Chapter

Analysis and Curation of the Database of a Colo-Rectal Cancer Screening Program

Rocio Aznar-Gimeno, Patricia Carrera-Lasfuentes, Vega Rodrigalvarez-Chamarro, Rafael del-Hoyo-Alonso, Angel Lanas and Manuel Doblare

Abstract

Data collection in health programs databases is prone to errors that might hinder its use to identify risk indicators and to support optimal decision making in health services. This is the case, in colo-rectal cancer (CRC) screening programs, when trying to optimize the cut-off point to select the patients who will undergo a colonoscopy, especially when having insufficient offer of colonoscopies or temporary excessive demand. It is necessary therefore to establish "good practice" guidelines for data collection, management and analysis. With the aim of improving the redesign of a regional CRC screening program platform, we performed an exhaustive analysis of the data collected, proposing a set of recommendations for its correct maintenance. We also carried out the curation of the available data in order to finally have a clean source of information that would allow proper future analyses. We present here the result of such study, showing the importance of the design of the database and of the user interface to avoid redundancies keeping consistency and checking known correlations, with the final aim of providing quality data that permit to take correct decisions.

Keywords: colo-rectal cancer screening program, health data, data analysis and curation, data coherence, data integrity

1. Introduction

Big Data and Artificial Intelligence are revolutionizing medicine, although they require large amounts of data [1]. Healthcare is an information-intensive activity that produces large quantities of structured (laboratory data) and non-structured (images, texts, etc.) data, from laboratories, wards, operating theaters, primary care organizations. Also, the amount of these data will surely highly increase in the near future due to the interconnection of medical devices via the Internet of Things [2].

Electronic Health Record Databases (EHRD) quality and interoperability [3] is one hot topic in Health Data Science. However, the data captured in EHRD is not available just from well-designed and maintained databases controlled by an administrator and curated by an IT Department, but, contrarily, it is composed of non-unified, redundant, and often replicated information that come from numerous independent e-health service providers (hospital, primary services, regional governments). Therefore, to assure the quality of the data, new processes are required from origin to final service generation through an appropriate data governance.

We can expect that new technologies such as blockchain will reduce this problem, introducing data interoperability, and security [4] and by the adoption of international standards for EHRD (Reference Model, ISO/DIS 13606–2, OpenEHR) [5–8]. However, we are just in the time when leveraging the power of health data to improve clinical or administrative decisions still requires an important effort to ensure the requested data quality. In this regard, many research studies discuss different approaches to improve quality and curation [9] and to deploy new advanced services [2].

There have been several works addressing this problem of data analysis and curation of health information systems [10–13]. This chapter focuses on data quality analysis of electronic medical records and, in particular, of the database of the colorectal cancer screening programme of the Spanish region of Aragón.

The colo-rectal cancer (CRC) screening program of Aragón started in 2014 and, as many other similar programs in the world, is based on the result of a fecal immunohisto-chemical test (FIT) and is focused on medium risk population (ages between 50 and 69 years, without family history of CRC and without the presence of colon diseases, colectomy, or irreversible terminal diseases such as Alzheimer's). The result of this test determines whether it is necessary to perform a colonoscopy (positive cases) or the patient will be screened again after a predefined period. The objective of the program is to diagnose the colo-rectal cancer in its early stage and/or to remove precancer polyps before they may evolve to potential malignant tumors.

One of the difficulties/limitations that this type of programs usually encounters is the insufficient offer of colonoscopies, or the excessive demand derived from positive FIT cases in the invited population. It is necessary therefore to analyze the historical information to define a set of data-based risk indicators than can support the decision-making process in public health services, trying to set the least harmful criteria for selecting the patients who will finally undergo the colonoscopy. It is then clear the importance of the quality of the information stored concerning the clinical data of the patients participating in the program as well as the information of the tests carried out, the results of the colonoscopies and the associated pathological data.

Data collection, if performed by humans, is prone to filling errors. These potential errors can be reduced by a proper design of the database and of the user interface, avoiding redundancies, keeping consistency and checking known correlations. Therefore, it is necessary to establish "good practice" guidelines for data collection and management. With the aim of improving the redesign of the current platform, we carried out an exhaustive analysis of the data collected in the Aragón's regional colo-rectal screening program from 2014 to 2018. This analysis revealed considerable data noise, so we proposed a list of recommendations to improve their quality. We also carried out a curation process of the available data in order to have a clean source of information that would allow proper future analyses.

The recommendations arose from the identification of a series of assumptions and restrictions that the platform should contain to comply with the integrity, coherence and consistency of the data and, therefore, to mitigate the noise. They covered from the default value of each variable, its range, its mandatory character and the redundancy control, to other types of suggestions that include possible constraints on their values, relationships between variables, creation of new variables that may facilitate the analysis and possible warnings or alerts that could help the user to perform a correct data filling.

The chapter is organized as follows. In the first section, the database analyzed is described. Section 2 introduces some basic principles with respect to data integrity, consistency and coherence that any data manager must adhere to in order to ensure the data quality and presents some examples of the data analysis undertaken and a number of recommendations with the aim of complying with these principles. Decisions taken retrospectively are then introduced into the data healing process in order to obtain a clean source of information from which to draw knowledge for further analyses. Finally, the last section includes the conclusions of the entire study.

2. Description of the database

All the information of the CRC screening program in the region of Aragón (Spain) is stored in a centralized database that is fed from other external databases and from the personal information of the patient that is filled by hospital staff through a user-interface (UI) tool. This UI is a web application on which the patient information is displayed and managed. The information that contains comes from the different public hospitals in Aragón: San Jorge Hospital, Barbastro Hospital, Miguel Servet University Hospital, Lozano Blesa University Clinical Hospital, Ernest Lluch Hospital, Obispo Polanco Hospital of Teruel and Alcaniz Hospital. Therefore, the staff who use the platform have different roles, belong to different hospitals and have different degree of training. This means that the application must be as intuitive as possible, as well as to comply with sufficient checks to handle the data relationships correctly, reducing possible errors as much as possible during the data collection. This translates the problem to the good design of the database.

Furthermore, it is quite common in public institutions to have several contracts with different private companies in a short period of time. This usually implies waste of time in understanding, adapting and changing the structure to the way of working of each of company. Therefore, it makes even more sense to establish a good database design and architecture that allows its correct growth and maintenance.

