digit identification number assigned to each lab result.
- Variable type: varchar (20)

Laboratory result version

- Field name: lab_ver
- Description: Two-digit number assigned to each lab result to identify modifications over the same lab result.
- Variable type: varchar (2)

Laboratory result episode identifier

- Field name: lab_ep
- Description: Episode assigned to a specific lab result. An episode can be assigned to more than one lab result, but a lab result can only be assigned to one episode.
- Variable type: varchar (10)

Laboratory result document

- Field name: lab_pdf
- Description: Lab result PDF document in Base64 format.
- Variable type: text

2.2.7. Medical service LOINCMedical service identifier

- Field name: ms_id
- Description: Unique seven-digit identifier assigned to each medical service which corresponds to the LOINC coding medical service system (See Annex)
- Variable type: varchar (7)

Medical service description

- Field name: ms_desc
- Description: Description of the medical service
- Variable type: text

Medical service units

- Field name: ms_unit
- Description: Measure units of the medical service.
- Variable type: text

Medical service reference value

- Field name: ms_ref
- Description: Higher and lower value interval considered normal for the medical service.
- Variable type: text

2.2.8. Document MS (Medical Service)Laboratory results identifier

- Field name: lab_id
- Description: Twenty-digit identification number assigned to each lab result.
- Variable type: varchar (20)

Laboratory result version

- Field name: lab_ver
- Description: Two-digit number assigned to each lab result to identify modifications over the same lab result.
- Variable type: varchar (2)

Medical service identifier

- Field name: ms_id
- Description: Seven-digit identifier assigned to each medical service which corresponds to the LOINC coding medical service system.
- Variable type: varchar (7)

Result value

- Field name: res_val
- Description: Numeric value of the result for a specific medical service.
- Variable type: decimal

Abnormal result value

- Field name: res_abn
- Description: Two-character code informing of an abnormal result value.
 - _A = High
 - AA = Very high
 - _B = Low
 - BB = Very low
- Variable type: varchar (2)

3. Queries

Four SQL views have been created in order to make the data retrieval easier. Two of them (*discharge_simple* and *lab_simple*) can be used for simple queries whereas the other two (*discharge_advanced* and *lab_advanced*) can be used for performing more advanced queries and statistic computation.

We present some examples:

- Get discharge reports from a given patient id:
 - o `SELECT * FROM discharge_simple WHERE pat_id='pat_id';`
- Get discharge reports from a given patient name and last name:
 - o `SELECT * FROM discharge_simple WHERE pat_name='name' AND pat_last1='last1' AND pat_last2='last2';`
- Get discharge reports by gender:
 - o `SELECT DISTINCT ON (dis_id, dis_ver) * FROM discharge_advanced WHERE pat_gen='M';`
- Get discharge reports by age range:
 - o `SELECT DISTINCT ON (dis_id, dis_ver) * FROM discharge_advanced WHERE ep_range='01';`
- Get discharge reports by service:
 - o `SELECT DISTINCT ON (dis_id, dis_ver) * FROM discharge_advanced WHERE dis_serv='CAR';`
- Get discharge reports by admission date:
 - o `SELECT DISTINCT ON (dis_id, dis_ver) * FROM discharge_advanced WHERE dis_adm='YYYY-MM-DD';`
- Get discharge reports by discharge date:
 - o `SELECT DISTINCT ON (dis_id, dis_ver) * FROM discharge_advanced WHERE dis_dis='YYYY-MM-DD';`
- Get discharge reports by admission type:
 - o `SELECT DISTINCT ON (dis_id, dis_ver) * FROM discharge_advanced WHERE dis_adtp='EN';`
- Get discharge reports by facility service destination:
 - o `SELECT DISTINCT ON (dis_id, dis_ver) * FROM discharge_advanced WHERE dis_dest='01';`
- Get discharge reports by diagnose id:
 - o `SELECT DISTINCT ON (dis_id, dis_ver) * FROM discharge_advanced WHERE dia_id='070.54';`

4. Histograms

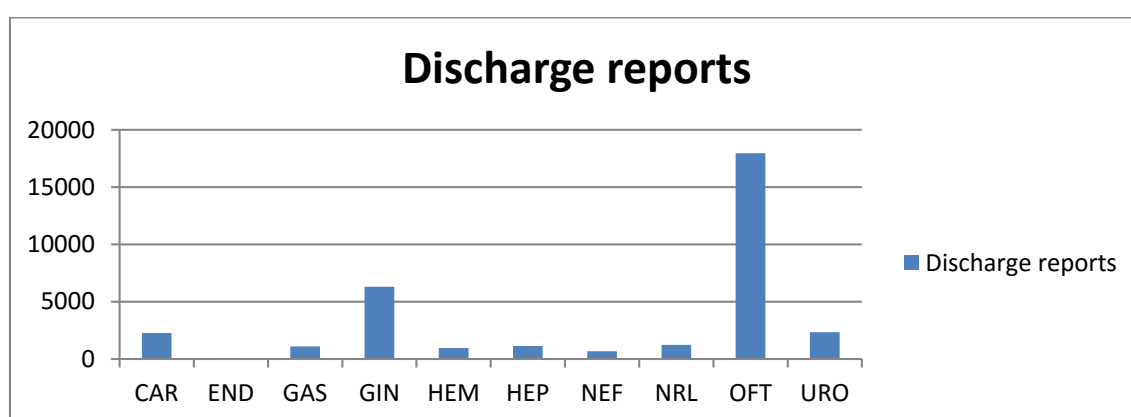
This dataset represents all the discharge reports released during 2014 for the following services: Cardiology (CAR), Endocrinology (END), Gastroenterology (GAS), Gynecology (GIN), Hematology (HEM), Hepatology (HEP), Nephrology (NEF), Neurology (NRL), Ophthalmology (OFT) and Urology (URO).

