French administrative health care database (SNDS) The value of its enrichment

Lucie-Marie Scailteux, Catherine Droitcourt, Frédéric Balusson, Emmanuel Nowak, Sandrine Kerbrat, Alain Dupuy, Erwan Drezen, André Happe, Emmanuel Oger

Thérapie

French administrative health care database (SNDS): the value of its enrichment.

short running title: enriching SNDS

Lucie-Marie SCAILTEUX[1,2], Catherine DROITCOURT[2,3], Frédéric BALUSSON[2], Emmanuel NOWAK[4], Sandrine KERBRAT[2], Alain DUPUY[2,3], Erwan DREZEN[2], André HAPPE[2,5], Emmanuel OGER[1,2]

1. Pharmacovigilance, Pharmacoepidemiology and Drug Information Center, Rennes University Hospital, 35000 Rennes, France
2. Univ Rennes, EA 7449 REPERES « Pharmacoepidemiology and Health Services Research », 35000 Rennes, France
3. Dermatology Department, Rennes Hospital University, 35000 Rennes, France
4. Université de Bretagne Loire, Université de Brest, INSERM CIC 1412, CHRU de Brest, 29200 Brest, France
5. Centre de Données Cliniques, CHRU de Brest, 29200 Brest, France

Corresponding author:
Lucie-Marie SCAILTEUX
CRPV de Rennes, CHU de Rennes
Rue Henri Le Guilloux – 35000 RENNES
Luciemarie.scailteux@chu-rennes.fr

Figures, tables: 0

Abstract (298 words)

SNIIRAM/SNDS, the French administrative health care database, covers around 99 % of the population. Its main limitation is the absence of clinical information and biological results. This report exposes the value of SNIIRAM/SNDS enrichment by external databases, and the linkage issues. It is illustrated by examples: the well-known population-based cohort CONSTANCES created to answer to epidemiological research questions with a specific interest on occupational and social factors, chronic diseases, and aging; the CANARI study, a regional-based study that collected Gleason score in all pathology laboratories in Brittany and then, linked pathology results to an *ad hoc* extraction from SNIIRAM database; the goal was to investigate the risk of high grade prostate cancer in patients treated by 5-alpha-reductase inhibitors for a symptomatic benign prostatic hyperplasia; the SACHA study, that identified and medically validated major bleeding event referred to emergency wards, then linked those clinical data to SNIIRAM; the goal was to minimize misclassification bias when estimating bleeding risk in patients who were prescribed antithrombotic drugs; the ISO-PSY study linked the SNIIRAM with the national cause of death registry (CépiDc) and the nationwide emergency department surveillance system (OSCOUR® Network) to investigate the potential link between isotretinoin and suicidal risk; the EFEMERIS cohort that assesses drugs prescriptions in French pregnant women who delivered in the Haute-Garonne region; the EPI-GETB-AM study that derived a SNIIRAM/SNDS-based algorithm to identify venous thromboembolism and linked SNIIRAM/SNDS to the EPI-GETBO-III survey for validation. Another perspective of SNDS enrichment is clinical trials' data for medico-economic assessment, and extended follow-up without attrition bias. Linkage is not straightforward. Apart from regulatory approbation and authorized data center issues, which could be solved by the Health Data Hub Initiative, a multidisciplinary team with medical, pharmacological and methodological knowledge, as well as with technical skills is essential to handle the whole process.

Key words: SNIIRAM, SNDS, administrative database, population-based study, linkage, matching.

Abbreviations

ALD: Affection de Longue Durée = Long term disease

BPH: benign prostate hyperplasia

CDW: Clinical Data Warehouses

CépiDC: Centre d'épidémiologie sur les causes médicales de décès

CNAMTS: Caisse Nationale d'Assurance Maladie des travailleurs salariés

CNAV: Caisse Nationale d'Assurance Vieillesse

CNIL: Commission Nationale de l'Informatique et des Libertés

EGB : Echantillon Généraliste des Bénéficiaires

GPRD: General Practice Research Database

HSC: Health Screening Center

ICD-10: International Classification of Diseases and related health problem, 10th revision

INDS : Institut National des Données de Santé

INSEE : Institut National de la Statistique et des Etudes Economiques

NIR: Numéro d'Inscription au Répertoire

NYHA: New York Heart Association

PCa: prostate cancer

PMSI: Programme de Médicalisation des Systèmes d'Information

PSA: prostate specific antigen

SNDS: Système National des Données de Santé

SNIIRAM: Système National d'Informations Inter-Régimes de l'Assurance Maladie

TNM: Tumor, lymph Nodes, Metastasis

UK: United Kingdom

VKA: vitamin K antagonist

Introduction

Taking into account various data on thousand patients in epidemiological studies arouses expectations as regards understanding disease risk factors, and improving access to care, disease prognosis or personalized medicine. Since 2000's, around the World, Pubmed includes an increasing number of publications using population-based study design. Whereas some questions aim to compare groups, through randomised trials for example, others require population representativeness. Population-based studies enter in this second category.

The specificity of French administrative health care database, called SNIIRAM (Système National d'Informations Inter-Régimes de l'Assurance Maladie) is that it is nationwide, covering around 99 % of the French population (1). In 2016, a modernisation healthcare system law extended SNIIRAM area establishing then the SNDS (Système National des Données de Santé)(2). Access to dedicated extraction from SNDS was authorised in 2017 pending a reinforced security process endorsement for data storage (3).

