ANNEX I. Characteristics and data quality control process of the EpiChron Cohort

The aim of this document is to provide detailed insight of the features of the EpiChron Cohort and the characteristics of the data collection, extraction and processing that are performed to ensure the quality and validity of the data contained. The data quality control process is conducted in three steps, which are described below together with the main characteristics of the EpiChron Cohort and of the regional health system in Aragon: 1) data collection, 2) data request and extraction, and 3) data processing.

1. Data collection

1.1. Scope of the cohort and context: The cohort is an exhaustive data repository of the public health service users in Aragon who represent 98% of total inhabitants in the region, and a demographically stable population. The population of Aragon is representative of the Spanish population in terms of sex, age and nationality distribution, although the population of Aragon is slightly older. As is the case at the national level, almost all inhabitants in Aragon (except for civil servants who are eligible to opt out of the public system, choosing fully private provision) are entitled to the public healthcare network and every individual has a family physician assigned who acts as the gatekeeper to the secondary and tertiary levels of care.

1.2. Data sources: The data collected in the EpiChron Cohort come directly from the primary sources of information, such as electronic health records and administrative databases filled in by healthcare professionals during daily practice. Specifically, the data stem from the Patient Index Database (BDU), the OMI-AP database from primary

care, the Basic Minimum Set of Data (CMBD) from hospital care, the registries from specialized care, the PCH database from the Emergency Room, and the Pharmaceutical Consumption Information System (Farmasalud) database. Common data collection software and procedures guarantee standardized data input by all healthcare professionals. The primary care information system in Aragon is regulated by legal order from 22nd September 2008 (Official Aragon Gazette number 165 from 9th October 2008). In the case of the hospital CMBD registry and the Farmasalud database, their completion is systematic, uniform and normative according to legal orders from 16th January 2001 (Official State Gazette number 11 from 26th January 2001), and from 17th December 2010 (Official Aragon Gazette number 17 from 20th January 2011), respectively.

1.3. Data collection and professionals involved: Those in charge of data collection and transcription are, among others, physicians, nurses, documentalists and administrative personnel. The main inaccuracies of these data arise from coding errors and/or data omission by professionals. Regarding the quality of primary care registries in Spain, the BIFAP Project (1) supported by the Ministry of Health and the Spanish Agency of Medicines and Medical Devices (AEMPS) deserves special attention. BIFAP aims to create a national public database for pharmacoepidemiological studies through the collaboration of general practitioner volunteers from different regions. For this purpose, BIFAP has been working for more than 15 years to train physicians to improve the quality of data registry, and they periodically (every six months) send feedback reports about data registry quality to each participant. A number of internal and external validation studies aimed at measuring the quality of diagnosis coding in

primary care and at comparing the incidence and prevalence of specific diseases, risk factors and drugs, have shown acceptable results (2). In Aragon, 10% of all general practitioners (85 family physicians and 35 pediatricians) who collect data that feed the EpiChron Cohort are also involved in the BIFAP Project.

Other studies beyond BIFAP conclude along the same lines, and have demonstrated that training of physicians can result in high codification quality in primary care (3). In Aragon, specific training and chart documentation on the use of electronic health record software is provided annually to physicians and nurses. The consistency of primary care electronic health record data is also favored by the high degree of work stability of general practitioners in Aragon.

The validity of Spanish primary care registries for use in research has been assessed in a number of studies. The use of gold standards to validate diabetes mellitus and hypertension diagnosis from primary care electronic health records showed a substantial agreement, which justifies the use of such data for epidemiological studies of these conditions (4). This is also the case of cardiovascular risk factors and vascular disease (5), although heart failure is not so properly codified in primary care (6).

The hospital CMBD registry is filled in by a centralized documentation service, although there is no individual-level ad-hoc study on the reliability and accuracy of hospital registries in Aragon. The need to link information from multiple databases to obtain reliable data for research and for routine monitoring of the prevalence of chronic diseases has been highlighted (7). This is the case of conditions such as COPD, diabetes, hypertension, cerebrovascular disease, ischemic heart disease, asthma,

epilepsy and heart failure, in which a low concordance between primary care and hospital diagnoses has been observed (8).

2. Data request and extraction

Data in the EpiChron Cohort are obtained under request from specific administrative bodies through annual data extraction waves. The systematics of this process is as follows.

