



HAL
open science

Initializing a hospital-wide data quality program. The AP-HP experience.

Christel Daniel, Patricia Serre, Nina Orlova, Stéphane Bréant, Nicolas Paris,
Nicolas Griffon

► **To cite this version:**

Christel Daniel, Patricia Serre, Nina Orlova, Stéphane Bréant, Nicolas Paris, et al.. Initializing a hospital-wide data quality program. The AP-HP experience.. Computer Methods and Programs in Biomedicine, 2019, 181, pp.104804 -. 10.1016/j.cmpb.2018.10.016 . hal-03488635

HAL Id: hal-03488635

<https://hal.science/hal-03488635>

Submitted on 20 Jul 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution - NonCommercial 4.0 International License

Initializing a hospital-wide data quality program. The AP-HP experience.

Christel Daniel, MD, PhD^{1,2}; Patricia Serre¹; Nina Orlova¹; Stéphane Bréant¹; Nicolas Paris¹; Nicolas Griffon, MD, PhD^{1,2}

¹DSI WIND, AP-HP, Paris, France, ²INSERM, U1142, LIMICS, F-75006, Paris, France; Sorbonne Universités, Paris, France

Corresponding author: Christel Daniel, MD, PhD
christel.daniel@aphp.fr

ABSTRACT

Background and Objectives: Data Quality (DQ) programs are recognized as a critical aspect of new-generation research platforms using electronic health record (EHR) data for building Learning Healthcare Systems. The AP-HP Clinical Data Repository aggregates EHR data from 37 hospitals to enable large-scale research and secondary data analysis. This paper describes the DQ program currently in place at AP-HP and the lessons learned from two DQ campaigns initiated in 2017.

Materials and Methods: As part of the AP-HP DQ program, two domains - patient identification (PI) and healthcare services (HS) - were selected for conducting DQ campaigns consisting of 5 phases: defining the scope, measuring, analyzing, improving and controlling DQ. Semi-automated DQ profiling was conducted in two data sets – the PI data set containing 8.8M patients and the HS data set containing 13099 consultation agendas and 2122 care units. Seventeen DQ measures were defined and DQ issues were classified using a unified DQ reporting framework. For each domain, actions plans were defined for improving and monitoring prioritized DQ issues.

Results: Eleven identified DQ issues (8 for the PI data set and 3 for the HS data set) were categorized into completeness (n=6), conformance (n=3) and plausibility (n=2) DQ issues. DQ issues were caused by errors from data originators, ETL issues or limitations of the EHR data entry tool. The action plans included sixteen actions (9 for the PI domain and 7 for the HS domain). Though only partial implementation, the DQ campaigns already resulted in significant improvement of DQ measures.

Conclusion: DQ assessments of hospital information systems are largely unpublished. The preliminary results of two DQ campaigns conducted at AP-HP illustrate the benefit of the engagement into a DQ program. The adoption of a unified DQ reporting framework enables the communication of DQ findings in a well-defined manner with a shared vocabulary. Dedicated tooling is needed to automate and extend the scope of the generic DQ program. Specific DQ checks will be additionally defined on a per-study basis to evaluate whether EHR data fits for specific uses.

Keywords: Data accuracy; data quality; Electronic Health Records; Data Warehousing; Observational Studies as Topic

1 Introduction

A Learning Health System (LHS) is defined by the Institute of Medicine(1) as a system in which “science, informatics, incentives, and culture are aligned for continuous improvement and innovation, with best practices seamlessly embedded in the delivery process and new knowledge captured as an integral by-product of the delivery experience.” LHS aim to maximize the potential of large-scale, harmonized data from variable quickly-developing digital sources including Electronic Health Records (EHRs). With the rapid adoption of EHRs, clinical information collected in structured, semi-structured and free-text notes during routine care is emerging as a powerful tool to facilitate discoveries that can improve health. EHR data is increasingly used for clinical research, effectiveness studies, innovation in digital health, quality improvement, and clinical decision support. Comparative effectiveness studies, patient-centered outcomes research, and pragmatic trials using EHR data from one or more practice settings are becoming an important complement to prospective randomized trials for generating new insights and knowledge. But the promised benefits of LHS can only be achieved if the quality of the data in the data sources is sufficient to support various secondary uses. EHR data are generated by electronic transactions in operational systems that are not primarily intended for research and secondary analysis. EHR systems rarely include predefined, unambiguous data definitions and uniform data collection procedures. They are optimized for administrative or clinical activities for which data collection is much less standardized and controlled than in clinical research and therefore can vary across healthcare facilities, even among users of the same systems. Data validity is one of the critical problems in reusing such “real-world” datasets. A number of studies have shown that EHR data contain errors that can have negative impacts on findings generated from these data and affect research results (2–4). With the development of hospital clinical data warehouses as well as large scale – regional, national or international – EHR-based data networks dedicated to clinical research (5–9) or public health (10), a key challenge is to determine for any specific purpose if the quality of the data is sufficient and fit the intended use. There is a need to set up consistent Data Quality (DQ) programs in practice settings and to report DQ findings to help secondary data users understand the potential impact of DQ on reusing data and interpreting findings(11).