Schematically, SNDS contains outpatient data (date of reimbursed drugs dispensation and number of unit; attribution of long-term chronic disease (ALD); date and nature of paramedical interventions, of laboratory tests; date of birth; gender; date of death…) linked with hospitalisation data from "PMSI" (admission date, duration, ICD-10 codes for main and associated diagnosis, medical acts…) through a unique personal identification number (NIR). Causes of death, which were previously collected by an independent national registry (CépiDC registry)(4), are progressively integrated to the SNDS. Massive datasets challenge data managers and researchers to understand complex and miscellaneous information. Drezen et al. has developed tools (ePEPS toolbox) to make easy data visualisation, and users exploration and analysis (5,6). Otherwise, the Echantillon Généraliste des Bénéficiaires (EGB) is a $1/97^{th}$ dynamic representative sample of the SNIIRAM (around 660 000 persons) as regards gender, age, occupation and health expenses (1); however, EGB could not be considered representative on a regional scale.

According to the research question and especially the number of subjects needed, the method design will require the use of the national database rather than the EGB database. For example, to estimate the association between brand-to-generic antiepileptic drugs substitution and seizure-related hospitalisation (7), considering further that seizure is a rare event, use the national database was the only way to answer to the question. In other

research questions as regards frequent disease and drug highly prescribed (8,9), EGB appears a more appropriate tool.

The main limitation of SNDS and EGB is the absence of clinical information and biological results. The lack of arterial blood pressure, haemoglobin rate, disease stage (TNM score for cancer, NYHA stage for heart failure…) or MRI results for example, considering they could be confounders or influence patients' selection, could limit results interpretation (10).

Clinical Data Warehouses (CDW), allowing secondary use of healthcare data produced during the care process, could bring a solution to enrich medico-administrative databases with relevant medical information. However, the numerous solutions (11–16) proposed, their lack of interoperability and their still modest coverage at a nationwide level prevent, at this stage, CDW to bridge that gap.

Some authors highlighted a potential selection bias due to unmeasured confounding variables (especially prostate specific antigen (PSA) rate in prostate cancer) and that some statistical tools, such as propensity score, were useless "when there are major differences in non-measurable confounding variables" (17,18). Another SNDS limit concerns hospitalised patients where no details on the drugs used during hospitalisation are available (except for some expensive drugs). Finally, the accuracy of some diagnoses could be questioned as well as the onset date of disease considering the way to identify them in SNDS: indeed, the diagnosis of most chronic diseases and of some acute events are based on ICD-10 codes of hospital stay identified through the PMSI. For note, when PMSI was created in 1982, the aim was to measure medical facility activity and thus to estimate the budget allocation. In 2005, a new remuneration system ("Tarification à l'activité") was introduced and based on several variables classified by hospital stay group which have potentially diluted some diagnosis information. Thus, studies to validate disease identification algorithm make senses but requires linkage of SNDS to reference data.

The aim of this report is to expose the value of enrichment of SNDS by external databases, and the linkage issues. It is illustrated by examples and two types of studies: association study used to estimate a risk or a benefit, and validation of disease identification algorithm studies.

The CONSTANCES cohort

In 2008, the population-based cohort CONSTANCES was set up to answer epidemiological research questions with a specific interest on occupational and social factors, chronic diseases, and aging. It is designed as a representative (on age, gender and socio-demographic status) sample of the 18-69 years French adult population affiliated to the national Health Insurance Fund (CNAMTS and CNAV) (19); self-employed and agricultural workers are not included due to a different health insurance affiliation. Eventually, this cohort plans to include around 200 000 voluntary participants, considered representative of the French population affiliated to Health Insurance general scheme. Patients were invited to join the closer health screening center to have full-examination, with health examination scheduled every 5 years; in parallel, an annual self-administered questionnaire is sent to patients. A passive follow-up of patients is allowed thanks to annual linkage with SNIIRAM made by CNAMTS and CNAV; causes of death were provided by the national cause of death registry (CépiDC). Thanks to the three data sources, several information are available: socio-demographic data (social position, education and income level, employment and marital status, geocoding of residency address…), health (personal and family disease history, incident and prevalent diseases, handicap, weight, height, blood pressure, electrocardiogram, vision test, laboratory tests, cognitive tests…), behaviour (smoking, cannabis use, alcohol consumption, sexual orientation, physical activity, dietary habits, physical performance), and results from biological samples (blood and urine samples were collected only for patients included after 2016 due to budget restriction). The main CONSTANCES strength is to include numerous clinical, biological, socio-demographic and behaviour data, which are not available in the SNIIRAM / SNDS; it allows to investigate more in depth of effect of personal behaviour, access to care disparity or social determinants of health inequality on some chronic diseases (diabetes, cardiovascular diseases...) for example (20–22). Among limitations, representativeness of the cohort could be debated: agricultural and self-employed workers are not included; yet, in other developed countries, studies have shown that rural populations have more prevalent asthma, untreated chronic pulmonary disease, have less vaccination or medical care provider (23,24). Otherwise, patient selection did not include younger nor elderly and frailty patients of whom behaviour and access to care are specific but also more complex to follow in the SNDS database (parents' health insurance affiliation for children; accommodation in nursing home for frailty people…). As a

voluntary participation, some specific behaviour (social excluded, heavy drinkers…) patients will be under-represented. Eventually, the cohort size does not allow to investigate rare outcomes: Zins et al. evoked a sufficient power for circumstances with at least 5 years of follow-up, prevalence exposure greater than 10 %, and outcomes with annual incidence greater than 10/100 000 and (19). Investigating the risk of ischemic events such as myocardial infarction or stroke for androgen deprivation therapy in men with prostate cancer as previously done thanks to SNIIRAM (25) would have lacked power using the CONTANCES cohort.