The data contained in each of the registries described previously is requested to a specific service of the Department of Health that centralizes all health information in Aragon, through a standardized protocol submitted during the month of January of every year. An independent committee subsequently assesses the data application and, in case of a favorable decision, the specific service performs a pseudonymization of the data to encrypt individual-level identification codes, protecting patients' privacy and complying with data protection laws. This new encrypted code is applied in all registries, enabling the linkage of data at the patient level. One copy of the databases is sent to the researcher and another copy is stored in a central computer server. Access to these files is restricted to members of the research group by a double entry password. This process is partially regulated by legal order from 1st April 2013 (Official Aragon Gazette number 88 from 8th May 2013).

3. Data processing

Given that original databases are in different formats (e.g. access, excel, plain text), the Structured Query Language (SQL) programming language is employed to extract the data that will later conform the EpiChron Cohort. The Stata Statistical Software

(Release 12. College Station, TX: StataCorp LP) is used for data processing, as explained below. Of note, the research group hosting the cohort is a multidisciplinary qualified team including public health specialists, epidemiologists, clinicians (from primary and specialized care), pharmacists, statisticians, and data managers. They are all trained in data management and patient data protection.

The data processing includes a number of systematic steps aimed at improving the quality, accuracy and reliability of the data for research purposes. All changes conducted in the cohort data are kept track of in Stata scripts, and are continuously revised and updated given the dynamic nature of the data processing.

3.1 Quality control of diagnoses: This step aims to verify the correspondence between a diagnostic code and its open-text descriptor (i.e. a separate section of the electronic health record where physicians describe the symptoms and specific features of a diagnostic episode). Non-existent codes are also redefined and/or deleted. It is also common that diagnostic codes assigned by physicians lack accuracy. Such is the case for diabetes, which is often coded in its generic form without specifying the type of diabetes, even if this information is available as free text. For 23,752 of a total of 76,784 diagnoses of diabetes mellitus in the EpiChron Cohort in 2011, the type of diabetes was not coded. Drug prescription and dispensation data can also be considered to determine the presence of a given chronic condition. In the case of diabetes, the absence of dispensation of insulin was used as a complementary criterion for the selection of patients with type II diabetes, as well as the dispensation of sulfonylureas, glucosurics, glitazones, or DDP4 inhibitors, or the treatment with lente

insulin therapy without another antidiabetic. This allowed for the relocation of 883 diagnoses of type I diabetes and 15,842 diagnoses of type II diabetes.

Specific algorithms are employed to search for specific key words or roots of words in open-text fields. At the moment, this is exclusively done for specific chronic diseases as it is the case of diabetes, COPD, heart failure, dementia, or stroke, given the time needed to complete this task and the specific focus of the cohort on the epidemiological study of chronic conditions.

The reliability of the diagnoses in the EpiChron Cohort is also enhanced by the combined use of primary care and hospital records. In the case of dementia (9), we have shown the added value of integrating different data sources feeding the cohort to obtain a global and more accurate view of the epidemiology of this chronic condition.

3.2. Quality control of patient general data (DGPs): It refers to data that are not systematically collected for all patients, such as clinical parameters from laboratory analytical tests (e.g. blood lipids –total cholesterol, HDL and LDL–, hematocrit, glycosylated hemoglobin –HbA1c–, urine albumin, and spirometry parameters such as forced vital capacity –FVC– and forced expiratory volume –FEV–), anthropometric measurements and health indicators (e.g. body mass index –BMI–, height, weight, blood pressure, and heart rate), and lifestyle factors (e.g. drinking and smoking habits). These variables need to be managed carefully given their high rates of missingness that rarely occurs completely at random. Therefore, they are only used in specific studies after appropriate multiple imputation procedures. The process of quality control of DGPs focuses mainly on the detection of outlier values for any of the parameters. The different thresholds and value ranges for each parameter have been agreed upon

within the multidisciplinary research team, taking into account the clinical experience and related literature.

In regard to drinking and smoking habits in 2011, there were 224,765 valid measurements corresponding to 194,907 patients regarding tobacco consumption (yes/no), whereas specific information about number of cigarettes per day was only available for 47,572 patients. There was information about alcohol consumption (yes/no) in 177,844 patients, whereas 34,377 patients had also quantitative records available (grams of alcohol per week). In these cases, one unit alcohol intake is assumed to correspond to 8 g of alcohol, and records of less than 8 g intake are transformed into cero units of alcohol intake.

Clinical thresholds and valid value ranges agreed for each variable are shown in Table 1. The outliers records and those that could not been converted to numerical format were considered as non-logical values and therefore treated as missing data. Regarding weight, height and BMI records, there were valid measurements available for 27%, 17% and 20% of the total cohort population, respectively. In the case of pulse records, and considering that resting heart rate in a healthy adult ranges from 60 to 100 beats/min, 20% of the cohort had at least one valid record. Similarly, there were valid total cholesterol and hematocrit values for approximately 20% of individuals. In regard to glycosylated hemoglobin and urine albumin values, only 4-5% of the individuals had a valid record, respectively. This percentage decreased to 0.2% for spirometry parameters, which were only available for approximately 3,000 individuals from the cohort.