4.1. Discharge reports

The following table represents the total number of reports for each service and the total amount of reports.

	CAR	END	GAS	GIN	HEM	HEP	NEF	NRL	OFT	URO	TOTAL
Discharge reports	2267	90	1095	6301	951	1134	671	1228	17954	2334	34025

Table 1

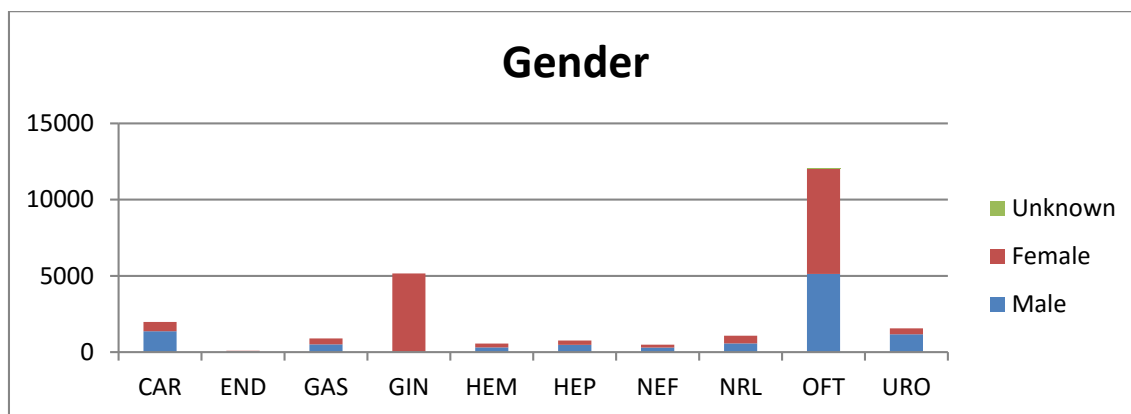


4.2. Gender

The table below represents the total number of male, female and unknown gender patients for each service.

	CAR	END	GAS	GIN	HEM	HEP	NEF	NRL	OFT	URO	TOTAL
Male	1363	40	505	20	303	479	292	569	5132	1171	9874
Female	615	45	390	5139	255	279	192	503	6934	383	14735
Unknown									1		1

Table 2

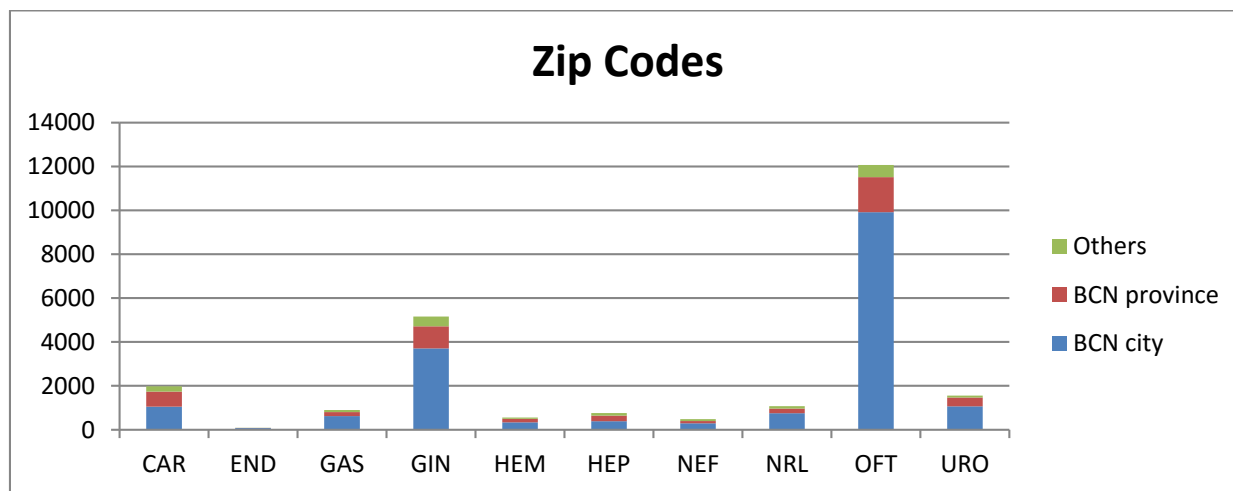


4.3. ZIP Codes

The distribution of ZIP codes by patient for each service is represented in the following table. ZIP codes have been grouped by zones to avoid much dispersion. For BCN city we have ZIP codes between 08000 and 08080; for BCN province, we have ZIP codes between 08080 and 08980, and for Others, ZIP codes from other parts of Catalonia, Spain or other countries are used.

	CAR	END	GAS	GIN	HEM	HEP	NEF	NRL	OFT	URO	TOTAL
BCN city	1051	70	619	3706	337	385	289	753	9914	1063	18187
BCN province	687	11	181	1008	158	257	106	213	1597	401	4619
Others	240	4	95	445	63	116	89	106	556	90	1804