SNIIRAM / SNDS linkage with *ad hoc* databases

Contrary to the CONSTANCES cohort which linked several data sources adapted to broad extent of research questions, we present thereafter several examples of very focused questions where the method was developed especially for one objective and used adequate data collection.

As a first example, we describe the CANARI study (26,27). It is a 2010-2013 study investigating the risk of high-grade prostate cancer in patients treated by 5-alpha-reductase inhibitors for a symptomatic benign prostatic hyperplasia (BPH) and we needed Gleason score: this score is provided after pathologist assessment on prostate biopsy sample, and not available in SNDS; furthermore, there is no national pathology registry. Considering it was infeasible on a national basis (even using the EGB database) and that BPH and prostate cancer are frequent diseases in men after 50 years, we decided to set up a regional-based study and to collect the Gleason score in all pathology laboratories in Brittany; then, NIR being not available, to link pathology results and SNIIRAM, we used two keys: 1) month and year of men birthday; 2) date of prostate sampling in SNIIRAM, and date of prostate sample reception in pathology laboratory database (or date of sampling in case of reception date missing) (26). We authorised a 3-day (then 4-day) relaxation between prostate sample date and prostate examination date to make linkage easier. From around 75,000 patients in SNIIRAM database (regional extraction) and 14,000 prostate examinations from pathology laboratories, we found 859 patients with incident prostate cancer with Gleason score of whom 767 were eligible for analysis (105 Gleason $\geq$ 8, and 662 Gleason < 8)(27). Even if we could discuss the representativeness of this study, conducting a similar design on a national scale (linking the whole SNIIRAM or the EGB representative random sample with all the

French pathology laboratories) would have been unrealistic owing to financial and human costs. Even, a yet to be built national network of hospital-based clinical data warehouse could not be the solution as most results came from private laboratories. As a reminder, the male population in Brittany contains around 1.5 million people.

As a second example, we describe the SACHA study which questioned the validity of the disease ICD-10 codes in the hospital stays. Considering that emergency departments are the front door for major bleeding outpatients and that hospital discharge diagnosis could not be accurate enough to categorise outpatient major haemorrhagic event, we set up a 2012-2015 prospective population-based cohort study linking SNIIRAM database with data from all emergency departments in Brittany to estimate the risk of major bleeding in patients treated with antithrombotic drugs (28). Previous studies have shown that ICD codes used to identify haemorrhagic complications in hospital discharge abstract could not be specific enough: according to a specific location (gastrointestinal tract, genitourinary tract, intracranial…) and hospital coding rules, positive predictive value could widely vary from less than 50 % to more than 90% (29–31). In SACHA study, each major bleeding included in the "emergency database" was medically validated and carefully checked afterwards. This minimized misclassification, which was one of the strength of this study compared to other observational studies (32,33). Regarding the linkage, several keys were used: 1) month and year of birthday; 2) date of hospital entry and date of discharge; 3) health facility identification number; 4) type of antithrombotic drug (platelet inhibitors, vitamin K antagonist (VKA), non-VKA oral anticoagulants). Despite this, some bleeding events remained unmatched. One of the explanations concerned dependant patients living in accommodation establishment where drugs are dispensed by internal use pharmacy and not identified in SNIIRAM.

SNIIRAM / SNDS linkage with existing databases

As an example, we describe through the ISO-PSY study the linkage of SNIIRAM with the national cause of death registry (CépiDc) to investigate the potential link between isotretinoin and suicidal risk (34,35). As a reminder, isotretinoin is a very efficacious oral treatment given for severe acne, a high prevalent skin condition affecting 15 to 20 % of the adolescents and young adults. Concerning the safety of isotretinoin, two risks are closely monitored: a teratogenic effect which is well-established, with the implementation of a

pregnancy prevention plan in 1997, and a potential and debated risk of psychiatric disturbances, in particular suicidal behaviours. Suicidal behaviours included suicidal ideations, suicide attempts and completed suicides. To answer the question, the SNIIRAM was used, over the 2010-1015 period, by assessing the risk of suicide attempt among patients aged 10 to 50 years with at least one delivery of oral isotretinoin. In the SNIIRAM, only the hospitalized suicide attempts could be studied, and it was considered as an incomplete approach of the suicidal risk assessment. Complementary analyses on completed suicides and on suicide attempts that do not lead to hospitalization were needed. Since the causes of death were not available within the SNIIRAM over this period, we planned to link it with the national causes of death registry (CepiDC registry) (4) using the following key variables: gender, month and year of birth, death date, ZIP code of residence. Among 1918 individuals who died, 1075 were perfectly matched and 248 were matched by removing one missing matching variable (mainly ZIP code of residence), leading to a matching rate of 69%.

In a second step, some specific situations of suicide attempts were investigated: individuals not seeking medical cares, individuals only seen in medical consultation, individuals seen in an emergency unit and who stay less than 8 hours (and then not coded as hospitalized). A French nationwide emergency department surveillance system, named OSCOUR® (OSCOUR® Network) and based on visits in an emergency unit, has been developed, for more than ten years, to detect unexpected public health events and follow trends of expected events (seasonal, for instance flu) (36). Data collection used ICD diagnostic codes as reported in emergency wards. Hence, in order to identify suicide attempts seen in an emergency unit, the OSCOUR database was linked to the SNIIRAM database using the following key variables: gender, month and year of birth, health facility identification number, entry date, and ZIP code of residence. Eventually, the linkage between SNIIRAM and these two databases (CépiDC and OSCOUR®) offers the possibility to study more widely and accurately suicidal risk under isotretinoin.