In particular, this section explains the characteristics of the database of the colo-rectal cancer screening program of Aragón between 2014, when the program started, and 2018. In the following sections, the inconsistences found in its design and, therefore, in the data quality are exposed and a series of recommendations (and “good practices”) are proposed to comply with data integrity, coherence and consistency as much as possible.

The existing database model is here explained in inverse order to the actual development of the database. First the final result is explained (relational model) and then the underlying model (entity-relationship model) is discussed [14].

2.1 Relational model

The database tables analyzed contain the following data information which is extracted from different sources of information:

- Patient: Basic demographic information on target patients. This information is extracted from the User Database (BDU) of the corresponding health area.

- Exclusion: Information on temporary or permanent exclusions to the program, which are similar to exclusion criteria of other CRC screening programs in Spain. Exclusions may be due to family history of

CRC, presence of colonic disease, colectomy, irreversible disease (e.g., Alzheimer), previous negative FIT result, or previous negative colonoscopy outcome. This information is automatically dumped into the table from different health system databases, such as OMI-AP (clinical information of patients attended in primary care), CMBDH (clinical information of patients attended in hospital), HP-His (clinical information of ambulatory patients) and BDU (User Database).

- Correspondence: Information from the letters sent to patients along their stay in the program. The process of sending letters is carried out manually by administrative staff through the platform, according to the hospital's criteria, using the target population (60–69 years), excluding those in the exclusion table. Administrative staff is in charge of setting dates, choosing the number of patients to send the letter and gathering positive results. This process is time costly and prone to errors, requiring additional validation.

- Test: Information about the tests performed on patients throughout the program. In particular, the tests carried out are the following:

    - Fecal immunochemical test (FIT): The result of this test comes from several laboratories, whose information is automatically uploaded to the table. This implies the need for a homogenization process for the information provided by the different labs, which might also provoke misunderstandings and associated errors.

    - Colonoscopy: The anatomo-pathological results of this test comes from several pathology laboratories, whose information is translated to the tables by health staff, which may also imply additional errors. Regarding the findings, the tables distinguish between the information about polyps and cancer lesions detected in the colonoscopy.

The whole information regarding the test procedure, preparation, exploration and findings is analyzed and entered into the platform by health professionals with different roles.

In summary, the information in the database comes from different external databases whose information is automatically dumped as well as from data filling by hospital staff with different roles. In these situations where several agents are involved and different information is crossed, we must ensure a good database design, proper data integration and an appropriate data checking and validation.

2.2 Entity-relationship model

The entity-relationship model facilitates the representation of the relationships between the entities. The main objectives of having an entity-relationship model are [14, 15]:

- To allow a high degree of independence between the application/platform and the internal representation of data.

- To provide a solid basis for addressing data consistency and redundancy.

Figure 1 shows the entity-relationship diagram of our database that represents the relationships between the entities. For simplicity, these relationships are shown

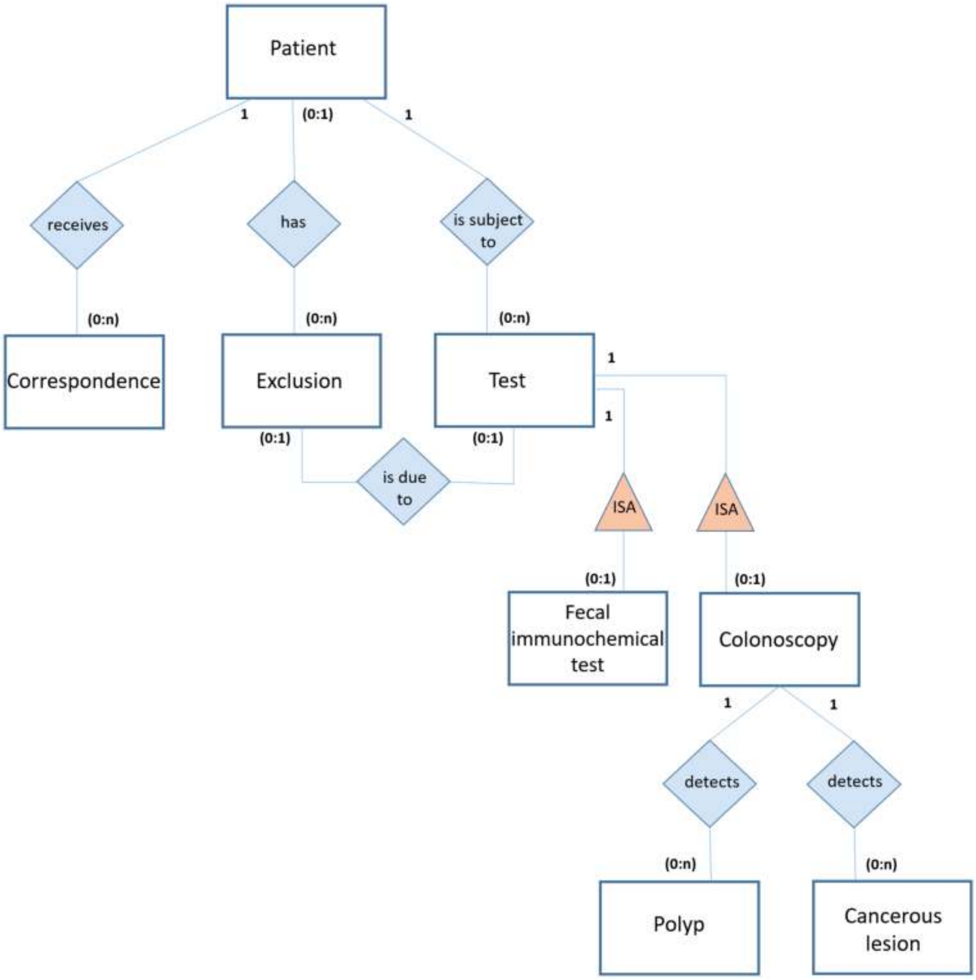

Figure 1.
Entity–relationship model using Chen notation.

in the diagram by means of the Chen notation [16]. The main relationships among data that may be found in the database of the colo-rectal cancer screening program are:

- One patient belonging to the target population may not have any invitation since their invitation to the program has not been processed yet, or may have received more than one invitation to the program over the years (1–0:n).