1.1 Healthcare data quality programs

In the healthcare domain, existing DQ frameworks usually address specific purposes and there are no generally accepted methods to best characterize the data and quantifying DQ (12). DQ is recognized as a multi-dimensional concept. Each dimension of DQ consists of a set of attributes characterizing specific DQ requirements. In a recent review about 39 studies of DQ in the public health domain, Chen et al stated that a broad number of different house-defined or standardized national or international frameworks were used (13). In the selected studies, a range of 1 to 30 data variables were selected for DQ analysis – usually including demographics such as age, gender, and birth date, and specific information such as diagnostic code or laboratory testing results – but the rationale for

selecting the set of DQ measures was not always given. They identified a total of 49 quality attributes and corresponding measures used in 39 studies to describe DQ. Completeness, conformance, plausibility and timeliness were the attributes measured most often. In order to ensure transparency and consistency in reporting DQ findings across sources, the Electronic Data Methods (EDM) Forum recommends using a unifying DQ reporting framework integrating harmonized existing published DQ terms and a set of 20 DQ reporting rules for studies that use observational clinical and administrative data for secondary data analysis (11,12). The EDM forum reporting framework identifies three main DQ categories: Completeness, conformance and plausibility. Completeness focuses on the frequencies of data attributes present in a data set without reference to data values (its structure or plausibility). Conformance focuses on the compliance of the representation of data against internal or external formatting, relational, or computational definitions. Plausibility focuses on the believability or truthfulness of data values. In addition, the EDM forum also recommends reporting both general and analysis-specific DQ features (11).

A data quality (DQ) program involves setting up organizations, processes, methods and tools to achieve and maintain data quality in line with healthcare issues. The DMAIC (Define Measure Analyze Improve Control) methodology is an iterative and continuously active approach, based on DQ campaigns, used for data quality management process. Each DQ campaign is a 5-step cycle consisting in: i) *defining* the scope, ii) *measuring* the quality of the data in the defined scope, iii) *analyzing* the results, iv) *improving* the quality of the data and v) *controlling* it over time.

Defining the scope of the DQ campaign consists in selecting the domain, the datasets, the variables of interest and the DQ measures to be provided accordingly to priorities depending on the intended use of the data. *Measuring DQ* consists in providing the selected DQ measures. Each DQ measure can be produced by different quantitative and/or qualitative methods; therefore, there is flexibility in methods used to measure DQ. DQ checks consist of computed DQ measures sometimes combined by qualitative observations using checklist and rating scales or informal observations (13). Qualitative methods based on interviews with key informants may be used to understand the context of quantitative DQ findings. Structured interviews or qualitative in-depth interviews provide end users' view on DQ (14,15). Statistical data analysis methods are determined by the purpose of the study and the types of data collected. *Analyzing* the DQ findings consists in characterizing DQ issues (DQI) and identifying their cause. *Improving DQ* requires prioritizing DQI, defining explicit objectives in terms of expected DQ level and setting up an action plan addressing the causes of the selected DQI. At the end of the action plan, *controlling DQ* consists in monitoring the outcome in order to prevent future issues and ensure that optimum DQ is maintained over time.

In the healthcare domain, few care practices or distributed data networks report on DQ programs ensuring the continuous monitoring of the quality of their datasets. The Canadian Institute for Health Information (CIHI) DQ process consisting of a manual process based on questionnaire to data stewards does not result in measures of DQ that are easily comparable across different data sets

(https://www.cihi.ca/en/data_quality_framework_2009_en.pdf). The Observational Health Data Sciences and Informatics (OHDSI) (<https://www.ohdsi.org/>) program is a multi-stakeholder, interdisciplinary collaborative partnership developing best practices for using observational health data through large-scale analytics (10). OHDSI has defined a common data model – the OMOP model – and a DQ process to ensure that all reported data meet certain DQ standards. A series of approximately 35 DQ rules are automatically applied to any OHDSI datasets. The OHDSI DQ rules are not easily transferable for assessing the quality of other data sets that do not conform to the OMOP data model. Similarly, MiniSentinel project conducted by the Food and Drug Administration to monitor the safety of medical products regulated by the agency has defined a checklist of over 2,000 DQ rules that must be satisfied for data to be acceptable (16). Besides these initiatives of continuous monitoring of DQ, most published DQ studies report retrospective, cross-sectional DQ assessments using variable sampling methods and data collection period considered.

1.2 AP-HP Electronic Healthcare Record (EHR) and Clinical Data Repository (CDR)

The Assistance Publique – Hôpitaux de Paris (AP-HP) (<https://www.aphp.fr>) is the largest hospital entity in Europe with 39 hospitals (22,474 beds) mainly located in the Greater Paris region. 92,000 dedicated professionals including 22,000 physicians offer advanced treatments to patients with 1.5M hospitalizations per year (10% of all hospitalizations in France). AP-HP is also the first biomedical research center in Europe with > 500 clinical trials and > 8,000 publications per year (25% of all medical research publications in France). Integrating health records across care delivery sites is critical to develop a more comprehensive and accurate picture of health and healthcare delivery for the individual, and in aggregate may provide clearer insight into the health of particular populations.