In 2004, the EFEMERIS cohort was built to assess, describe and follow the drugs prescriptions in pregnant women who delivered in the French Haute-Garonne region (37). Main interest of this cohort is to provide data as regards children handicap, neonatal diseases and prevent malformation following pregnancy drug exposure. Several data sources were merged: 1) SNIIRAM database to obtain the drug prescribed before and during pregnancy (outpatients) and mother hospitalization (including medical pregnancy

interruption, spontaneous abortion, fetal death) data through PMSI; 2) the newborn health status (APGAR score, height, weight, congenital malformation…) was obtain using data coming from the mother and child protection center database; 3) data as regards medical pregnancy interruptions were obtained through the multidisciplinary prenatal diagnosis center database. Data anonymization was made in a similar way in the databases and provided the same encoded number allowing a not too complex database linkage. Annually, around 10 000 new pregnancies are included in the EFEMERIS cohort. Among the studies recently conducted on drug exposure during pregnancy, a particular focus was made on antidepressants, psychotropic drugs or asthma medications (38–40); other published works concern the description of the newborn's growth or the risk of infection in the first year of life after *in utero* exposure to drugs acting on immunity (41,42). On the 2005-2014 period, around 89 000 pregnant women living in Haute-Garonne were included in the cohort (38).

Validation of disease identification algorithm

Using SNDS, whereas some diseases can easily be identified using drugs claims data combined with ICD-10 codes (diabetes for example), others appear more complicated to identify due to care evolution during years or outpatient care provided without specific proxies.

In the EPI-GETB-AM example, we showed the value of SNDS enrichment to validate venous thromboembolism (VTE) identification in SNDS. Indeed, except in United States of America and Northern Europe, few studies have estimated the incidence of VTE. Regarding the French population, VTE incidence in 2000 (per 1,000 person-year, 95%CI) was estimated to be 1.83 (1.69-1.98) in the Brest population (Bretagne, France) (43). In order to obtain more recent data, a new survey, EPI-GETBO-III, was set up in 2013 to include all the VTEs in the Brest district (ClinicalTrials.gov Identifier: NCT02895971). Such a survey could not be conducted at a national level or with a long-term follow-up, considering financial and organizational constraints. Furthermore, VTE diagnosis process has changed over time: it is nowadays made more in an outpatient basis than through hospitalization. In order to construct and validate a VTE identification algorithm, SNIIRAM was linked to the EPI-GETBO-III using the last one as gold-standard. As a first step, a nine-month period (from the April 1ST 2013 to December 31st 2013) was used to construct the algorithm, and a twelve-month period (January 1st to December 31st 2014) was used for validation. The statistical unit used

was a VTE event defined by the occurrence of a deep vein thrombosis (DVT) or a pulmonary embolism (PE); other thrombotic events (muscle thrombosis, superficial venous thrombosis…) were not considered. The EPI-GETBO-III database (reference diagnosis) contained information such as year of birth, gender, ZIP code at the time of VTE occurrence, VTE date and type (EP and / or DVT), acts performed as well as place and date of VTE diagnosis.

Three ways were explored during the construction phase to identify VTE: 1) Identification of the beginning of a curative treatment (a first delivery was defined by no dispensing in the previous 3 months) along with an imaging act (thought as a VTE diagnosis) within a pre-specified time-frame; 2) Identification of a specific therapeutic intervention: thrombolysis or thrombectomy of adequate localization; 3) Identification of a hospital stay with a VTE code as main diagnosis. Over the period, the numbers of events identified by the algorithm and in EPI-GET-BO-III were compared, and the validation of the cases identified in the SNIIRAM required a linking to EPI-GETBO-III database. The key variables used for linking were the year of birth, gender, date of VTE and ZIP code of residency. The sensitivity of the algorithm has been defined as the number of events from EPI-GETBO-III with a hit among those identified by the algorithm, divided by the total number of events identified in EPI-GET-BO-III. The number of events identified by the algorithm but with no correspondence among EPI-GETBO-III, was also determined but no denominator was available to estimate specificity.

Discussion and conclusion

Among the numerous SNIIRAM/SNDS enrichments already made in France, we present here some examples of large scale enrichment and others based on specific questions, and thus showed the miscellaneous use in the medical context.

In the CANARI example, the Gleason score was the main not available data in the SNIIRAM database; linkage with a local pathology database was made to provide this data. Other studies investigating the same question and using health care databases obtained Gleason score thanks to the Finnish or Swedish Cancer Registry or directly from mailed baseline questionnaire (44–46).

The accuracy of major hemorrhagic events diagnosis was the cornerstone of the SACHA study; in the majority of previous studies, no medical validation was made (33) and there were few real-life studies with medical diagnosis validation (47–50).

In the ISOPSY example, the potential isotretinoin suicide risk was better apprehended by studying both hospitalized suicide attempts, completed suicide, and less severe suicidal behaviours such as suicide attempts leading to visits to emergency departments without hospitalization. The most important study on the issue was based on the data of a compassionate use program in Sweden to identify patients with isotretinoin dispensation. This study population was linked to the national patient register of in-hospital care and to the cause of death register, by using a unique personal identification number, maintained by the National Board of Health and Welfare, to identify hospitalized suicide attempts and suicides (51). This study published in 2010 was based on isotretinoin dispensation data from the eighties, a period where this potential risk was not known.

Other enrichments of administrative health care databases are used for general epidemiologic researches (CONTANCES cohort) or to investigate economic or social aspects. This is the case of the French “CARE-ménage” survey whose outcomes include the following of autonomy loss since 2000’s, the estimation of the family involvement with the elderly and of the economic burden left in elderly (52).