- A patient may (or may not) be excluded from the program (0:1–0:n) for several reasons. As mentioned above, this exclusion may (or may not) be due to a negative FIT result or colonoscopy findings (0:1–0:1).

- The target patient decides whether (or not) to undergo an FIT and also a colonoscopy if the FIT result is positive (1–0:n).

- In each colonoscopy, findings can be detected (or not) such as cancer lesions and/or polyps (1–0:n).

The fields for each entity are presented below. In total, the database contains about 140 fields.

Patient entity. The patient entity contains 12 attributes with basic patient demographic information in addition to his/her identifier. These fields are related to the date of birth, sex, the round in which the program is at the time in which the patient was enrolled, as well as the place of residence, the health district and the hospital to which the patient belongs.

Exclusion entity. The exclusion entity contains 6 attributes related to the period of exclusion from the program (date of exclusion and, in the case of temporary exclusion, the date of inclusion), the reason for exclusion, a number that determines the priority of exclusion, a binary field that determines whether the exclusion was entered manually and a free text field for comments. If a patient had more than one exclusion, the one with the highest priority prevails. As shown in **Figure 1**, in addition to these fields, the table contains the unique exclusion identifier, the patient identifier, and the test identifier if the exclusion was due to such test.

Correspondence entity. The correspondence entity contains 8 attributes which are as follows: the time when the correspondence was sent (date and time), the type of correspondence sent to the patient (invitation to the program, FIT result, date of scheduled colonoscopy), the round in which the patient was enrolled at the time when the correspondence was received, a binary field that determines if the patient agreed to participate in the program, a binary field that determines if the test recipient was received successfully, a binary field that determines if the patient was included in the program on demand and a free text field for notes. In addition, the table contains the unique identifier of the correspondence and the patient identifier.

Test entity. The entity related to the tests of the screening program contains a large number of attributes (>80). For simplicity, we present here only a high-level description with additional detail for the most relevant aspects. Specifically, the entity contains attributes related to the patient's condition prior to the test and after ending the patient's cycle (round, if the cycle ended, the reason for such ending and the patient's situation after finishing the round).

Regarding the FIT inheritance table, the fields are associated with the date of the interview at primary care, the date of the test and the result of the test. In particular, a field for continuous values of blood concentration in feces (ng/ml), a binary field that determines whether the test was positive and fields that determine whether the sample and the test were correct.

Concerning the colonoscopy inheritance table, it contains a big number of fields related to the following information: basic colonoscopy information (date and time of the scheduled colonoscopy, whether it was performed or not, actual date and time of the colonoscopy and the reason for being performed), colonoscopy preparation (drugs, tolerance, modality, colonic preparation and Boston scale [17]), the process during the colonoscopy (tolerance, which zone was reached, the duration, adequacy of the colonoscopy, etc.), the treatment used during the colonoscopy (type of sedation, type of endoscopic treatment, etc.), the findings found during the colonoscopy (main result: normal colonoscopy, non-neoplastic pathology, polyps, polyposis, cancer, cancer associated with polyposis; risk degree: no risk, low risk, medium risk, high risk and cancer; number of polyps, adenomas, cancer lesions), with the possible complications after the operation (type of complication, whether hospitalization was required and if the patient passed away within the following 30 days…) and possible repetitions of the colonoscopy if required.

Polyp entity. The polyp entity contains, in addition to the identifier of each polyp and the colonoscopy test identifier, 12 attributes concerning the order, size, histology, dysplasia, shape and location of the detected polyp as well as the method for the polypectomy performed, the treatment, the removal performed, etc.

Cancer lesion entity. The cancer lesions entity contains, in addition to the identifier of each lesion and the colonoscopy test, 12 attributes related to the order, size, histology, location of the lesion detected, as well as the stage of the cancer lesion, presence of occluding structure, the type of primary resection and the type of chemotherapy or radiotherapy if applied.

The entity relation model is therefore clear and well defined. However, we should not forget that, in these situations where several agents are involved and different information is crossed, we must ensure proper data integration from a correct database design. This is analyzed in the next section.

3. Analysis and recommendations

An incorrect design of the database and/or the platform often ends up with deficiencies, noise and mistakes in the data, which might prevent a rigorous analysis. In order to have information of sufficient quality to guide appropriate clinical decisions, it is necessary to follow basic principles for data collection and management.

The underlying overall objective of the chapter, and of this section in particular, is to establish from a general perspective, a set of basic principles regarding the integrity, consistency and coherence of data that must be met by any data management system. In particular, for our case study, we thoroughly analyzed whether each of these principles was met and, if not, we proposed a series of recommendations to mitigate the noise of the data and improve the quality of data management. In this analysis it was fundamental to work in a multidisciplinary team with biomedical, statistical and database experts.

The derived recommendations correspond to prospective improvement actions related to data filling, platform characteristics and database design. However, if the information is intended to be used retrospectively, it is also necessary to carry out an additional data curation action. In the following section we explain this curation process in our case study.

In summary, the underlying objective is to highlight the importance of an effective data governance [18, 19], a concept that refers to the ability of an organization to guarantee high quality data throughout its lifecycle, ensuring principles such as availability, easy use, consistency, integrity and security of data. The data manager must ensure such data governance principles and processes.

This concept is crucial as organizations rely more and more on data analysis to optimize their processes and to take relevant decisions [20]. In our particular case, quality data are essential to extract statistical information such as the screening program indicators, or to carry out studies with the objective of improving the overall healthcare system. Some examples are the establishment of the cut-off point to undergo a colonoscopy, a risk analysis to identify risk factors and decisions taken to minimize the undetected lesions, but always based on data evidence.

3.1 Principles

Some of the basic principles, related to data integrity, coherence and consistency, that we analyzed are the following:

- Information utility: All fields defined in the database (or variable to be introduced in the platform) must be filled for some entity (or record).

- Maintenance of consistency: The database or data manager must ensure the stability of the information to any change in the procedure/process and/or to any data dump from an external database.

- Redundancy control: Each register should be uniquely identified. A good database design avoids having more than one field identifying the same event.
- Clarity of the data dictionary: The information of each field of the database (or variable to be introduced in the platform) must be clearly established without any doubt for any user. The information of the field is related to:
    - Name of the field
    - Description of the field (unambiguity of the information)
    - Mandatory
    - Data type
    - Default value
    - Range of values
    - Primary key or foreign key
    - Table to which it belongs
- Management of relationships: The relationships between the fields of the database (or variables in the platform) must be established clearly.
- Control fields in the tables: Fields that identify the creation date, last change date, deletion date, deletion bit, creation user, last change user/process and deleted user/process allow to control the process changes in the data management.