Table 3



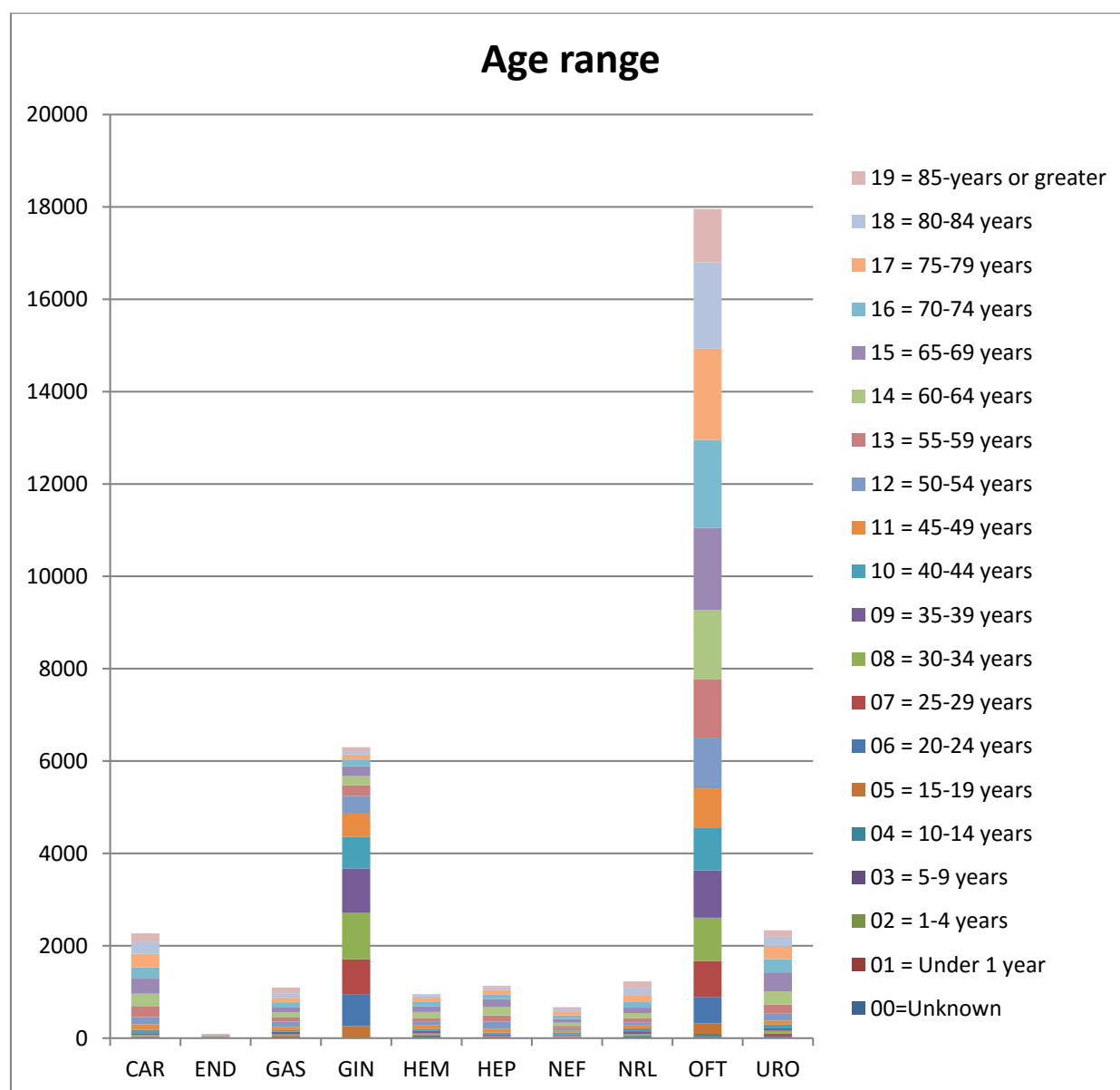
4.4. Age

The age range distribution is represented in the following table. It is obtained from the different episodes of each patient into the different services.

	CAR	END	GAS	GIN	HEM	HEP	NEF	NRL	OFT	URO	TOTAL
00=Unknown			1	5	1		1		2	1	11
01 = Under 1 year									2		2

02 = 1-4 years									15		15
03 = 5-9 years								1	29		30
04 = 10-14 years	3			4				1	45		53
05 = 15-19 years	5	1	2	253	11	3	12	7	228	17	539
06 = 20-24 years	17	9	15	692	21	10	7	29	572	35	1407
07 = 25-29 years	18	6	30	754	31	15	8	14	779	49	1704
08 = 30-34 years	23	10	33	1002	38	14	29	34	935	55	2173
09 = 35-39 years	43	5	57	966	53	44	21	64	1023	54	2330
10 = 40-44 years	74	6	37	683	50	39	39	61	917	83	1989
11 = 45-49 years	115	10	72	503	83	80	49	62	863	87	1924
12 = 50-54 years	157	11	111	382	76	152	49	69	1098	150	2255
13 = 55-59 years	241	11	90	232	74	133	49	97	1257	197	2381
14 = 60-64 years	267	7	111	198	129	181	70	106	1506	288	2863
15 = 65-69 years	325	6	116	212	121	174	80	115	1781	409	3339
16 = 70-74 years	246	2	101	147	93	103	76	120	1904	287	3079
17 = 75-79 years	293	3	97	106	97	103	89	154	1974	269	3185
18 = 80-84 years	266	2	97	75	57	59	62	157	1868	216	2859
19 = 85-years or greater	174	1	125	87	16	24	30	137	1156	137	1887