With the EPI-GETB-AM example, we showed the need of SNDS enrichment to validate VTE identification algorithm and the way to made it. Other examples of disease identification algorithms (peripheral arterial and venous thrombosis, infections responsible for hospitalization, acute coronary syndrome) were established using SNIIRAM database and validated using medical charts from regional hospitals in France (53–55).

Regarding primary care database, the best known example concerns the United Kingdom (UK) where the CPRD (Clinical Practice Research Datalink) and THIN databases provide information as regard less than 10 % of UK population with data registered by general practitioners (56,57); for note, a portion of patients included in the THIN database overlap those in the GPRD database. In France, such primary care databases do not exist but researchers tried to enrich SNIIRAM with primary care data: in 2014, Perlbarg et al. investigated the use of a research tool in ambulatory care matching the SNIIRAM database with primary care medical data (58). From 800 general practitioners (GP) belonging to the Société Française de Médecine Générale, a sample of 30 GP users of a specific software was selected in 2008 considering their years of experience, quality and exhaustiveness of medical case coding. The medical data included several sections: “patient” (date of birth, gender, living area…), “consultation” (context of care, date and type of consultation…), “diagnosis”

(symptoms and diagnosis with ICD-10 corresponding codes), and “decision” (prescription of drug, biological analyses, imaging, consultation in a specialist…). Six variables were used to link the two databases: professional practice number, consultation date, type of consultation, month and year of birth, gender. Eventually, around 80% of patients had been linked (around 29,000 patients and 89,000 consultations). SNIIRAM / SNDS does not include cause of GP or specialist consultation; only the date and type of medical doctor are available. Using symptoms and diagnosis data in primary care allow detecting more prematurely disease beginning rather than using SNDS. In the latter, disease diagnosis could be estimated through different ways: long term disease registration, hospital diagnoses, some specific drugs and laboratory tests. Taking the example of the CANARI study (27), BPH was identified using SNIIRAM through reimbursement of drug licensed for symptomatic or complicated BPH; the use of primary care national database if it had existed, would have made it possible to confirm the BPH diagnosis and give more details as regards the beginning of the disease and of medical treatment (it was not possible to estimate the drug initiation date for duration of use of 2 years or more), and the results of PSA dosage as well as the digital rectal examination. Taking the example of ISOPSY study, suicidal ideations or depressive disorders could be detected through diagnosis data in primary care database such as GPRD and allowed to assess more broadly the potential risk of psychological disorders under isotretinoin.

Another perspective of SNDS enrichment is clinical trials' data: such a linkage could offer a medico-economic assessment, and a longer follow-up of patients included, as seen in an US study by Unger et al. (59): the Prostate Cancer Prevention Trial (PCPT) (60) investigated the use of finasteride (5-alpha-reductase inhibitor) in prostate cancer prevention with a seven-year follow-up; in order to investigate wether the finasteride effect in the prostate cancer risk reduction was maintained, especially after finasteride discontinuation, authors linked PCPT clinical records with Medicaire’s participants to detect long-term outcomes (59). Such a linkage between clinical trials’ data and medico-administrative data would make easier the long-term follow-up as regards the financial and organizational aspects; furthermore, using especially SNDS data, attrition bias is limited contrary to trials.

In parallel of the medical aspect and data analysis, we have to keep in mind that SNDS database access and enrichment requires a data anonymization and a legal approbation, a not so easy process. Since the SNDS set up in 2016, a single office, INDS (Institut National des

Données de Santé), constitutes the front door to administrative health care databases access (61). INDS has four missions: 1) to improve and make easier data access procedure; 2) information exchange between data producers and data managers / users; 3) to assess the public interest of the proposed researches projects; 4) to provide data with poor risk of people re-identification after project examination and approbation by the Commission Nationale de l'Informatique et des Libertés (CNIL). Furthermore, with the multiplication of SNDS enrichment requests in recent years and the new security process established in 2017, we observed a reinforcement of data access and use, which forced researches teams to modify their infrastructure.

The question of SNDS enrichment makes sense in the current politics of the French government. Indeed, in recent years, we observe an enthusiasm for big data and artificial intelligence with specific application in medical field. Furthermore, in June 2018, the French Secretary of Health, Agnès Buzyn, mandate the INDS president, Dominique Polton, and the professor of medical informatics, Marc Cuggia, to collaborate to think about a "Health Data Hub" (62,63): the goal is to enrich SNDS, especially with hospital clinical data increasing the type of available data. Several public and private research partners are involved.

In conclusion, we present herein several examples of SNDS enrichment and show their values. One however should keep in mind that such process requires a legal approval and a specific regulatory circuit (but made easier since the last SNDS creation); furthermore, considering the specificity of population based-studies, the constitution of multidisciplinary teams including people with medical knowledge, as well as people with technical and methodological skills to handle the particular data and the data volume is essential.

Conflicts of interest

None.