Without loss of generality, we present in this section the analysis of these principles for the fields analyzed and introduce general and specific recommendations to comply with these principles and guarantee good data quality.

Information utility. First, the completeness of the fields in the database tables (about 140 fields) was analyzed. We detected two fields that were not filled, and eight fields defined in the database that were not filled for any entity. The latter can be variables that were defined at the beginning but were never used. To comply with the principle of useful information and to maintain a clean database, these variables should be removed.

Consistency of the information. First and regarding procedure changes, we detected some variables that were no longer used from a certain time; in particular, examples are the binary field that determines whether the patient agrees to participate in the program and the field that determines whether the patient was included in the program on request. **Figure 2** displays the time graph that represents the completeness of these variables over time, where the value 1 indicates completion. As can be seen, from mid-2016, the variables were not completed even once. In this case, and in order to maintain the consistency of the information, from that date of change, the variable should disappear from the platform and the values in the database should be filled to null by default.

Another example concerns the type of correspondence. Currently (from 2017) 3 types of correspondence are delivered: invitations to the program, notification of negative FIT result and notification of positive FIT result along with the scheduled

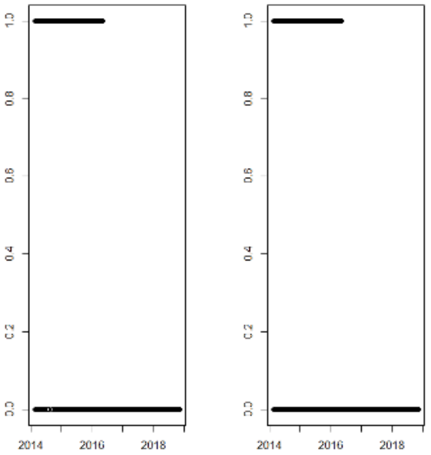

Figure 2.
Time chart of completeness.

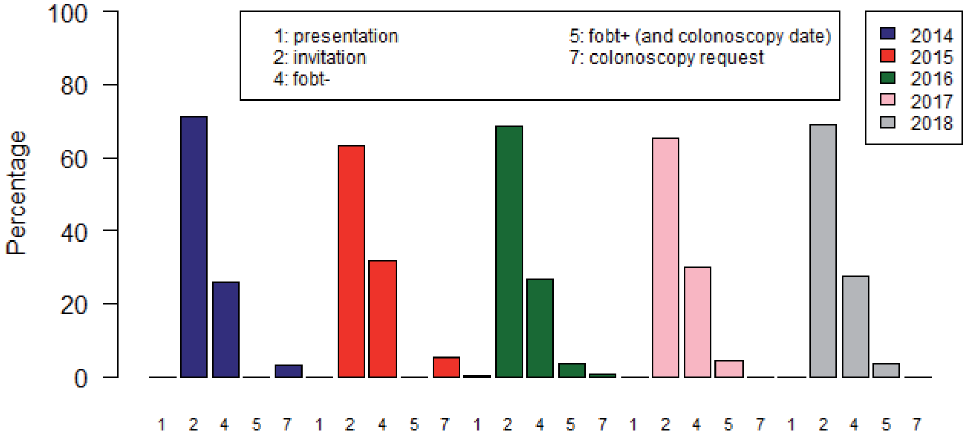

Figure 3.
Distribution of type of correspondence over the years.

date for the colonoscopy. However, previously, the procedure for the FIT positive result and the colonoscopy request was different, as shown in **Figure 3**.

Therefore, currently the value of the field corresponding to the positive FIT notification together with the scheduled date for the colonoscopy (value = 5) refer to a different type of correspondence than in previous years. These changes in the definition of the fields are not recommended since they do not ensure the stability of the information. If they are eventually made, they should be documented and to keep in mind that the historical information should be translated into its current equivalent.

As discussed in the previous section, some of the information in the screening program were extracted from external databases. Therefore, it is also important to analyze its source and the quality of such external sources. Consequently, the quality of the information in the external databases was analyzed and some shortcomings were identified.

For example, in the exclusion process, those exclusions due to findings of cancer lesions in the colonoscopy were considered as temporary exclusions (for 10 years), when it should be a permanent exclusion since the patient as part of the “high risk” group is transferred to the digestive service specialists. Another deficiency found was related to the date of exclusion and, consequently, of inclusion in the program. In particular, in the interview in primary care (OMI database), when a patient fulfills some reason for exclusion, the date of exclusion, that is stored is the one of the interview, and not the actual date when the reason for exclusion was detected. This is important since an erroneous date of inclusion leads to the patient being (falsely) part of the target population at a certain time or the opposite, i.e., not being part of the target population when he/she should be.

These shortcomings involve importing incorrect information into the database and, at best, manual human correction. Ideally, these deficiencies should be corrected from these external databases but since this control can be more difficult and limited, the recommendation regarding our database in these cases would be to correctly identify the possible cases and relationships that must be met (requirements of the screening program database). Also, it is important to make a procedure where only those records that do not induce conflict are updated in the database while the other cases should be reported to allow the user their modification and import/store them correctly. Finally, the automatic generation of reports is also desirable.

In addition to the above deficiencies, in particular in the information from the laboratory database (fecal immunohistochemical test), some records were detected whose information in some fields was crushed or deleted. An incremental import of the information from the external databases would guarantee and ensure the correct storage of the manual changes and would prevent their deletion (since the original unmodified information would not be reloaded).

Redundancy control. Another basic principle for a good database (or data manager) design is the control of redundancy. In particular, the database analyzed does not fully comply with this principle as it contains several fields that identify the same event and, therefore, with redundant/repeated information.

Some examples are the fields related to the result of the FIT: on the one hand, there is a binary field to determine whether the test is positive (> = 117 ng/ml) or negative and, on the other hand, the field representing the quantitative value of the test (ng/ml). Another example detected was the variables related to colonic preparation: colonic preparation in the left colon, colonic preparation in the right colon, colonic preparation in the transverse colon and the Boston scale (from 0 to 9). The latter is, by definition, the sum of the values of the three previous ones.