Table 4

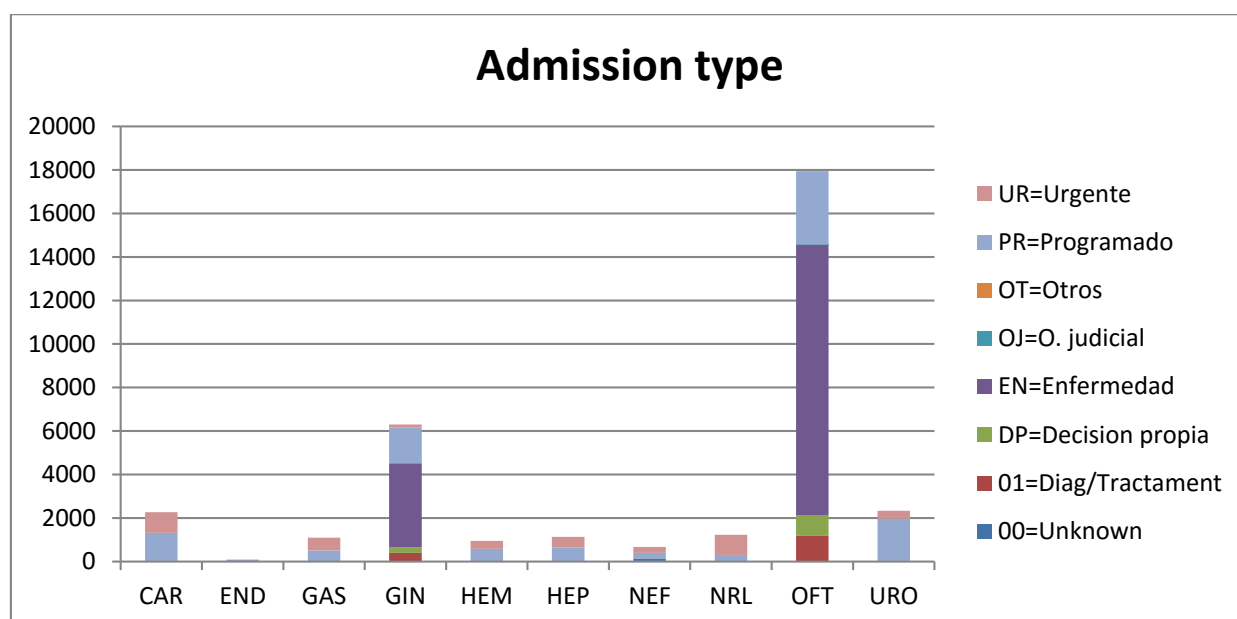


4.5. Admission type

The admission type distribution is represented in the next table. It is obtained from the different discharge reports from each service.

	CAR	END	GAS	GIN	HEM	HEP	NEF	NRL	OFT	URO	TOTAL
00=Unknown	33	53	3	26	25	5	141	6	21	36	349
01=Diag/Tractament				376					1178		1554
DP=Decision propia				248					912		1160
EN=Enfermedad			1	3870				3	12444	1	16319
OJ=O. judicial									12		12
OT=Otros	1			1					1		3
PR=Programado	1284	11	506	1644	560	648	255	296	3377	1932	10513
UR=Urgente	949	26	585	136	366	481	275	923	9	365	4115

Table 5

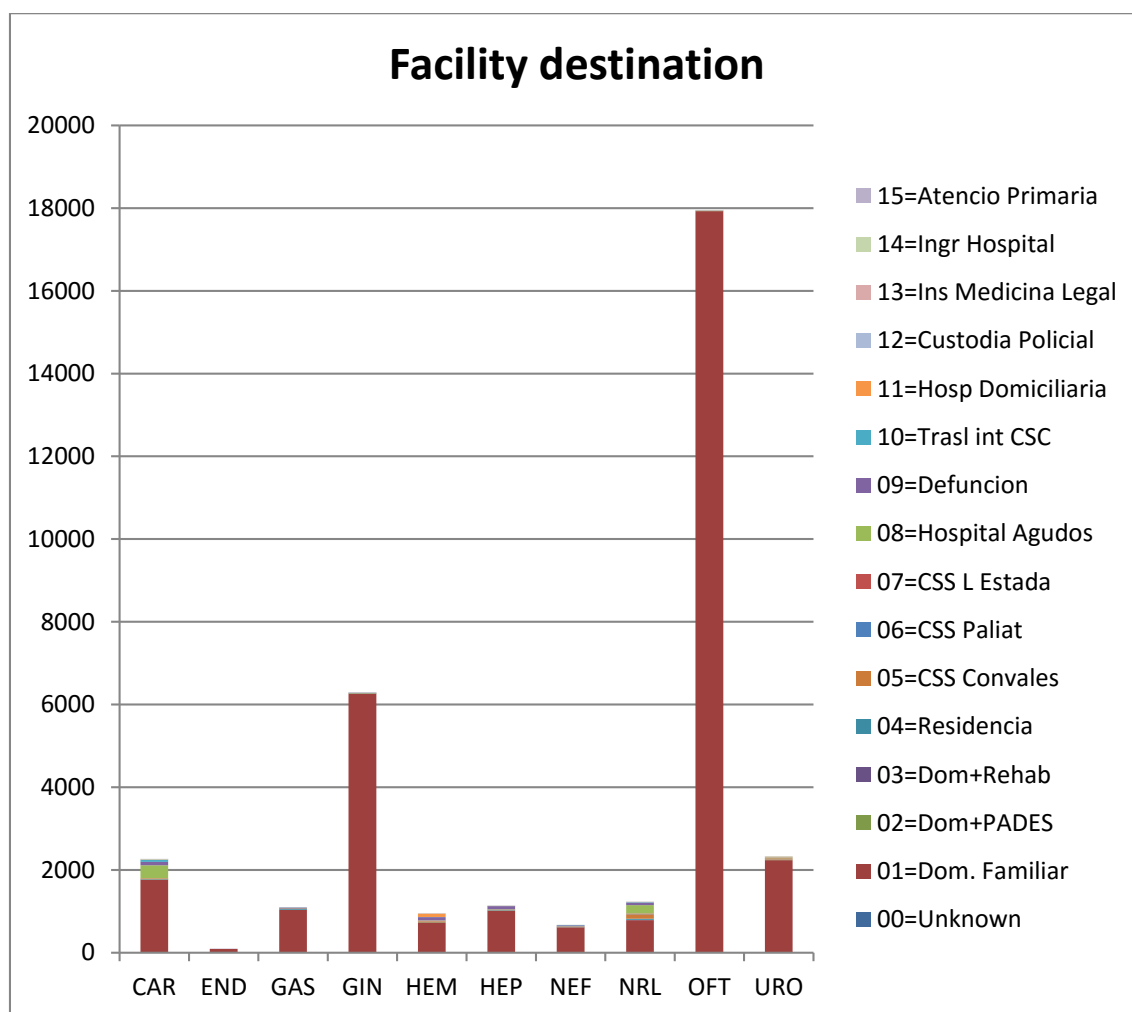


4.6. Facility destination

The facility destination is distributed as follows. It's obtained in the same way as the admission type.