Reference

1. Bezin J, Duong M, Lassalle R, Droz C, Pariente A, Blin P, et al. The national healthcare system claims databases in France, SNIIRAM and EGB: Powerful tools for pharmacoepidemiology. Pharmacoepidemiol Drug Saf. 2017 Aug;26(8):954–62.
2. LOI n° 2016-41 du 26 janvier 2016 de modernisation de notre système de santé. 2016-41 Jan 26, 2016.
3. Arrêté du 22 mars 2017 relatif au référentiel de sécurité applicable au Système national des données de santé.
4. Rey G. Les données des certificats de décès en France : processus deproduction et principaux types d'analyse. Rev Med Interne. 2016 Oct;37(10):685–93.
5. Drezen E, Guyet T, Happe A. From medico-administrative databases analysis to care trajectories analytics: an example with the French SNDS. Fundam Clin Pharmacol. 2018 Feb;32(1):78–80.
6. Happe A, Drezen E. A visual approach of care pathways from the French nationwide SNDS database - from population to individual records: the ePEPS toolbox. Fundam Clin Pharmacol. 2018 Feb;32(1):81–4.
7. Polard E, Nowak E, Happe A, Biraben A, Oger E, GENEPI Study Group. Brand name to generic substitution of antiepileptic drugs does not lead to seizure-related hospitalization: a population-based case-crossover study. Pharmacoepidemiol Drug Saf. 2015 Nov;24(11):1161–9.
8. Bezin J, Pariente A, Lassalle R, Dureau-Pournin C, Abouelfath A, Robinson P, et al. Use of the recommended drug combination for secondary prevention after a first occurrence of acute coronary syndrome in France. Eur J Clin Pharmacol. 2014 Apr;70(4):429–36.
9. Montastruc F, Bénard-Laribière A, Noize P, Pambrun E, Diaz-Bazin F, Tournier M, et al. Antipsychotics use: 2006-2013 trends in prevalence and incidence and characterization of users. Eur J Clin Pharmacol. 2018 May;74(5):619–26.
10. Moulis G, Lapeyre-Mestre M, Palmaro A, Pugnet G, Montastruc J-L, Sailler L. French health insurance databases: What interest for medical research? Rev Med Interne. 2015 Jun;36(6):411–7.
11. Murphy SN, Mendis ME, Berkowitz DA, Kohane I, Chueh HC. Integration of clinical and genetic data in the i2b2 architecture. AMIA Annu Symp Proc AMIA Symp. 2006;1040.
12. Murphy SN, Weber G, Mendis M, Gainer V, Chueh HC, Churchill S, et al. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). J Am Med Inform Assoc JAMIA. 2010 Apr;17(2):124–30.
13. Hripcsak G, Duke JD, Shah NH, Reich CG, Huser V, Schuemie MJ, et al. Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers. In: Studies in Health Technology and Informatics [Internet]. IOS Press; 2015 [cited 2018 Jun 28]. p. 574–8. Available from: https://tmu.pure.elsevier.com/en/publications/observational-health-data-sciences-and-informatics-ohdsi-opportun
14. Cuggia M, Garcelon N, Campillo-Gimenez B, Bernicot T, Laurent J-F, Garin E, et al. Roogle: an information retrieval engine for clinical data warehouse. Stud Health Technol Inform. 2011;169:584–8.
15. Garcelon N, Neuraz A, Salomon R, Faour H, Benoit V, Delapalme A, et al. A clinician friendly data warehouse oriented toward narrative reports: Dr. Warehouse. J Biomed Inform. 2018 Apr 1;80:52–63.
16. Delamarre D, Bouzille G, Dalleau K, Courtel D, Cuggia M. Semantic integration of medication data into the EHOP Clinical Data Warehouse. Stud Health Technol Inform. 2015;210:702–6.

17. Williams SB, Huo J, Chamie K, Smaldone MC, Kosarek CD, Fang JE, et al. Discerning the survival advantage among patients with prostate cancer who undergo radical prostatectomy or radiotherapy: The limitations of cancer registry data. Cancer. 2017 Jan;123(9):1617–24.
18. Mack Roach 3rd. Re: Morbidity and Mortality of Locally Advanced Prostate Cancer: A Population Based Analysis Comparing Radical Prostatectomy Versus External Beam Radiation. Eur Urol. 2018 Apr;73(4):638–9.
19. Zins M, Goldberg M, CONSTANCES team. The French CONSTANCES population-based cohort: design, inclusion and follow-up. Eur J Epidemiol. 2015 Dec;30(12):1317–28.
20. Feral-Pierssens A-L, Carette C, Rives-Lange C, Matta J, Goldberg M, Juvin P, et al. Obesity and emergency care in the French CONSTANCES cohort. PloS One. 2018;13(3):e0194831.
21. Merle BMJ, Moreau G, Ozguler A, Srour B, Cougnard-Grégoire A, Goldberg M, et al. Unhealthy behaviours and risk of visual impairment: The CONSTANCES population-based cohort. Sci Rep. 2018 Apr 26;8(1):6569.
22. Wiernik E, Meneton P, Empana J-P, Siemiatycki J, Hoertel N, Vulser H, et al. Cardiovascular risk goes up as your mood goes down: Interaction of depression and socioeconomic status in determination of cardiovascular risk in the CONSTANCES cohort. Int J Cardiol. 2018 Jul 1;262:99–105.
23. Earle-Richardson G, Scribani M, Scott E, May J, Jenkins P. A comparison of health, health behavior, and access between farm and nonfarm populations in rural New York state. J Rural Health Off J Am Rural Health Assoc Natl Rural Health Care Assoc. 2015;31(2):157–64.
24. Deligiannidis KE. Primary Care Issues in Rural Populations. Prim Care. 2017 Mar;44(1):11–9.
25. Scailteux L-M, Vincendeau S, Balusson F, Leclercq C, Happe A, Le Nautout B, et al. Androgen deprivation therapy and cardiovascular risk: No meaningful difference between GnRH antagonist and agonists-a nationwide population-based cohort study based on 2010-2013 French Health Insurance data. Eur J Cancer. 2017 May;77:99–108.
26. Scailteux L-M, Balusson F, Vincendeau S, Rioux-Leclercq N, Nowak E. Rationale and design of the CANARI study: a case-control study investigating the association between prostate cancer and 5-alpha-reductase inhibitors for symptomatic benign prostate hypertrophy by linking SNIIRAM and pathology laboratories in a specific region in France. Fundam Clin Pharmacol. 2018 Feb;32(1):120–9.
27. Scailteux L-M, Rioux-Leclercq N, Vincendeau S, Balusson F, Nowak E, Oger E, et al. Use of 5α-reductase inhibitors for benign prostate hypertrophy and risk of high grade prostate cancer: a French population-based study. BJU Int. 2018 Jul 19; doi: 10.1111/bju.14495
28. Major Bleeding Risk Associated With Antithrombotics - the SACHA study (Survey of acute haemorrhage with antithrombotic drugs) [Internet]. 2013 [cited 2018 Jun 17]. Available from: https://clinicaltrials.gov/ct2/show/NCT02886533
29. Arnason T, Wells PS, Walraven C van, Forster AJ. Accuracy of coding for possible warfarin complications in hospital discharge abstracts. Thromb Res. 2006 Jan 1;118(2):253–62.
30. Pisa F, Castellsague J, Drigo D, Riera-Guardia N, Giangreco M, Rosolen V, et al. Accuracy of International Classification of Diseases, 9th Revision, Clinical Modification codes for upper gastrointestinal complications varied by position and age: a validation study in a cohort of nonsteroidal anti-inflammatory drugs users in Friuli Venezia Giulia, Italy. Pharmacoepidemiol Drug Saf. 2013 Nov;22(11):1195–204.
31. Delate T, Jones AE, Clark NP, Witt DM. Assessment of the coding accuracy of warfarin-related bleeding events. Thromb Res. 2017 Nov;159:86–90.