These examples are fields with a deterministic relationship, where some are the result of the information of others. Therefore, if redundancy control is not fulfilled and the redundant fields are maintained, at least it should be guaranteed that these relations are fulfilled in a deterministic way both in the database and in the platform, self-calculating the fields and/or restricting their values according to the information of the rest of the related fields. However, the analysis carried out revealed that these relationships were not considered in the platform or in the database. This can be observed in **Tables 1** and **2**. **Table 1** shows the qualitative variable of the FIT and the transformation to a categorical variable from the quantitative variable given the current cut-off point (117 ng/ml). **Table 2** shows the variable

Cuantitative/Cualitative	Negative FIT	Positive FIT	Total
Concentration < 117 ng/ml	59.95	0.02	59.97
Concentration ≥ 117 ng/ml	0.02	10.99	11.01
Empty value	27.42	1.59	29.01
Total	87.39	12.6	100

Table 1.
Contingence table (%): Fetal occult blood concentration.

Left+right+transverse/ Boston scale	0	1	2	3	4	5	6	7	8	9	Empty
0	6557	0	0	0	0	0	0	0	0	0	0
1	2	1	0	0	0	0	0	0	0	0	0
2	1	0	10	0	0	0	0	0	0	0	0
3	16	0	0	38	0	0	0	0	0	0	0
4	40	0	0	0	65	0	0	0	0	0	0
5	55	0	0	0	0	158	0	0	0	0	0
6	280	0	0	0	0	0	884	0	0	0	0
7	234	0	1	0	0	0	0	450	0	0	2
8	394	0	1	0	0	0	0	0	673	0	1
9	589	0	0	0	0	0	0	0	0	1100	2
Empty	1462	3	10	39	41	97	411	259	297	314	80211

Table 2.
Contingence table (absolute frequency): sum of the colonic preparation of the left, right and transverse part of the colon vs. Boston scale.

calculated as the sum of the values of the variables of the colonic preparation of the left, right and transversal part of the colon versus the qualitative variable of the Boston scale.

If the database and the platform guarantee these deterministic relations of redundant fields, the above tables should be diagonal matrices. However, the analysis showed this weakness in data consistency and showed that neither the platform nor the database considers these relationships, permitting the user an unrestricted completion of those fields, which could lead to data inconsistencies.

Therefore, the analysis carried out showed not only a lack of redundancy control, but also a lack of consistency in the data management. As a recommendation, redundant information should either be removed or, if not, these deterministic restrictions should be established both in the platform and in the database in such a way as to ensure that the relevant information entered is consistent and not confusing.

Data dictionary and relations. As mentioned above, in order to be clear about the meaning of each of the tables and their fields, it is advisable to prepare a priori a data dictionary where each of the fields of each table and the relationships to be established between them are clearly defined. The minimum information to establish, whenever possible, is the following: name and description of the field, mandatory (or not) field, type of data, default value and range of values. However, the analysis performed showed the non-existence of an explicit data dictionary.

An example of ambiguity in the definition would be the variable “round” which appears in both the correspondence table and the test table as well as in the patient’s table. Its name is ambiguous since its meaning leads to confusion, having two possible alternatives: it indicates either the patient’s round in the program or the current hospital’s call round.

It would be natural to think that the “round” variable in the correspondence table refers to the program round and the “round” variable in the test and patient tables refers to the patient round. However, after an exhaustive analysis of these variables, it was concluded that no clear definition of the variable could be extracted from either table. Specifically, if it were “round by patient” the following basic hypothesis should be fulfilled: if a patient has round 2, he/she must also have had round 1. However, patients with round 2 who had not had round 1 were detected in the data. If it were "round by program" the following basic hypothesis should be fulfilled: for the same patient the variable "round" should be increasing over time, that is, if a patient has round 2 at a certain moment, at later dates he or she should have round greater than or equal to 2. However, this was not fulfilled either. Therefore, as commented, we concluded that there is no clear definition of the variable and its completion may be ambiguous. Furthermore, this variable is of great importance since it allows the temporal follow-up of the patient in the program. Thus, the recommendation is crucial in this particular case: to establish a consistent definition of the round variable that is implemented both in the database and in the platform.

Another field information to be established is its mandatory character, if any. The information collected in these fields is the minimum information required to have quality information. In our case, the mandatory variables would be those containing the minimum information of the screening program. However, the analysis showed that there were no mandatory fields established (neither in the platform nor in the database). As an example, each FIT should have the minimum information of its date and its quantitative result (ng/ml), however this does not always happen, as shown in **Figure 4**.

Some fields, like those above, are of a permanent mandatory character, while others may have this nature depending on the values of other fields. For example, the field indicating the findings found in the colonoscopy should be mandatory if the colonoscopy was performed. A good database design should establish this obligation permanently or with restrictions in all necessary cases. As far as the platform is concerned, this obligation should also be established in such a way that all completed information cannot be saved if the minimum necessary information is not filled in.

The data type is another kind of fundamental information to be established for each field. The analysis carried out revealed a lack of consistency in this regard: there are several free text fields in the platform with no restrictions. In particular, the field that determines the time of the next colonoscopy is a free text field that provokes that each user is free to interpret the type of data, filling it in with three different types of data: strings, integers or dates. Examples are the following:

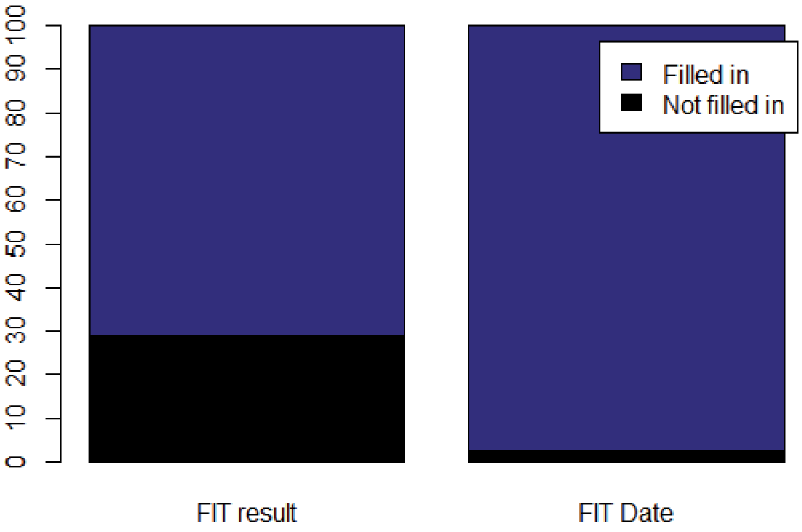

Figure 4.
FIT filling distribution.