Table 1. Clinical thresholds used to validate patient general data (DGP) in the EpiChron Cohort in 2011.

DGP	Lower-upper valid thresholds	Initial number of records	Number of records transformed into missing	Final number of valid records	Final number of patients with valid records
Weight[a] (kg)	25 - 180	884,553	1,849	882,704	336,646
Height[a] (cm)	140 - 216	464,160	26,317	437,843	219,065
BMI[b] (kg/m^2)	10 - 75	460,445	1,547	458,898	246,135
Pulse (beats/min)	30 - 170	786,008	842	785,166	255,171
Total cholesterol (mg/dl)	43 -970	373,151	598	372,553	291,131
HDL[c] (mg/dl)	9 - 200	343,517	411	343,106	271,975
LDL[d] (mg/dl)	30 - 500	330,196	3,536	326,660	262,011
Hematocrit (%)	7 - 70	323,403	533	322,870	262,708
HbA1c[e] (%)	4 - 20	80,674	349	80,325	62,414
Urine albumin (mg/g)	≥ 0	74,166	6,367	67,799	57,145
FVC[f] (%)	≥ 14	3,478	349	3,129	2,962
FEV[g] (%)	≥ 20	3,513	241	3,272	3,082
Ratio FEV/FVC (%)	26 - 100	3,538	604	2,934	2,743

[a]Lower thresholds were only applied for individuals over 15 years of age.

[b]Body mass index.

[c]High-density lipoproteins.

[d]Low-density lipoproteins.

[e]Glycosylated hemoglobin.

[f]Forced vital capacity.

[g]Forced expiratory volume.

3.3. Grouping of chronic diseases into broader categories: Data are originally coded according to international codification systems both in the hospital registry (International Classification of Diseases, Ninth Revision, Clinical Modification, ICD-9-CM) and in primary care records (International Classification of Primary Care, First edition, ICPC-1). Mapping algorithms between different classification systems (e.g. ICD-9 to ICD-10, ICD-9 to ICPC-1) are occasionally employed to establish cross-national comparisons. Diagnoses are moreover grouped into Expanded Diagnostic Clusters (EDCs) using the ACG System®, which is used internationally in clinical management and health services research (10). Although the original diagnostic codes are available in the EpiChron Cohort, ECDs are better suited for the study of multimorbidity, since similar diagnoses are merged using data both from primary care and the hospital setting, increasing the manageability and reliability of the diagnostic data.

3.4. Creation of new variables: New variables are continuously generated based on existing variables in the EpiChron Cohort, according to their relevance for specific studies on chronic diseases and multimorbidity. Below are some few examples.

I) Socio-demographic variables. Area of residence and immigrant status are two examples of newly created variables. In the first case, the population is classified as living in a rural or urban area according to the location of a specific primary care health center. In the second case, the population is classified as native or immigrant taking into account other variables such as nationality, country of birth and length of stay in the host country. The specific algorithms for data transformation are based on the team's expertise.

II) Clinical variables. The presence of multimorbidity and polypharmacy are another two examples that required research team consensus for their operationalization. The number and type of conditions or drugs, but also the level of aggregation of diseases (e.g. chronic respiratory diseases versus separate codes for asthma, COPD, bronchitis, or emphysema), are common decisions taken within the research group.

III) Use of health services. Some examples are the total number of specialist visits, the number of visits to different specialties, the number of 7-day and 30-day hospital re-admissions, and/or the number of hospital admissions due to Ambulatory Care Sensitive Conditions –ACSC– (11). These variables are commonly used as proxies of negative health outcomes or inappropriate use of health services.

IV) Drug prescription and dispensation. Thanks to the expertise of clinicians and pharmacists within the group, different variables have been created related to the adherence to drugs and the level of compliance and persistence, which can be indirectly calculated based on the medication possession ratio.

3.5. Regular monitoring of aggregate clinical and drug data: Figures of the prevalence and incidence of chronic conditions, the use of health services, and drug prescription and dispensation are assessed periodically to evaluate their consistency with the literature and official national and regional reports, as a means of external validation of the cohort. Below we show the prevalence of some of the most frequent chronic diseases in the EpiChron Cohort (year 2011), standardized to the Spanish population and compared with those reported in the literature for Spain (Table 2). In general, prevalence rates in the EpiChron Cohort are similar to those reported in other sources.