Since 2010, AP-HP is deploying a common Electronic Healthcare Record (EHR), based on the Orbis solution (AGFA Healthcare), for sharing information across the AP-HP hospitals and communicating with the hospitals and physicians outside AP-HP. To date, 37 AP-HP hospitals are using Orbis. Since 2017, in order to improve patient experience, AP-HP has been deploying a common online medical booking service across all AP-HP hospitals (<https://www.aphp.fr/rendez-vous>). As part of the Orbis project, AP-HP has set up an Enterprise Master Patient Index service to assign a single common identifier for a patient across hospitals using a linkage algorithm that minimize the need for human review (17,18). The linkage process will end up in 2018 with the adoption of Orbis by the 2 last AP-HP hospitals. In addition, ongoing efforts focus on defining common vocabularies and semantic interoperability solutions for health information exchange within AP-HP as well as within regional/national/international health information networks (regional TerriS@nté project and national personal health record).

In parallel, since 2014, AP-HP is building a LHS in order to answer medical questions more efficiently and to constantly extract new knowledge from the data and propagate it for quality

improvement. The analytics platform is based on a clinical data repository (CDR), aggregating day-to-day clinical data from 8,8M patients captured by clinical databases. The AP-HP analytics platform consists of a PostgreSQL database, a Big Data computing cluster composed of both hadoop distributed CPU and nvidia GPU technologies. The ecosystem allows access to dedicated data marts and to a pool of analytics tools in order to explore and analyze AP-HP datasets. Dedicated applications are deployed to specifically support the different uses of the data. For research or innovation purposes, an open-source cohort builder - the Integrating Biology and the Bedside (i2b2) software – deployed since 2016 is used to identify patient population of interest for conducting studies. Since 2017, Jupyter is used in addition as a portal to access massive clinical data that are needed for the research studies. The AP-HP CDR and I2b2/jupyter solutions currently support both observational studies (21 on-going studies) and prospective randomized controlled trials (feasibility studies, patient recruitment, data collection). The data analytics platform is also used to develop and/or evaluate algorithms based on new Artificial Intelligence techniques (deep learning). Pilots are being built on medical images and real-time data generated by intensive care monitors. Multilingual semantic interoperability services have been developed (e.g. OMOP standardization <https://www.youtube.com/watch?v=eJpnrPPErq0>) and pilot integration with international research platforms are under evaluation (EHR2EDC project (EITHealth)). For hospital performance monitoring, the IBM Cognos solution is under deployment. Since 2017, different dashboards have been developed and used by clinical units for monitoring their activity. Figure 1 depicts the timeline of the deployment of both the AP-HP EHR (for patient care) and AP-HP CDR (for research and business intelligence).

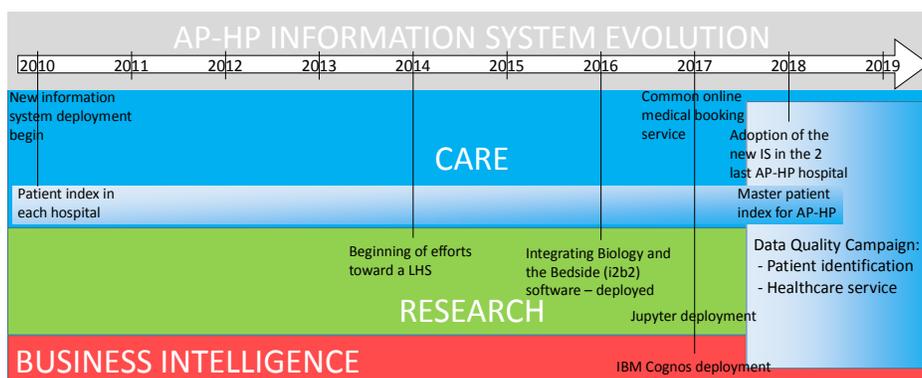


Figure 1 : Timeline of the deployment of the common Electronic Health Record (AP-HP EHR) (for patient care) and of the Clinical Data Repository (i2b2, Jupyter, IBM Cognos)(for research and business intelligence).

In this context, patient identification and healthcare services catalogue were identified as two high-priority domains of the DQ program. Patient identification is a key concern in our process of linking EHR data from a large network of hospitals having overlap in patient populations. In this context, being unable to link records for individuals who visits multiple AP-HP hospitals may lead to double-counting of individuals when aggregating health events across AP-HP hospitals. The healthcare

services catalogue describes the list of consultations and hospital services across the 39 hospitals of the AP-HP network. The quality of this catalogue is important in order to characterize the organization and performance of healthcare delivery in the institution.

The objective of this paper is to describe the DQ program currently in place at AP-HP and to report the preliminary results of two DQ campaigns initiated in 2017 in the domains of Patient Identification (PI) and Healthcare Services (HS).