	CAR	END	GAS	GIN	HEM	HEP	NEF	NRL	OFT	URO	TOTAL
00=Unknown	7		1	18	3	1	3	2	2	1	38
01=Dom. Familiar	1756	89	1033	6243	729	1011	609	786	17924	2236	32416
02=Dom+PADES	1		2	3	13	3			10	3	35
03=Dom+Rehab	2	1	1		2	1	1	6		3	17
04=Residencia	12		19	1	6	10	4	25		9	86
05=CSS Convaless	19		11	3	13	11	12	111	2	26	208
06=CSS Paliat			3	2	10	9	2	6			32
07=CSS L Estada	3		1		6	1	1	12		1	25
08=Hospital Agudos	314		1	7	4	7	7	200	4	8	552
09=Defuncion	82		18	2	75	72	24	63	2	3	341
10=Trasl int CSC	53		3	7	4	1	6	11		13	98
11=Hosp Domiciliaria	8				79	1		1		15	104
12=Custodia Policial	1			1		1		1	2		6
13=Ins Medicina Legal				3						1	4
14=Ingr Hospital	8		2	3	1	3	2	2	6	11	38
15=Atencio Primaria	1			8	6	2		2	2	4	25

Table 6



4.7. Diagnoses

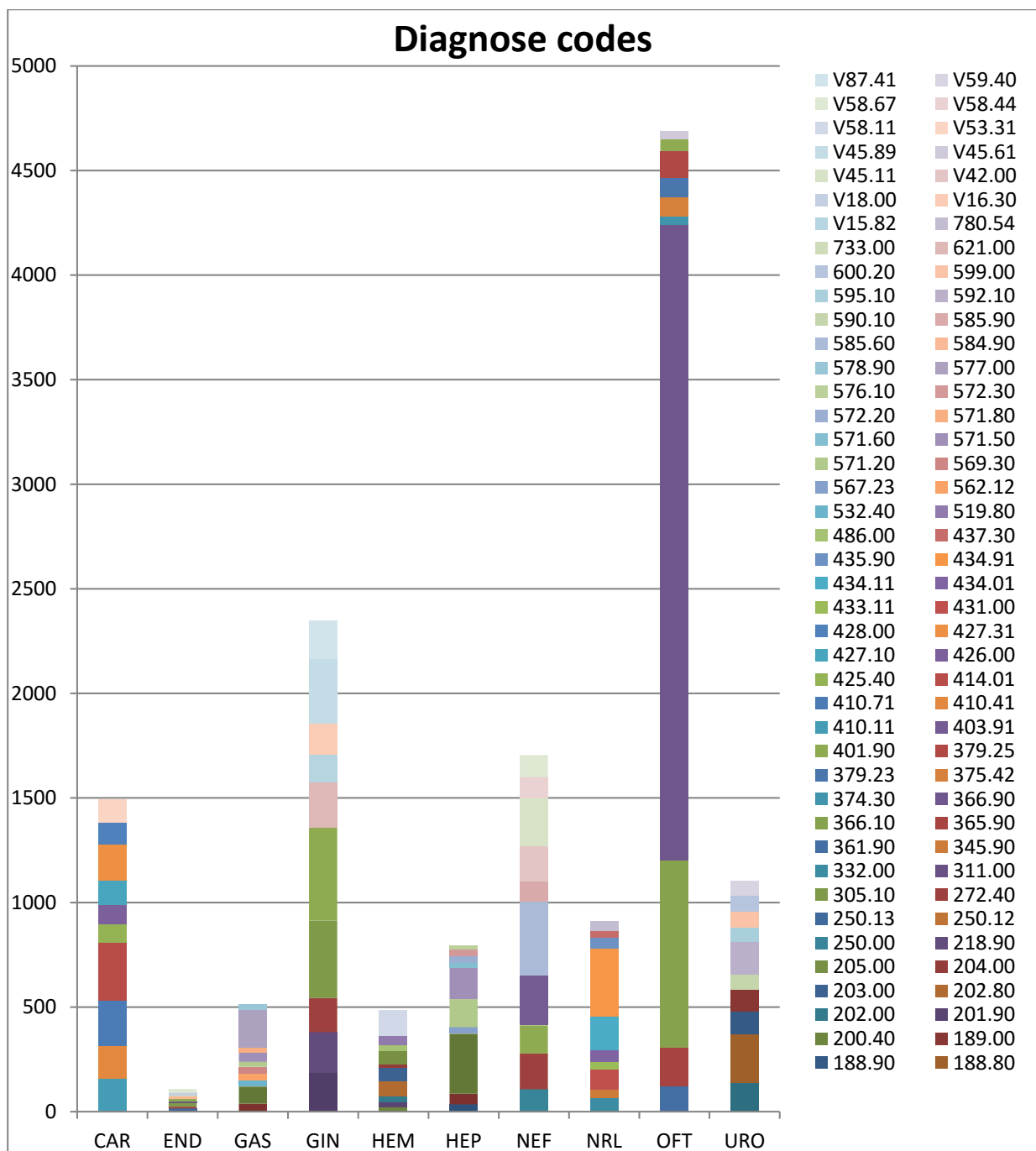
The diagnose codes for each diagnose report from each service are distributed as follows.