"In 5 years", "Not required", "-5 years", "3", "09/05/2016", etc. This means that, in order to use this information, it is necessary to standardize it in the same format, which entails certain difficulties and limitations. For example, the user who filled the value "3" may have referred to months or years and, if this is not established in the type of data or in the definition, this information cannot be used in an analysis. In addition to this, the normalization process of a text-type field takes a great effort [21]. In our particular case, for example, one user entered "In 5 years", another filled the field ("Not required") when the message shows that it should not have been filled (he misuses it as a note) and another uses the mathematical minus sign ("- 5 years") which could mean that the next colonoscopy should be performed in less than 5 years or it could be a simple filling error.

It is therefore necessary to consciously establish the type of data to avoid problems of ambiguity that are difficult to deal with. In addition, the number of free text fields should be limited and, a training effort should be made for the staff who handle and fill the data in order to standardize and unify their interpretation.

The analysis also found that there were fields without a fixed default value or an inadequate default value either in the database or in the platform. For example, it was observed that in some numerical fields both the value 0 and the null value were used indistinctly as default values, which leads to ambiguity in the interpretation of the information for the value 0 which may indicate either the value itself or its default value.

An appropriate default value should be established for each field to avoid ambiguity in the subsequent interpretation of the information and to ensure adequate data quality.

In addition to a clear definition of the field, its mandatory nature, its data type and its default value, restricting the field to a certain range of values is also important in the definition of the data dictionary as it limits the information to possible values and mitigates noise, i.e. possible filling errors and ambiguity problems. If this restriction of the range of possible values is not contemplated, a series of warnings or alerts should be at least implemented to notify the user of an outlier value and the need to revise it.

At this point, both the platform and the database showed weaknesses as there is also a lack of constraints in this regard and no alerts were implemented. An example can be seen in **Table 3** that shows the distribution of values (minimum, quartiles (Q1, Q2, and Q3), mean and maximum value) in the field "weight (kg)" of the patient. On the one hand, weights of 0 kgs in the screening program are not possible, while at least 50% of the filled values took this value. This is an error that can potentially come from not setting the default value or from an incorrect default value, as mentioned above. This would imply that the value 0 was taken incorrectly as default value, distorting the statistics. On the other hand, very high weights (e.g. 81,700 kg) are also inconsistent and may come from human error in the filling process. This example shows that if the fields included a range of possible concrete values or outlier alerts, these errors would be mitigated and, consequently, better quality data is got.

One of the key principles for ensuring consistency in data is to look at the relationships between fields through appropriate constraints. Some of these

Min	Q1	Q2	Mean	Q3	Max	Num. NA's
0	0	0	34	70	81,700	80,415

Table 3.
Patient weight distribution.

relationships may be deterministic, such as those between fields that identify the same event and, as discussed above, in these cases, the constraint must be clear, and preferentially the associated values should be self-calculated. Other relations correspond to restrictions in the values of a field depending on the value or values of other fields and others to restrictions related to the mandatory filling of a field depending on the filling of others.

In relation to this, the analysis carried out found a lack of constraints in the fields which may lead to data inconsistencies, sometimes difficult to correct. For example, if this principle were fulfilled, the following should occur: if the FIT concentration is greater than or equal to 117 ng/ml (cut-off point), the FIT result variable should not be "positive"; however, this restriction was not always considered. A characteristic example related to the restriction of values depending on the value of another field is the one of the monitoring dates that should follow a chronological order (e.g. date of invitation<date of sample reception<FIT result date<colonoscopy date…), however, these inequalities were not always met. Other example is the following: if the field that determines whether the colonoscopy was performed is equal to "No", then the variables related to the colonoscopy should not be filled.

Establishing these constraints, both in the database and in the platform by activating or not the fields in the platform, is fundamental to avoid possible inconsistencies in the data which, on some occasions, can be remedied by curing the data and, on other cases, it is unfeasible to know what the real information in the data is. These restrictions can also be accompanied by alerts or warnings in the platform to help the user and avoid mistakes.

These constraints must be implemented not only in the data filling but also in the deletion, that is, they must guarantee that when the user the value of a variable, the data related to such value must be deleted. For example, if the variable indicating whether a colonoscopy was performed changes its value from "Yes" to "No", then all variables related to the colonoscopy should be set to their default or null value, thus deleting their last filled-in values.

In summary, the analysis showed that a conscious establishment of the values for each field, the data dictionary and a good training of the staff who handle the data is crucial. The more limited and defined the information to be entered is, the better the data will be processed, resulting in fewer errors and less problems of ambiguity, many of which are difficult to deal with subsequently. In addition, the implementation of alerts in the platform could also help to mitigate those filling errors. It is also crucial to thoroughly analyze all possible relationships between all fields in the database and to establish these constraints in the database or in the data manager.

This section has highlighted the inconsistencies, incoherence and errors (some difficult to fix) that can occur in a database if it does not comply with the basic principles of good data management, especially when different agents are involved (external databases, staff with different roles, etc.). As a first conclusion, a good data governance is required to guarantee data quality permitting the extraction of reliable knowledge.

4. Data curation process

The recommendations suggested are referred to improvement measures to comply with the basic principles for a correct design of the database, with the aim of improving the quality of data in the future. However, on many occasions, such data are needed to be used retrospectively. In such cases, a previous curation process is required to eliminate as many errors as possible. In our particular case, the information in the CRC screening database was used to obtain annual indicators of the screening program [22] and to analyze different scenarios for decision making based on the FIT cut-off point, the colonoscopies offered, the target population and the risk factors in order to minimize undetected lesions.

The data curation process carried out was done in the most conservative way possible, and it was mainly based on the relationships that can be established between the fields, recalculating the inconsistent values according to the values of the most reliable/secure reference fields and, if this was not possible, either setting the values that produce inconsistencies to null or finally by removing the whole record from the data set to be analyzed. To carry out this process it was necessary to have a multidisciplinary team composed of statisticians and clinical staff so that statistical knowledge and decision making was supported by knowledge on the environment.