2 Material and methods

2.1 Patient identification

Patient identification (PI) is the first act of any administrative or medical activity at registration time, during any visit or stay and at the patient's exit. Concerning all hospital stakeholders - administrative or healthcare professionals - and requiring patient involvement, identity vigilance is a key factor of patient safety and hospital management. A correct identification of the patient guarantees that the right care is delivered to the right patient and the corresponding invoice sent to the right payer.

The Identity Vigilance committee defined the scope of the DQ campaign for the Patient Identification (PI) domain and decided to focus on 10 identification features to be evaluated on the data set of 8,8M patients' identities aggregated across 37 AP-HP hospitals as part of the ORBIS project (table 1). Among the variables of interest, five identification features – birth last name, usual last name, first name, gender, birth date – were selected because they are used by the linkage algorithm to identify duplicates i.e. match identities that are likely to correspond to the same patient. The other identification features were selected because of their interest for epidemiological studies (birth zip code, birth country code and residence country code) or for the capacity of AP-HP to conduct massive information campaigns towards the patients for engaging them in both their personal care or research activities (address, email).

Thirteen DQ measures grouped into three DQ categories: completeness (n=9), conformance (n=2) and plausibility (n=2) were considered (table 1). The completeness DQ measures evaluate the proportion of valued identification features without any reference to the structure or plausibility of the values. The conformance DQ measures allow checking that the values of the identification features meet the expected data representation and value set constraints described in the data dictionary. Value sets constraints are either locally defined or imposed by external terminology standards (e.g. HL7 FHIR value set for administrative gender, ISO 3166 value set for country codes). Two DQ measures of plausibility were defined to evaluate the believability or truthfulness of PI data. The first one is the identity confirmation rate. It measures the percentage of patient identities that were controlled using a trustworthy document. The AP-HP has two identity statuses – provisional or confirmed – related to the degree of confidence of the patient identification features. In the current identity management system, at registration time, an identity is initialized with the provisional status and is confirmed as soon as the strict identification features – birth last name, usual last name, first name, date of birth and

administrative gender – have been verified by the registration agent based on a reference identification document. The second plausibility DQ measure is the duplicate rate, which is the proportion of patients having multiple identities (and multiple medical records), as measured by an algorithm based on the five above mentioned strict identification features. This DQ measure, related to uniqueness plausibility, allows ensuring that data values that identify a single object are not duplicated. DQI were classified using the EDM forum unifying DQ reporting framework and an action plan was defined for improving and monitoring prioritized DQI.

2.2 Healthcare services

The catalog of AP-HP healthcare services (HS) is maintained in order to provide a clear description of all institutional consultation and hospitalization services. Each consultation or care unit is associated to a medical or paramedical specialty. In the context of the deployment of the online consultation booking service, the quality of the AP-HP HS catalog is important for both patients and physicians: the former needs must match the latter competencies. The deployment of the online consultation booking service generated dissatisfaction expressed by several patients and healthcare professionals, some of the latter decided to give up the service.

The Healthcare Services (HS) catalog committee defined the scope of the DQ campaign for the domain focusing on two variables - care unit specialty and consultation specialty - to be evaluated respectively on 2122 active AP-HP care units and a subset of 825 active consultation agendas used in the Necker hospital. Four DQ measures grouped into two DQ categories: completeness (n=2) and conformance (n=2) were considered (table 3). The evaluation of the conformance of HS specialty is based on the compliance against a locally defined AP-HP specialty catalog. DQI were analyzed and an action plan was defined for improving and monitoring them.

3 Results

This section includes, for both PI and HS domains, the results of three steps of the DQ campaign cycle: *measuring*, *analyzing* and *improving* the quality of the data. For each domain, the DQ measures are provided, the DQI identified and the action plan defined for acting on the cause of the DQI and therefore improving the DQ measures is described.

3.1 Patient identification

Table 1 describes the list of DQ measures evaluated during the DQ campaign in the Patient Identification (PI) domain. A completeness rate of 100% is observed for four identification features - usual last name, first name, gender and birth date - for both confirmed and provisional identities. We identified eight DQI regarding completeness (n=5), conformance (n=1) and plausibility (n=2). The completeness rate of the birth last name is only of 55.7% for provisional identities. This identification feature was not mandatory in the previous identity management systems used in the AP-HP hospital before the deployment of the Orbis solution and this information needs to be completed when the patients come back to any AP-HP hospital. The completeness rate for other identification features –

birth zip, birth country code, address and email – is not satisfactory, especially in the case of provisional identities. The completeness rate of email is less than 8%. The conformance rate evaluated on gender and residence country code is respectively of 100% and 22.8%. The low conformance rate of country code was due to a programming error of the ETL process. Only 18% of identities are confirmed with a great disparity between AP-HP hospitals. This low rate of identity confirmation, observed since the linkage process, is related to the fact that the identification system used currently at AP-HP considers an additional strict identification feature - birth last name – which was not systematically collected with the previous identification system (considering only usual last name, first name, date of birth and administrative gender as strict identification features). Therefore, all identities processed during the linkage period have a provisional status until they are match to a confirmed identity or confirmed at registration time when the patient come back to any AP-HP hospital. Systematic identity confirmation by registration agents is a big challenge to be addressed to achieve an identity confirmation rate >50% defined as initial short term target.