32. Maura G, Blotière P-O, Bouillon K, Billionnet C, Ricordeau P, Alla F, et al. Comparison of the short-term risk of bleeding and arterial thromboembolic events in nonvalvular atrial fibrillation patients newly treated with dabigatran or rivaroxaban versus vitamin K antagonists: a French nationwide propensity-matched cohort study. Circulation. 2015 Sep 29;132(13):1252–60.
33. Weeda ER, White CM, Peacock WF, Coleman CI. Rates of major bleeding with rivaroxaban in real-world studies of nonvalvular atrial fibrillation patients: a meta-analysis. Curr Med Res Opin. 2016;32(6):1117–20.
34. Droitcourt C, Nowak E, Rault C, Le Nautout B, Happe A, Oger E, et al. Risk of suicide attempt associated with isotretinoin: a nationwide cohort and case-time-control study. 2nd European Dermato-Epidemiology (EDEN) Network - Berlin, 15-16 March 2018 [Internet]. 2018 [cited 2018 Jun 26]. Available from: http://www.dermepi.eu/?page_id=471
35. Droitcourt C, Nowak E, Rault C, Happe A, Le Nautout B, Kerbrat S, et al. Le risque suicidaire est-il majoré sous isotrétinoïne ? Analyse des données nationales de l'assurance maladie (SNIIRAM 2009–2016) - Journées Dermatologiques de Paris 2017. Ann Dermatol Vénérologie. 2017 Dec;144(12S):S58.
36. Josseran L, Fouillet A, Caillère N, Brun-Ney D, Ilef D, Brucker G, et al. Assessment of a syndromic surveillance system based on morbidity data: results from the Oscour network during a heat wave. PloS One. 2010 Aug 9;5(8):e11984.
37. Damase-Michel C, Lacroix I, Hurault-Delarue C, Beau A-B, Montastruc J-L, les partenaires d'EFEMERIS. Évaluation des médicaments chez la femme enceinte : à propos de la base de données française EFEMERIS. Therapie. 2014 Feb;69(1):91–100.
38. Hurault-Delarue C, Lacroix I, Bénard-Laribière A, Montastruc J-L, Pariente A, Damase-Michel C. Antidepressants during pregnancy: a French drug utilisation study in EFEMERIS cohort. Eur Arch Psychiatry Clin Neurosci. 2018 May 26;
39. Beau A-B, Didier A, Hurault-Delarue C, Montastruc J-L, Lacroix I, Damase-Michel C. Prescription of asthma medications before and during pregnancy in France: An observational drug study using the EFEMERIS database. J Asthma Off J Assoc Care Asthma. 2017 Apr;54(3):258–64.
40. Hurault-Delarue C, Damase-Michel C, Finotto L, Guitard C, Vayssière C, Montastruc J-L, et al. Psychomotor developmental effects of prenatal exposure to psychotropic drugs: a study in EFEMERIS database. Fundam Clin Pharmacol. 2016 Oct;30(5):476–82.
41. Beau A-B, Tauber M, Chollet C, Hurault-Delarue C, Bouilhac C, Montastruc J-L, et al. A contemporary description of French newborns' growth using the Efemeris cohort. Arch Pediatr Organe Off Soc Francaise Pediatr. 2017 May;24(5):424–31.
42. Palosse-Cantaloube L, Hurault-Delarue C, Beau A-B, Montastruc J-L, Lacroix I, Damase-Michel C. Risk of infections during the first year of life after in utero exposure to drugs acting on immunity: A population-based cohort study. Pharmacol Res. 2016;113(Pt A):557–62.
43. Oger E. Incidence of venous thromboembolism: a community-based study in Western France. EPI-GETBP Study Group. Groupe d'Etude de la Thrombose de Bretagne Occidentale. Thromb Haemost. 2000 May;83(5):657–60.
44. Murtola TJ, Tammela TLJ, Määttänen L, Ala-Opas M, Stenman UH, Auvinen A. Prostate cancer incidence among finasteride and alpha-blocker users in the Finnish Prostate Cancer Screening Trial. Br J Cancer. 2009 Sep 1;101(5):843–8.
45. Robinson D, Garmo H, Bill-Axelson A, Mucci L, Holmberg L, Stattin P. Use of 5α-reductase inhibitors for lower urinary tract symptoms and risk of prostate cancer in Swedish men: nationwide, population based case-control study. BMJ. 2013 Jun 18;346:f3406.
46. Preston MA, Wilson KM, Markt SC, Ge R, Morash C, Stampfer MJ, et al. 5α-Reductase inhibitors and risk of high-grade or lethal prostate cancer. JAMA Intern Med. 2014