	CAR	END	GAS	GIN	HEM	HEP	NEF	NRL	OFT	URO	TOTAL
070.44						35					35
070.54			38			51					89
155.00			82			286					368
174.8				185							185
185.00										136	136
188.80										236	236
188.90										107	107
189.00										104	104
200.40					21						21
201.90					24						24
202.00					27						27
202.80					76						76
203.00					61						61
204.00					19						19

205.00					63						63
218.90				195							195
250.00							107				107
250.12		7									7
250.13		8									8
272.40		9		163			171				343
305.10		18		370							388
311.00		6									6
332.00								66			66
345.90								39			39
361.90									123		123
365.90									182		182
366.10									898		898
366.90									3035		3035
374.30									44		44
375.42									90		90
379.23									93		93
379.25									128		128
401.90		15		444			135		57		651
403.91							240				240
410.11	158										158
410.41	156										156
410.71	220										220
414.01	273										273
425.40	89										89
426.00	92										92
427.10	118										118
427.31	172										172
428.00	105										105
431.00								96			96
433.11								38			38
434.01								54			54
434.11								160			160
434.91								328			328
435.90								51			51
437.30								32			32
486.00					28						28
519.80					45						45
532.40			31								31
562.12			32								32
567.23						32					32
569.30			31								31
571.20			24			134					158
571.50			44			149					193

571.60						25					25
571.80			26								26
572.20						31					31
572.30						33					33
576.10						19					19
577.00			179								179
578.90			26								26
584.90		9									9
585.60							352				352
585.90							95				95
590.10										74	74
592.10										153	153
595.10										72	72
599.00										77	77
600.20										74	74
621.00				219							219
733.00		6									6
780.54								45			45
V15.82				133							133
V16.30				146							146
V18.00		12									12
V42.00							171				171
V45.11							227				227
V45.61									38		38
V45.89				311							311
V53.31	110										110
V58.11					124						124
V58.44							101				101
V58.67		19					106				125
V59.40										71	71
V87.41				181							181

Table 7



5. Synthetic regeneration of data

Because the original healthcare data cannot be released, the eHealth dataset has been synthetically created. To obtain a plausible synthetic dataset, values have been generated by following the distribution of original values which have been given by the data provider Hospital Clinic of Barcelona (HCB). The distributions we have considered to regenerate the attribute values are:

- Number of reports for each service (Table 1)
- Gender patients for each service (Table 2)
- ZIP codes by patient for each service (Table 3)
- Age range for each service (Table 4)
- Admission type for each service (Table 5)
- Facility destination for each service (Table 6)
- Diagnose code for each service (Table 7)

The distributions of these attributes are available in the previous section “Histograms”. According to these distributions of original attribute values, first, we generated the table “PATIENT” following the distributions given in tables 1, 2 and 3 for each service. Then, we generated the tables “EPISODES” and “DISCHARGE_REPORT” following the distributions given in tables 4, 5, 6 and 7. An episode corresponds to a discharge report. Taking the services into account, we randomly assigned a patient to each episode. A patient can be assigned to more than one episode because there are more episodes than patients. The values in tables “DOCUMENT_DIAGNOSE” and “DIAGNOSE_CIE9MC” are randomly generated according to the domains defined in the annex, because no information about data distributions is available. Likewise, table “MEDICAL_SERVICE_LOINC” is randomly generated with domains defined in the annex. Finally, tables “LAB_RESULT” and “DOCUMENT_MS” are completely random. Through this process, the synthetic eHealth dataset exactly preserves the marginal attribute distributions of data and their relative distribution toward the medical service source, which is important to maintain the coherence of the data at a record level.

6. Attribute Characterization and privacy requirements

Database attributes have been characterized according to their sensitiveness so we can define our privacy requirements and the appropriate protection measures to undertake.

From a data protection perspective, attributes can be classified as:

- *Identifiers*: attributes that, individually, identify the individual univocally (e.g., Social Security numbers, names, etc.). To avoid *identity disclosure*, the values of these attributes must be removed or encrypted in the released dataset.
- *Quasi-identifiers*: those combinations of attributes (e.g., place of birth + age + job) that, even though they do not cause disclosure individually, they may re-identify an individual when considered in aggregate. The values of these attributes should be protected (masked) prior releasing the dataset so that i) re-identification inferences are no longer possible and ii) the masked values still preserve their analytical utility.
- *Confidential* attributes are those attributes that reveal sensitive information about an individual (e.g., diagnoses), and thus, may cause *attribute disclosure*. If the former attributes have been properly protected with respect to these, confidential attributes can be left in clear (because the confidential data they provide cannot be linked to any specific individual); this is crucial, since most research tasks focus on this kind of data.
- *No confidential* attributes are the remaining non-sensitive attributes that do not require any specific treatment.

We have characterized the attributes in the eHealth dataset according to the notion of re-identificative and confidential data stated in current legislations on healthcare data protection. On the one hand, Article 8 par.1 of the EU Directive 95/46/EC prohibits the processing of personal data concerning health or sex life, unless the conditions under paragraph 2 of Article 8 are met³. In particular, the EU Directive 95/46/EC requires a higher level of protection for health data. On the other hand, in the USA there is the Health Insurance Portability and Accountability Act (HIPAA⁴), which defines the attributes in electronic healthcare records which should be protected because they may re-identify individuals. These rules cover strict identifiers (such as names and social security numbers) but also quasi-identifiers, such as combinations of demographic data. According the HIPAA, the confidential data refers to the individual's past, present or future physical or mental health or condition, the provision of health care to the individual, or the past, present, or future payment for the provision of health care to the individual. In Europe, these matters are currently regulated on a national level, in the sense that national legislations which transposed the EU Directive 95/46 apply, until the General Data Protection Regulation (GDPR)⁵ comes into force on the 25th May 2018⁶.