As of March 2018, the duplicate rate in the database is only of 2.5% of the total number of identities in the Orbis database. To date, duplicates mainly result from the still ongoing linkage process across the AP-HP hospitals and to a lesser extent to errors during data entry at registration time (unsuccessful search for already existing identity and erroneous data entry for newcomers). This DQI need to be fixed by better quality of data entry at registration time and more efficient merge of duplicates by data stewards of the identity vigilance units.

Table 1: DQ campaign in the Patient Identification (PI) domain – DQ measuring phase: DQ measures, DQI and targeted values. (: identification features used by the linkage algorithm to identify duplicates). Conf. Id: confirmed identity; Prov. Id: provisional identity.*

DQ measure			Value (%)			Objective	DQI / Target (%)
			(March 2018)				
Completeness							
			All Id (%)	Conf. Id (%)	Prov. Id (%)		
Birth last name*			63.6	100	55.7	DQI 1/80	
Usual last name*				100		Features used for duplicate detection shall not be null	
First name*				100			
Gender*				100			
Birth date*	Completeness rate			100			
Birth zip code			32.3	52.3	27.9	DQI 2/100	
Birth country code			71.2	82.5	68.7	DQI 3/100	
Address			93.7	93.7	94.3	Features used to contact patient shall not be null	DQI 4/100
Email			8.0	21.1	4.8		DQI 5/40
Conformance							
Residence code	country	Rate of compliance with the data dictionary		22.8		Country codes shall conform to ISO 3166 country codes	DQI 6/100
Gender*				100		Gender values shall conform to the code system	

			http://hl7.org/fhir/administrative-gender (male, female, other, unknown)	
Plausibility				
Patient uniqueness	Duplicate rate	2.5%	Minimizing the duplicate rate.	DQI 7/ < 1
Identity confirmation	Identity confirmation rate	18.0%	Maximizing the identity confirmation rate	DQI 8/ > 50
DQI: Data Quality Issue				

In summary, DQI are caused by unintended errors from data originators – registration agents or healthcare professionals, from data stewards in identity vigilance units and/or from the patient. Staff-related errors are mainly related to the absence or defects of verification due to the lack of identity document or unreliable identity documents. The patient's role in the identification process should not be neglected, (change of usual names e.g. after marriage, inability to communicate, non-disclosure of identity, identity theft). DQI may also be caused by limitations of the EHR data entry tool (inadequacy of value sets constraints, lack of DQ checks) or be false positives due to programming errors of the ETL process.

Based on the analysis of the type and causes of DQI, the identity vigilance committee defined a set of actions for improving the DQ measures and reaching explicit objectives in terms of expected level of quality.

Table 2 summarizes the action plan established to ensure a better completeness of the patient identification features and more systematic verification of identities with trustworthy documents. The actions are grouped in four axes: communication and teaching, engagement of registration agents or health professionals, patient engagement and information system. Improving patient identification features and the rate of confirmed identities is likely to improve the result of the linkage algorithm and reduce the workload of the manual check of duplicates distinguishing true and false matches.

Table 2: DQ campaign in the Patient Identification (PI) domain – DQ improvement phase: action plan

Axes and actions	DQI impacted	Involved actors				
		Patients	Data originator	Data steward	Data consumer	Results consumer
Axis 1: Communication and teaching						
Leaflets	DQI 1-8	x	x	x	x	x
Completed Procedures Guide	DQI 1-8		x	x		
Educational videos on identity vigilance issues	DQI 1-8		x	x		
Feedback process	DQI 1-8		x	x		
Axis 2: Engagement of registration agents or health professionals						

Patient identity confirmation campaign	DQI 7	x	
Duplicate merging campaign	DQI 8	x	
Axis 3: Patient engagement			
Involvement of the patient (re-reading by the patient of the seized elements and labels)	DQI 1-6	x	
Axis 4: Information system			
Basic equipment (computers, tablets, etc.) and use of new technologies (barcode scanner, etc.)	DQI 1-8	x	x
Data quality checks (completeness, conformance)	DQI 6	x	
DQI: Data Quality Issue (see table 1)			

Some of the actions of the action plan have been engaged. For example, the programming error of the ETL regarding country codes has been fixed. In order to monitor DQ improvement within the current DQ quality campaign, a dashboard has been developed using the IBM Cognos solution. This dashboard allows the data stewards in the identity vigilance units to follow the progress of the action plan at both the global AP-HP and hospital level.