This section explains the main steps and difficulties of the curation process carried out chronologically in order to obtain a clean dataset. The variables presented are the most representative and important ones to carry out the studies required: related to FIT, colonoscopy and follow-up.

4.1 Fecal Immunohistochemical test (FIT)

As commented in the previous section, the relationship between the quantitative variable of the FIT concentration and the qualitative variable (negative, positive FIT...) is not fulfilled in a deterministic way as it should be. Below we present the most representative cases of incoherence and how they were cured:

- Records with concentration = 117 ng/ml but with "negative" FIT result and records with negative concentration were detected. Cooperation with health staff was key here. With regard to the first case, after several meetings, it was concluded that they were values from the laboratory where the report specified the value at "-117", referring to "less than 117". For the second, it was deduced that a hospital considered the cut-off point of 117 ng/ml as negative. In order to standardize the information from all hospitals it was decided to re-establish the value of the quantitative concentration at 116 ng/ml in these records.

- Records were detected with concentration > 117 ng/ml but with negative FIT result. It was found that these records were two tests for the same patient in which the second test overwrote the qualitative value of FIT but maintained the old value of the quantitative one. In these cases, it was decided to follow the more conservative decision and take the information of the first one (with its quantitative value) and recalculate the qualitative value.

- Records with concentration < 117 ng/ml were detected but with positive FIT results: This was one of the errors that were not possible to re-establish and, therefore, after several meetings, it was decided to establish the values of the concentration at null and to save the identification of those patients in order to establish the correct value of the concentration in the future in case they are identified.

- Once the concentration values were corrected, the qualitative variable of the FIT was recalculated in the cases where the concentration was different from zero.

4.2 Colonoscopy

The variable that indicates the performance of the colonoscopy (colonoscopy variable) is also strictly related to the result of the FIT, and, obviously, with the variables associated with the colonoscopy. Below we present the inconsistencies found and the steps followed for their curation:

- Records were detected in which the FIT result was negative, and the colonoscopy variable took the value "Yes". After studying it, it was concluded that this error was likely due to an overwriting or crushing of the data. Since this error does not allow the recovery of realistic values from the record, it was decided to remove these records from the data set.

- Records were detected that had a value in the variable that indicates the result of the colonoscopy (normal colonoscopy, polyps, cancerous lesion...) and with a completed colonoscopy date, and yet the colonoscopy variable did not take the value "Yes". These cases are a clear example of error because the colonoscopy variable was not defined as mandatory in the database. In these cases the value of the colonoscopy variable was reset to "Yes".

- All records with no colonoscopy value were reset to 0. This is due to the lack of definition of the default value of this field in the database.

- Records with no information on colonoscopy-related variables were detected, yet the colonoscopy variable was equal to "Yes". This error was possible thanks to the lack of constraints in the database. In these cases the colonoscopy variable was reset to "No".

- The variable that determines the reason for the exploration in the cases in which the colonoscopy variable was equal to "No" was reset to null.

4.3 Follow-up

When the records have all their process information filled, they should have the "end of cycle" variable filled to "Yes". Therefore, in the analysis we must take the records that have finished their cycle ("end of cycle" = "Yes"). However, this variable was not defined as mandatory in the database and, therefore, presents inconsistencies that were solved according to the following rules:

- All records with the colonoscopy variable equal to "No" should be closed cycles and, therefore, the variable "end of cycle" should take the value "Yes".

- All records with the variable that determines the reason for the end of the cycle completed refer to the patient's closed cycles and, therefore, the variable "end of cycle" should take the value "Yes".

- All colonoscopies prior to the last year (2018) with "end of cycle" equal to "No" would really be considered as closed cycle ("end of cycle" = "Yes") since the information should take less than one year to be filled in.

The variables related to dates are important since they allow the information recorded to be followed up by years. Therefore, their completion should be ensured as far as possible, in particular the date of the sample result and the date of the colonoscopy which are the most crucial ones. The dates considered for completion are the following: date of invitation to the program, date of interview at primary care, date of reception of the sample, date of result of FIT, date scheduled for the colonoscopy, date of colonoscopy. Chronological completion was done as follows:

- In the records where the FIT result date was not completed, it was set as follows in order of preference: result date = colonoscopy date –1-month, result date = sample receipt date, result date = OMI date, result date = colonoscopy schedule date-1 month, result date = program invitation date+1 month.

- In the records where the colonoscopy date was not completed, it was established considering the same date as the date programmed or considering the result date of the FIT-1 month.

In this section the most important aspects of the curation process carried out in order to obtain a clean data set with reliable information on which to carry out future analyses have been introduced. Despite they are specific, it has been shown, the wide range of potential bugs that may appear due to a wrong design of the database.

5. Conclusions

Good data management and consequently good data quality must comply with some basic principles. This is especially important when those data are used to support decision makers in public health services. In this chapter, we analyze the database of a regional CRC screening program to identify the weaknesses in the process of data collection, providing some guidelines for future maintenance. We also identified incorrections in the database design that may lead to data errors. General and specific recommendations were suggested to meet the requirements of data integrity, consistency and coherence.

However, most of these recommendations are forward-looking suggestions, i.e. they will improve the quality of future data from the moment they are considered. Simultaneously, and in order to be able to exploit the information retrospectively, it was necessary to make a data curation of the historical information. To do this, a clean-up process was followed in the most conservative way possible, re-establishing values, cleaning-up some data and discarding repetitive or non-essential data, trying to eliminate as many errors as possible and guarantee good quality data both prospectively and retrospectively. This process is time costly and tedious, but it is an essential first step in data governance to extract reliable knowledge and taking correct decisions.

In summary, this analysis showed the importance of data quality and curation to get a robust, consistent and reliable database, as well as the need for a good design of the data acquisition process and, finally, a proper and coherent maintenance system, especially in health systems where the decisions derived from the analysis of databases may be critical.

Acknowledgements

We acknowledge the help provided by the Department of Health of the Aragón Government for allowing us the access to anonymized patient data and to the EU Program for Employment and Social Innovation (EaSI) for partial financial support.