The differences observed might be due to the fact that in those reports only population over 15 years of age is taken into account.

Table 2. Comparison of prevalences of specific chronic conditions in the EpiChron Cohort (2011) standardized to the Spanish population in terms of sex and age to those reported for Spain in different reports. Prevalences of external reports refer only to population aged 15 years and over.

Chronic conditions	**EpiChron Cohort (%)**	**Spain (%)**	**Source, year**
Hypertension	16.1	18.4	SNIE[a], 2011
Cholesterol	15.4	16.5	SNIE, 2011
Diabetes	5.7	6.8	SNIE, 2011
Eczema	9.3	4.8	SNIE, 2011
Varicose veins	6.8	9.2	SNIE, 2011
Chronic heart failure	0.83	0.99	Galindo et al., 2011 (12)
Chronic allergies	11.3	13.3	SNIE, 2011
Asthma	4.8	4.5	OECD[b], 2014 (13)
COPD	2.1	3.3	OECD, 2014 (13)
Dementia	1.2	1.8	OECD, 2015 (13)

[a]SNIE: Spanish National Institute of Statistics.

[b]OECD: Organisation for Economic Co-operation and Development.

Bibliography

1. Salvador Rosa A, Moreno Pérez JC, Sonego D, García Rodríguez LA, Abajo Iglesias FJ de. El Proyecto BIFAP: Base de datos para la Investigación Farmacoepidemiológica en Atención Primaria. *Aten Primaria*. 2003;31(10):655–661.

2. Agencia Española de Medicamentos y Productos Sanitarios. BIFAP » Base de datos para la Investigación Farmacoepidemiológica en Atención Primaria [Internet]. [cited 2017 Sep 15]. Available from: http://www.bifap.org/publicaciones.php

3. Orueta JF, Urraca J, Berraondo I, Darpón J. ¿Es factible que los médicos de primaria utilicen CIE-9-MC? Calidad de la codificación de diagnósticos en las historias clínicas informatizadas. *Gac Sanit*. 2006 May;20(3):194–201.

4. Burgos-Lunar C de, Salinero-Fort MA, Cárdenas-Valladolid J, et al. Validation of diabetes mellitus and hypertension diagnosis in computerized medical records in primary health care. *BMC Med Res Methodol*. 2011;11(1):146.

5. Ramos R, Balló E, Marrugat J, et al. Validity for Use in Research on Vascular Diseases of the SIDIAP (Information System for the Development of Research in Primary Care): the EMMA Study. *Rev Esp Cardiol*. 2012 Jan 1;65(1):29–37.

6. Verdú-Rotellar JM, Frigola-Capell E, Alvarez-Pérez R, et al. Validation of heart failure diagnosis registered in primary care records in two primary care centres in Barcelona (Spain) and factors related. A cross-sectional study. *Eur J Gen Pract*. 2017 Oct 2;23(1):107–113.

7. Orueta JF, Nuño-Solinis R, Mateos M, Vergara I, Grandes G, Esnaola S. Monitoring the prevalence of chronic conditions: which data should we use? *BMC Health Serv Res*. 2012 Dec 22;12(1):365.

8. Revilla-lópez C, Calderón-Larrañaga A, Enríquez-Martín N, Prados-Torres A. Baja concordancia entre la información clínica de atención primaria y hospital. *Aten Primaria*. 2016;48(4):244–50.

9. Marta-Moreno J, Obón-Azuara B, Gimeno-Felíu L, Achkar-Tuglaman NN, Poblador-Plou B, Calderón-Larrañaga A, et al. Concordancia del registro de demencia en las principales fuentes de información clínica. *Rev Esp Geriatr Gerontol*. 2016;51(5):276–9.

10. Starfield B, Kinder K. Multimorbidity and its measurement. *Health Policy*. 2011;103;3–8.

11. Purdy S, Griffin T, Salisbury C, Sharp D. Ambulatory care sensitive conditions: terminology and disease coding need to be more specific to aid policy makers and clinicians. *Public Health*. 2009;123(2):169–73.

12. Galindo Ortego G, Esteve IC, Gatius JR, Santiago LG, Lacruz CM, Soler PS. Pacientes con el diagnóstico de insuficiencia cardiaca en Atención Primaria: envejecimiento, comorbilidad y polifarmacia. *Aten Primaria*. 2011 Feb 1;43(2):61–67.

13. OECD/EU (2016). Health at a Glance: Europe 2016: State of Health in the EU Cycle, OECD Publishing, Paris.