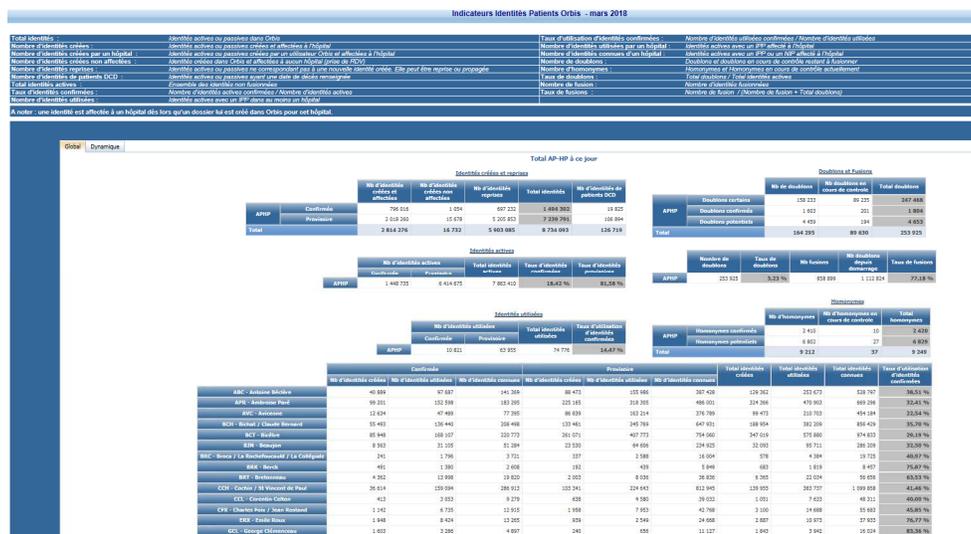


Figure 1: DQ control - Patient Identification (PI) dashboard for data stewards in the identity vigilance units.

3.2 Healthcare services

Table 3 describes the list of DQ measures evaluated during the DQ campaign in the Healthcare Services (HS) domain classified in the completeness and conformance categories. A completeness rate of 100% is observed for specialty in consultation agendas. We identified three DQI regarding completeness (n=1) and conformance (n=2). The completeness rate of specialty in care unit is of 97.4%. The conformance rate of specialties for the AP-HP care units and consultation agendas are respectively of 46.9% and 85.7%. The low score of conformance rate of specialty in the AP-HP care units is related to the use of a financially-oriented specialty catalog defined to optimize rather the billing than care process. The financially-oriented specialty catalog needs to be aligned to the

medically-oriented specialty catalog and the latter needs to be updated to provide missing codes. The cause of suboptimal score of conformance rate of specialty in the consultation agendas has been analyzed in the hospital with the lowest score. This suboptimal score was due to human errors (typos (18%), formalism errors (41%)) and inadequacy of the specialty catalog (41%).

Table 3: DQ campaign in the Healthcare Services (HS) domain – DQ measuring phase: DQ measures, DQI and targeted values

Check data	DQ measure	Value (%) (March 2018)	Objective	DQI / Target (%)
Completeness				
Consultation specialty	Completeness rate	100	Healthcare services shall be associated to a specialty	DQI 9/100
Care unit specialty		97.4		
Conformance				
Consultation specialty	Conformance rate	85.7	Specialties associated to healthcare services shall be compliant with the AP-HP specialties catalogue	DQI 10/100
Care unit specialty		46.9		DQI 11/100

In order to correct the identified DQI, the Healthcare Services (HS) committee defined a set of actions grouped in three axes: communication and teaching, engagement of the staff and information technology support (table 4). Some of the actions have been already engaged. In order to address the DQI 10 (non-conformance of specialty consultation agenda), new specialties were added to the AP-HP catalog to better describe the specialty of the healthcare professionals.. In order to address the DQI 11 (non-conformance of specialty in the AP-HP care units), the mapping from the financially-oriented to the medically-oriented specialty catalogs of respectively 342 and 169 terms has been done. Thirty-one specialties of the financially-oriented specialty catalog will be introduced in the medically-oriented catalog and 110 specialties are under consideration. One difficulty in managing the specialty catalog, as for any terminology, is to find the right level of granularity allowing both a correct and usable description of the domain. A new version of the data dictionary and procedure guide will be available soon.

Table 4: DQ campaign in the Healthcare Services (HS) domain – DQ improvement phase: action plan

Axes and actions	DQI impacted	Involved actors				
		Data originator	Data steward	Data consumer (Healthcare Professionals)	Data consumer (Patient)	Results consumer
Axis 1 : Communication and teaching						
Update of the Procedures Guide	DQI 9-11		x			x
Communication/dissemination of the procedure guide	DQI 9-11	x				
Feedback process	DQI 10-11	x		x	x	x
Axis 2 : Engagement of the staff (vocabulary team, IT and healthcare professionals)						

Update of the specialty catalog	DQI 10-11	x	x			x
Update of the consultation agendas	DQI 10	x				
Update of the care unit catalog	DQI 11	x				
Axis 3 : Information system						
Data quality checks (completeness, conformance)	DQI 9-11	x				
DQI: Data Quality Issue (see table 3)						

4 Discussion

Reflecting an important part of healthcare practice, EHR data are essential to research and public health. Improving EHR data quality toward longitudinal health records is a key challenge for patient care as well as for clinical effectiveness research, clinical research, phenotyping, drug safety, population health, pharmaceutical surveillance, etc. Personal health records, health information exchange as well as secondary use of EHR data for research or public health are key enabling technologies for improving EHR data quality.