Author details

Rocio Aznar-Gimeno[1], Patricia Carrera-Lasfuentes[2],
Vega Rodrigalvarez-Chamarro[1], Rafael del-Hoyo-Alonso[1], Angel Lanas[2]
and Manuel Doblare[3]*

1 Technological Institute of Aragon (Itainnova), Zaragoza, Spain

2 Aragon Institute of Health Research (IISAragon) and CIBERehd, Zaragoza, Spain

3 Aragon Institute of Health Research (IISAragon) and CIBERbbn, Zaragoza, Spain

*Address all correspondence to: mdoblare@unizar.es

References

[1] Abadi, D., et al. The Beckman report on database research. Communications ACM, 2016, vol. 59(2), p. 92-99

[2] da Costa, C.A., Pasluosta, C.F., Eskofier, B., Bandeirada, D., Rodrigoda, S. and Righi, R. Internet of Health Things: Toward intelligent vital signs monitoring in hospital wards. Artificial intelligence in medicine, 2018 vol. 89, p. 61-69

[3] Bhalla, S., Sachdeva, S. and Batra, S. Semantic interoperability in electronic health record databases: Standards, architecture and e-health systems. In 5th International Conference on Big Data Analytics, Hyderabad, India, 2017. Lecture Notes in Computer Science book series (LNCS, volume 10721)

[4] Biswas, S., Sharif, K., Li, F., Latif, Z., Kanhere, S.S. and Mohanty, S.P. Interoperability and Synchronization Management of Blockchain-Based Decentralized e-Health Systems, in IEEE Transactions on Engineering Management, 2020, vol. 67(4), p. 1363-1376, doi: 10.1109/TEM.2020.2989779.

[5] Dipak, K., Beale, T. and Sam Heard. The openEHR foundation. Studies in health technology and informatics, 2005, vol. 115, p. 153-173. PMID: 16160223.

[6] Pathak, J., Bailey, K.R., Beebe, C.E., Bethard, S., Carrell, D.S., Chen, P.J., … and Chute, C.G. Normalization and standardization of electronic health records for high-throughput phenotyping: the SHARPn consortium. Journal of the American Medical Informatics Association, 2013, vol. 20(e2), ep. 341-e348 doi: 10.1136/amiajnl-2013-001939.

[7] Sachdeva, S. and Bhalla, S. Semantic interoperability in standardized electronic health record databases. J. Data Inf. Qual. (JDIQ), 2012 vol. 3(1), p. 1 https://doi.org/10.1145/2166788.2166789

[8] Hoffman, S. and Podgurski. A. Big bad data: law, public health, and biomedical databases. The Journal of Law, Medicine & Ethics, 2013 vol. 41, p. 56-60 https://doi.org/10.1111/jlme.12040

[9] Batra, S. and Sachdeva, S. Pre-Processing Highly Sparse and Frequently Evolving Standardized Electronic Health Records for Mining. Handbook of Research on Disease Prediction Through Data Analytics and Machine Learning. IGI Global, 2020. P. 8-21 doi: 10.4018/978-1-7998-2742-9.ch002

[10] Satti, F. A., Ali, T., Hussain, J., Khan, W. A., Khattak, A. M., and Lee, S. Ubiquitous Health Profile (UHPr): a big data curation platform for supporting health data interoperability. Computing, 2020, vol. 102(11), 2p. 409-2444. https://doi.org/10.1007/s00607-020-00837-2

[11] Pezoulas, V. C., Kourou, K. D., Kalatzis, F., Exarchos, T. P., Venetsanopoulou, A., Zampeli, E., … and Fotiadis, D. I. Medical data quality assessment: On the development of an automated framework for medical data curation. Computers in biology and medicine, 2019, vol. 107, p. 270-283. doi: 10.1016/j.compbiomed.2019.03.001

[12] Feder, S.L. Data quality in electronic health records research: quality domains and assessment methods. Western journal of nursing research, 2018, vol. 40(5), p. 53-766. doi: 10.1177/0193945916689084

[13] Weiskopf, N. G., and Weng, C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research.

Journal of the American Medical Informatics Association, 2013, vol. 20(1), p. 44-151. doi:10.1136/amiajnl-2011-000681

[14] Elmasri, R. and Navathe, S.B. (eds) The relational data model and relational database constraints. In Fundamentals of Database Systems, Pearson Addison-Wesley, 2013. ISBN-0133970779

[15] Codd E.F. A Relational Model of Data for Large Shared Data Banks. In: Software Pioneers (Broy M., Denert E. (eds)). Springer Verlag, 2002 https://doi.org/10.1007/978-3-642-59412-0_16

[16] Chen, P.P-S. The entity-relationship model—toward a unified view of data. ACM Transactions on Database Systems, 1976, vol. 1(1), p. 9-36. Doi:10.1145/320434.320440

[17] Calderwood, A.H. and Jacobson, B.C. Comprehensive Validation of the Boston Bowel Preparation Scale. Gastrointestinal Endoscopy, 2010 vol. 72(4) p. 686-692. Doi: 10.1016/j.gie.2010.06.068.

[18] Dama International. Dama-DMBOOK: Data Management Body of Knowledge. Technics Publications, LLC, 2017 ISBN-1634622340

[19] Khatri, V. and Brown, C.V. Designing data governance. Communications of the ACM, 2010, vol. 53, no 1, p. 148-152. Doi: 10.1145/1629175.1629210

[20] Wieten, E., Schreuders, E.H., Nieuwenburg, S.AV., Hansen, B.E., Lansdorp-Vogelaar, I., Kuipers, E.H., Bruno, M.J. and Spaander, M.C.W. Effects of increasing screening age and fecal hemoglobin cutoff concentrations in a colo-rectal cancer screening program. Clinical Gastroenterology and Hepatology, 2016, vol. 14, no 12, p. 1771-1777. Doi:10.1016/j.cgh.2016.08.016

[21] Kreimeyer, K., Foster, M., Pandey, A., Arya, N., Halford, G., Jones, S.F., Forshee, R., Walderhaug, M. and Botsis, T. Natural language processing systems for capturing and standardizing unstructured clinical information: a systematic review. Journal of biomedical informatics, 2017, vol. 73, p. 14-29. Doi: 10.1016/j.jbi.2017.07.012

[22] Llop, E.S., Cano del Pozo, M., García Montero, J.I., Carrera-Lasfuentes, P. and Lanas A. Colo-rectal cancer screening program in Aragon (Spain): preliminary results Gaceta sanitaria, 2018, vol. 32, no 6, p. 559-562. doi: 10.1016/j.gaceta.2017.05.014