Aug;174(8):1301–7.
47. Yavuz B, Ayturk M, Ozkan S, Ozturk M, Topaloglu C, Aksoy H, et al. A real world data of dabigatran etexilate: multicenter registry of oral anticoagulants in nonvalvular atrial fibrillation. J Thromb Thrombolysis. 2016 Oct;42(3):399–404.
48. Becattini C, Franco L, Beyer-Westendorf J, Masotti L, Nitti C, Vanni S, et al. Major bleeding with vitamin K antagonists or direct oral anticoagulants in real-life. Int J Cardiol. 2017 Jan 15;227:261–6.
49. Xu Y, Schulman S, Dowlatshahi D, Holbrook AM, Simpson CS, Shepherd LE, et al. Direct Oral Anticoagulant- or Warfarin-Related Major Bleeding: Characteristics, Reversal Strategies, and Outcomes From a Multicenter Observational Study. Chest. 2017;152(1):81–91.
50. Larsen TB, Skjøth F, Kjældgaard JN, Lip GYH, Nielsen PB, Søgaard M. Effectiveness and safety of rivaroxaban and warfarin in patients with unprovoked venous thromboembolism: a propensity-matched nationwide cohort study. Lancet Haematol. 2017 May;4(5):e237–44.
51. Sundström A, Alfredsson L, Sjölin-Forsberg G, Gerdén B, Bergman U, Jokinen J. Association of suicide attempts with acne and treatment with isotretinoin: retrospective Swedish cohort study. BMJ. 2010 Nov 11;341:c5812.
52. Carrère A. Les enrichissements prévus pour l'enquête CARE-Ménages. Direction de la recherche, des études, de l'évaluation et des statistiques [Internet]. 2017 [cited 2018 Jun 20]. Available from: https://hal.archives-ouvertes.fr/hal-01617706
53. Bezin J, Girodet P-O, Rambelomanana S, Touya M, Ferreira P, Gilleron V, et al. Choice of ICD-10 codes for the identification of acute coronary syndrome in the French hospitalization database. Fundam Clin Pharmacol. 2015 Dec;29(6):586–91.
54. Sahli L, Lapeyre-Mestre M, Derumeaux H, Moulis G. Positive predictive values of selected hospital discharge diagnoses to identify infections responsible for hospitalization in the French national hospital database. Pharmacoepidemiol Drug Saf. 2016;25(7):785–9.
55. Prat M, Derumeaux H, Sailler L, Lapeyre-Mestre M, Moulis G. Positive predictive values of peripheral arterial and venous thrombosis codes in French hospital database. Fundam Clin Pharmacol. 2018 Feb;32(1):108–13.
56. Lawson DH, Sherman V, Hollowell J. The General Practice Research Database. Scientific and Ethical Advisory Group. QJM Mon J Assoc Physicians. 1998 Jun;91(6):445–52.
57. THIN Database [Internet]. [cited 2018 Jun 18]. Available from: http://www.ucl.ac.uk/pcph/research/thin-database/database
58. Perlbarg J, Allonier C, Boisnault P, Daniel F, Le Fur P, Szidon P, et al. Faisabilité et intérêt de l'appariement de données individuelles en médecine générale et de données de remboursement appliqué au diabète et à l'hypertension artérielle. Sante Publique Vandoeuvre--Nancy Fr. 2014 Jun;26(3):355–63.
59. Unger JM, Hershman DL, Till C, Tangen CM, Barlow WE, Ramsey SD, et al. Using Medicare Claims to Examine Long-term Prostate Cancer Risk of Finasteride in the Prostate Cancer Prevention Trial. J Natl Cancer Inst. 2018 Mar 9;
60. Thompson IM, Goodman PJ, Tangen CM, Lucia MS, Miller GJ, Ford LG, et al. The influence of finasteride on the development of prostate cancer. N Engl J Med. 2003 Jul 17;349(3):215–24.
61. Présentation de l'INDS | INDS Institut National des Données de Santé [Internet]. [cited 2018 Jun 20]. Available from: https://www.indsante.fr/fr/presentation-de-linds
62. Nomination de Dominique POLTON en qualité de co-pilote du Health Data Hub | INDS Institut National des Données de Santé [Internet]. [cited 2018 Jun 20]. Available from: https://www.indsante.fr/fr/actualite/nomination-de-dominique-polton-en-qualite-de-co-pilote-

du-health-data-hub
63. Agnès Buzyn lance la mission de préfiguration du « Health Data Hub » un laboratoire d'exploitation des données de santé [Internet]. Ministère des Solidarités et de la Santé. 2018 [cited 2018 Jun 20]. Available from: http://solidarites-sante.gouv.fr/actualites/presse/communiques-de-presse/article/agnes-buzyn-lance-la-mission-de-prefiguration-du-health-data-hub-un-laboratoire