Like more and more institutions maturing in clinical data warehousing, AP-HP has set up a Clinical Data Repository aggregating data from 8,8M patients collected during the past 10 years in 37 hospitals of the greater Paris area. In the context of this project, we emphasized the importance of systematic data quality assessment considering that data quality is a component of the quality and effectiveness of the EHR. Data quality is influenced by technical, organizational, behavioral and environmental factors and covers large information systems contexts, specific knowledge and multi-disciplinary techniques (19).

As first step of the AP-HP DQ program, DQ campaigns were initiated by the DQ manager focusing on the patient identification (PI) and healthcare services (HS) domains.

Eleven DQI have been identified (8 for the PI data set and 3 for the HS data set) and categorized into completeness (n=6), conformance (n=3) and plausibility (n=2) DQI. The adoption of a unifying DQ reporting framework enabled the communication of DQ findings in a well-defined manner with a shared vocabulary across domains. DQI were caused by errors from data originators, ETL issues or limitations of the EHR data entry tool (inadequate value set constraints, lack of DQ checks). Although related to a very limited scope, these results are consistent with prior studies describing major causes of data quality issues (20).

Action plans defined in both PI and HS domains included respectively nine and seven actions. Though only partially implemented, the DQ campaigns already resulted in significant improvement of DQ measures. For example, in the PI domain, the implementation of the first actions of the plan resulted in an increase of 11,7% of the identification confirmation rate (from 18% in May to 20.12% in July). In the HS domain, in order to address the DQI 10 (non-conformance of specialty consultation agenda) in the hospital with the lowest score, seven new specialties were added to the AP-HP catalog. The update of the value set constraints – based on the April version of the specialty catalog - resulted in increase of 30.8% of the conformance rate of the specialty consultation agenda (from 71.6% to 93.7 %). These

preliminary results illustrate the benefit of the engagement into a DQ program. The major achievement was the development of manual and automated DQ checks and the increased awareness about DQI within the AP-HP direction and across AP-HP hospitals. The DQ campaigns contributed to position data quality at the forefront of data governance.

4.1 Limitations

This study reports preliminary results of DQ assessment conducted on a very limited scope related to only two non-clinical domains. The future DQ campaigns will have to address clinical data including diagnoses, conditions, procedures, observations, etc.

We used only quantitative DQ attributes and simple analysis methods. Data quality assessment should use mixed methods (e.g., qualitative and quantitative assessment methods) to assess data from multiple sources (e.g., records, organizational documentation, data collection process and data users) and used at different levels of the organization (13). Even though, we considered data generators' and data stewards' perspective (collecting their feedback on the DQ campaign), our analysis of DQI did not extensively consider all DQ documentation and assessment performed at each step along the chain of data stewardship. Moreover, our DQ tooling is still at its infancy and most of the DQ checks require manual process. Dedicated tooling is needed to automate and extend the scope of the generic DQ program. Evaluating the quality of data for secondary use is challenging. Although DQ should be defined from an end user's perspective, it is often implemented from the perspective of the data producer. Determining if a data set is, in fact, fit for its intended use, is highly context specific. For example, a generic DQ summary is unlikely to provide a comprehensive analysis of the data quality for a specific study which is necessarily dependent on the analysis plan for the study. We started to take into account the feedback from data consumers including research study investigators in order to define a method for developing specific DQ checks defined on a per-study basis.

4.2 Future work

Future DQ campaigns will cover an extended scope of clinical variables with the challenging objective of validating automated phenotype identification algorithms. We also plan to include assessment of unstructured clinical data and validation against external established gold standard datasets.

Dedicated tooling still needs to be developed to automatically generate DQ annotations. AP-HP is currently building a data mart based on the OMOP CDM as an internal data model. We will implement the data quality checks that are designed accordingly and intend to leverage the lessons learned by other users of OMOP-based CDMs. It is important to deploy the DQ monitoring tool and to engage data generators and data stewards in regular check of the DQ dashboard. Brown et al. noted that a previously validated dataset does not necessarily guarantee that there will be no new DQI over time (21).

“Fitness for use” is a key concept in determining if a data set can be used in a specific task. While generic DQ checks are important in an initial assessment of fitness for use for a specific analysis or

study, these checks may not be sufficient to understand the fitness for a highly specialized use. We will develop detailed DQ assessment metrics to be incorporated into analytic programs using study-specific data sets.