³ Article 8 of Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data of 24 October 1995, available at <http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:31995L0046>

⁴ HIPAA Privacy Rule, available at <https://www.gpo.gov/fdsys/pkg/PLAW-104publ191/html/PLAW-104publ191.htm>.

⁵ Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation).

⁶ See also Addendum of D2.4 Legal and Ethical Requirements.

According to these guidelines, the following table states the attributes in the eHealth dataset that are either identifiers, quasi-identifiers or confidential.

Attribute	Type
pat_id	Identifier
pat_name	Identifier
pat_last1	Identifier
pat_last2	Identifier
pat_gen	Quasi identifier
pat_zip	Quasi identifier
ep_age	Quasi identifier
ep_range	Quasi Identifier
dis_adm	Quasi Identifier
dis_dis	Quasi Identifier
dis_days	Quasi Identifier
dis_id	Confidential
dis_ver	Confidential
dia_id	Confidential

The remaining attributes are considered non-confidential.

Annex

ICD-9-CM/CIE9MC diagnoses

Code	Description
070.44	Chronic hepatitis C with hepatic coma.
070.54	Chronic hepatitis C without mention of hepatic coma.
155.00	Malignant neoplasm of liver and intrahepatic bile ducts. Liver, primary.
174.8	Malignant neoplasm of female breast. Other specified sites of female breast.
185.00	Malignant neoplasm of prostate.
188.80	Malignant neoplasm of bladder. Other specified sites of bladder.
188.90	Malignant neoplasm of bladder. Bladder, part unspecified.
189.00	Malignant neoplasm of kidney and other and unspecified urinary organs. Kidney, except pelvis.
200.40	Mantle cell lymphoma.
201.90	Hodgkin's disease, unspecified.
202.00	Nodular lymphoma.
202.80	Other lymphomas.
203.00	Multiple myeloma.
204.00	Lymphoid leukemia. Acute.
205.00	Myeloid leukemia. Acute.
218.90	Leiomyoma of uterus, unspecified.
250.00	Diabetes mellitus without mention of complication.
250.12	Diabetes with ketoacidosis. Type II or unspecified type, uncontrolled.
250.13	Diabetes with ketoacidosis. Type I [juvenile type], uncontrolled.
272.40	Other and unspecified hyperlipidemia.
305.10	Tobacco use disorder.
311.00	Depressive disorder, not elsewhere classified.
332.00	Parkinson's disease. Paralysis agitans.
345.90	Epilepsy, unspecified.
361.90	Unspecified retinal detachment.
365.90	Unspecified glaucoma.
366.10	Senile cataract.
366.90	Unspecified cataract.
374.30	Ptosis of eyelid.
375.42	Chronic dacryocystitis.
379.23	Vitreous hemorrhage.
379.25	Vitreous membranes and strands.
401.90	Essential hypertension. Unspecified.
403.91	Hypertensive chronic kidney disease. Unspecified. With chronic kidney disease stage V or end stage renal disease.
410.11	Acute myocardial infarction. Other anterior wall. Initial episode of care.
410.41	Acute myocardial infarction. Other bottom wall. Initial episode of care.
410.71	Acute myocardial infarction. Subendocardial infarction. Initial episode of care.

414.01	Coronary atherosclerosis. Native coronary artery.
425.40	Other primary cardiomyopathies.
426.00	Atrioventricular block, complete.
427.10	Paroxysmal ventricular tachycardia.
427.31	Atrial fibrillation and flutter.
428.00	Congestive heart failure, unspecified.
431.00	Intracerebral hemorrhage.
433.11	Occlusion and stenosis of the precerebral arteries. Carotid artery. With cerebral infarction.
434.01	Cerebral thrombosis. With cerebral infarction.
434.11	Cerebral embolism. With cerebral infarction.
434.91	Cerebral artery occlusion. With cerebral infarction.
435.90	Unspecified transient cerebral ischemia.
437.30	Cerebral aneurysm, nonruptured.
486.00	Pneumonia, organism unspecified.
519.80	Other diseases of respiratory system, not elsewhere classified.
532.40	Duodenal ulcer. Chronic or unspecified with haemorrhage. Without mention of obstruction.
562.12	Diverticula of intestine. Colon. With bleeding.
567.23	Spontaneous bacterial peritonitis.
569.30	Hemorrhage of rectum and anus.
571.20	Alcoholic cirrhosis of liver.
571.50	Cirrhosis of liver without mention of alcohol.
571.60	Biliary cirrhosis.
571.80	Other chronic nonalcoholic liver disease.
572.20	Hepatic coma.
572.30	Portal hypertension.
576.10	Cholangitis.
577.00	Acute pancreatitis.
578.90	Hemorrhage of gastrointestinal tract, unspecified.
584.90	Acute renal failure, unspecified.
585.60	End-stage renal disease.
585.90	Chronic kidney disease, unspecified.
590.10	Acute pyelonephritis. With renal medullary necrosis injury.
592.10	Calculus of ureter.
595.10	Chronic interstitial cystitis.
599.00	Urinary tract infection, site not specified.
600.20	Benign localized hyperplasia of prostate.
621.00	Polyp of corpus uteri.
733.00	Osteoporosis, unspecified.
780.54	Hypersomnia, unspecified.
V15.82	History of tobacco use.
V16.30	Family history of malignant neoplasm. Breast.
V18.00	Diabetes mellitus.
V42.00	Organ or tissue replaced by transplant. Kidney.
V45.11	Renal dialysis status.

V45.61	Cataract extraction status.
V45.89	Other postprocedural status. Other.
V53.31	Cardiac pacemaker.
V58.11	Encounter for antineoplastic chemotherapy.
V58.44	Aftercare following organ transplant.
V58.67	Long-term (current) use of insulin.
V59.40	Donors. Kidney.
V87.41	Personal history of antineoplastic chemotherapy.