Acknowledgments

The authors wish to gratefully acknowledge the help of P.Vallet, F.Jauffret, Francine Ahouangonou, Elodie Ricaud and M.Guerrier

Author Contributions

CD conceptualized the study and developed the conceptual framework with the support of PS. PS carried out the design of the study with the guidance of CD and NG. NO and PS developed the DQ measures for the PI domain and CD and NG for the HS domain. NO developed the dashboard for the PI domain. NP and SB developed and deployed the AP-HP Clinical Data Repository. PS and NG collected the data, performed the data analysis. CD, PS and NG discussed the study and all participated in the synthesis processes. CD drafted the first manuscript. All authors made intellectual input through critical revision to the manuscript. All authors read and approved the final manuscript.

Conflicts of Interest

The authors declare no conflict of interest.

Reference

1. Friedman CP, Wong AK, Blumenthal D. Achieving a nationwide learning health system. *Sci Transl Med.* 10 nov 2010;2(57):57cm29.
2. Cheng P, Gilchrist A, Robinson KM, Paul L. The risk and consequences of clinical miscoding due to inadequate medical documentation: a case study of the impact on health services funding. *Health Inf Manag J Health Inf Manag Assoc Aust.* 2009;38(1):35-46.
3. Chen H, Yu P, Wang N. Do we have the reliable data? An exploration of data quality for AIDS information system in China. *Stud Health Technol Inform.* 2013;192:1042.
4. Wilkinson DL, McCarthy M. Use of comparative data for integrated cancer services. *BMC Health Serv Res.* 17 déc 2007;7:204.
5. McMurry AJ, Murphy SN, MacFadden D, Weber G, Simons WW, Orechia J, et al. SHRINE: enabling nationally scalable multi-site disease studies. *PloS One.* 2013;8(3):e55811.
6. De Moor G, Sundgren M, Kalra D, Schmidt A, Dugas M, Claerhout B, et al. Using electronic health records for clinical research: the case of the EHR4CR project. *J Biomed Inform.* févr 2015;53:162-73.
7. Coorevits P, Sundgren M, Klein GO, Bahr A, Claerhout B, Daniel C, et al. Electronic health records: new opportunities for clinical research. *J Intern Med.* déc 2013;274(6):547-60.

8. Delaney BC, Curcin V, Andreasson A, Arvanitis TN, Bastiaens H, Corrigan D, et al. Translational Medicine and Patient Safety in Europe: TRANSFoRm--Architecture for the Learning Health System in Europe. *BioMed Res Int*. 2015;2015:961526.
9. Collins FS, Hudson KL, Briggs JP, Lauer MS. PCORnet: turning a dream into reality. *J Am Med Inform Assoc JAMIA*. août 2014;21(4):576-7.
10. 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. *Stud Health Technol Inform*. 2015;216:574-8.
11. Kahn MG, Callahan TJ, Barnard J, Bauck AE, Brown J, Davidson BN, et al. A Harmonized Data Quality Assessment Terminology and Framework for the Secondary Use of Electronic Health Record Data. *EGEMS Wash DC*. 2016;4(1):1244.
12. Kahn MG, Raebel MA, Glanz JM, Riedlinger K, Steiner JF. A pragmatic framework for single-site and multisite data quality assessment in electronic health record-based clinical research. *Med Care*. juill 2012;50 Suppl:S21-29.
13. Chen H, Hailey D, Wang N, Yu P. A review of data quality assessment methods for public health information systems. *Int J Environ Res Public Health*. 14 mai 2014;11(5):5170-207.
14. Chiba Y, Oguttu MA, Nakayama T. Quantitative and qualitative verification of data quality in the childbirth registers of two rural district hospitals in Western Kenya. *Midwifery*. juin 2012;28(3):329-39.
15. Hahn D, Wanjala P, Marx M. Where is information quality lost at clinical level? A mixed-method study on information systems and data quality in three urban Kenyan ANC clinics. *Glob Health Action*. 29 août 2013;6:21424.
16. Andrade SE, Toh S, Houstoun M, Mott K, Pitts M, Kieswetter C, et al. Surveillance of Medication Use During Pregnancy in the Mini-Sentinel Program. *Matern Child Health J*. avr 2016;20(4):895-903.
17. Weiner M, Stump TE, Callahan CM, Lewis JN, McDonald CJ. A practical method of linking data from Medicare claims and a comprehensive electronic medical records system. *Int J Med Inf*. août 2003;71(1):57-69.
18. Grannis SJ, Overhage JM, Hui S, McDonald CJ. Analysis of a probabilistic record linkage technique without human review. *AMIA Annu Symp Proc AMIA Symp*. 2003;259-63.
19. Aqil A, Lippeveld T, Hozumi D. PRISM framework: a paradigm shift for designing, strengthening and evaluating routine health information systems. *Health Policy Plan*. mai 2009;24(3):217-28.
20. Botsis T, Hartvigsen G, Chen F, Weng C. Secondary Use of EHR: Data Quality Issues and Informatics Opportunities. *AMIA Jt Summits Transl Sci Proc AMIA Jt Summits Transl Sci*. 1 mars 2010;2010:1-5.
21. Brown JS, Kahn M, Toh S. Data quality assessment for comparative effectiveness research in distributed data networks. *Med Care*. août 2013;51(8 Suppl 3):S22-29.