LOINC codes

LOINC	Description
00704-7	Basophils [# /volume] in Blood by Automated count
00706-2	Basophils/100 leukocytes in Blood by Automated count
00711-2	Eosinophils [# /volume] in Blood by Automated count
00713-8	Eosinophils/100 leukocytes in Blood by Automated count
00718-7	Hemoglobin [Mass/volume] in Blood
00731-0	Lymphocytes [# /volume] in Blood by Automated count
00736-9	Lymphocytes/100 leukocytes in Blood by Automated count
00742-7	Monocytes [# /volume] in Blood by Automated count
00751-8	Neutrophils [# /volume] in Blood by Automated count
00770-8	Neutrophils/100 leukocytes in Blood by Automated count
00777-3	Platelets [# /volume] in Blood by Automated count
00785-6	Erythrocyte mean corpuscular hemoglobin [Entitic mass] by Automated count
00787-2	Erythrocyte mean corpuscular volume [Entitic volume] by Automated count
00789-8	Erythrocytes [# /volume] in Blood by Automated count
01742-6	Alanine aminotransferase [Enzymatic activity/volume] in Serum or Plasma
01755-8	Albumin [Mass/time] in 24 hour Urine
01920-8	Aspartate aminotransferase [Enzymatic activity/volume] in Serum or Plasma
01968-7	Bilirubin.direct [Mass/volume] in Serum or Plasma
01971-1	Bilirubin.indirect [Mass/volume] in Serum or Plasma
01975-2	Bilirubin.total [Mass/volume] in Serum or Plasma
01988-5	C reactive protein [Mass/volume] in Serum or Plasma
02039-6	Carcinoembryonic Ag [Mass/volume] in Serum or Plasma
02075-0	Chloride [Moles/volume] in Serum or Plasma
02085-9	Cholesterol in HDL [Mass/volume] in Serum or Plasma
02093-3	Cholesterol [Mass/volume] in Serum or Plasma
02157-6	Creatine kinase [Enzymatic activity/volume] in Serum or Plasma
02160-0	Creatinine [Mass/volume] in Serum or Plasma
02161-8	Creatinine [Mass/volume] in Urine
02162-6	Creatinine [Mass/time] in 24 hour Urine
02164-2	Creatinine renal clearance in 24 hour
02324-2	Gamma glutamyl transferase [Enzymatic activity/volume] in Serum or Plasma

02345-7	Glucose [Mass/volume] in Serum or Plasma
02498-4	Iron [Mass/volume] in Serum or Plasma
02571-8	Triglyceride [Mass/volume] in Serum or Plasma
02731-8	Parathyrin.intact [Mass/volume] in Serum or Plasma
02777-1	Phosphate [Mass/volume] in Serum or Plasma
02823-3	Potassium [Moles/volume] in Serum or Plasma
02842-3	Prolactin [Mass/volume] in Serum or Plasma
02857-1	Prostate specific Ag [Mass/volume] in Serum or Plasma
02951-2	Sodium [Moles/volume] in Serum or Plasma
02986-8	Testosterone [Mass/volume] in Serum or Plasma
03024-7	Thyroxine (T4) free [Mass/volume] in Serum or Plasma
03034-6	Transferrin [Mass/volume] in Serum or Plasma
03053-6	Triiodothyronine (T3) [Mass/volume] in Serum or Plasma
03084-1	Urate [Mass/volume] in Serum or Plasma
03091-6	Urea [Mass/volume] in Serum or Plasma
03167-4	Volume of 24 hour Urine
05193-8	Hepatitis B virus surface Ab [Units/volume] in Serum or Plasma by Immunoassay
05334-8	Rubella virus IgG Ab [Units/volume] in Serum or Plasma by Immunoassay
05388-4	Toxoplasma gondii IgG Ab [Units/volume] in Serum or Plasma by Immunoassay
05763-8	Zinc [Mass/volume] in Serum or Plasma
05821-4	Leukocytes [# /area] in Urine sediment by Microscopy high power field
05905-5	Monocytes/100 leukocytes in Blood by Automated count
06301-6	INR in Platelet poor plasma by Coagulation assay
06690-2	Leukocytes [# /volume] in Blood by Automated count
06768-6	Alkaline phosphatase [Enzymatic activity/volume] in Serum or Plasma
06875-9	Cancer Ag 15-3 [Units/volume] in Serum or Plasma
10334-1	Cancer Ag 125 [Units/volume] in Serum or Plasma
11011-4	Hepatitis C virus RNA [Units/volume] (viral load) in Serum or Plasma by Probe and target amplification method
12841-3	Prostate Specific Ag Free/Prostate specific Ag.total in Serum or Plasma
13457-7	Cholesterol in LDL [Mass/volume] in Serum or Plasma by calculation
13980-8	Albumin/Protein.total in Serum or Plasma by Electrophoresis
14158-0	Cortisol [Mass/time] in 24 hour Urine
14805-6	Lactate dehydrogenase [Enzymatic activity/volume] in Serum or Plasma by Pyruvate to lactate reaction
15067-2	Follitropin [Units/volume] in Serum or Plasma
17849-1	Reticulocytes/100 erythrocytes in Blood by Automated count
19123-9	Magnesium [Mass/volume] in Serum or Plasma
20447-9	HIV 1 RNA [# /volume] (viral load) in Serum or Plasma by Probe and target amplification method
24108-3	Cancer Ag 19-9 [Units/volume] in Serum or Plasma
30341-2	Erythrocyte sedimentation rate
50560-2	pH of Urine by Automated test strip
53962-7	Alpha-1-fetoprotein.tumor marker [Mass/volume] in Serum or Plasma
56477-3	Thyroperoxidase Ab [Units/volume] in Serum or Plasma by Immunoassay

56536-6	Thyroglobulin Ab [Units/volume] in Serum or Plasma by Immunoassay
----------------	---