Data Models for an Imaging Bio-bank for Colorectal, Prostate and Gastric Cancer: the NAVIGATOR Project

Andrea Berti*[††], Gianluca Carloni*[††], Sara Colantonio*, M. Antonietta Pascali*, Paolo Manghi*, Pasquale Pagano*, Rossana Buongiorno*[†], Eva Pachetti*[†], Claudia Caudai*, Domenico di Gangi*, Emanuele Carlini*, Zeno Falaschi[‡], Esther Ciarrocchi[‡], Emanuele Neri[‡], Elena Bertelli[§], Vittorio Miele[§], Roberto Carpi[¶], Giulio Bagnacci[||], Nunzia di Meglio[||], M. Antonietta Mazzei[||] Andrea Barucci**

*Institute of Information Science and Technologies - ISTI-CNR, Pisa, Italy Email: [name.surname]@isti.cnr.it

[†]Department of Information Engineering, University of Pisa, Pisa, Italy

[‡]Dpt of Translational Research and of New Surgical and Medical Technology, University of Pisa, Pisa, Italy

[§]Azienda Ospedaliera Universitaria Careggi, Florence, Italy

[¶]Azienda USL Toscana Centro, Florence, Italy

[||]Azienda Ospedaliera Universitaria Senese, Siena, Italy

** Institute of Applied Physics - IFAC-CNR, Florence, Italy

[††]*These authors share the first authorship*

***Abstract*—Researchers nowadays may take advantage of broad collections of medical data to develop personalized medicine solutions. Imaging bio-banks play a fundamental role, in this regard, by serving as organized repositories of medical images associated with imaging biomarkers. In this context, the NAVIGATOR Project aims to advance colorectal, prostate, and gastric oncology translational research by leveraging quantitative imaging and multi-omics analyses. As Project's core, an imaging bio-bank is being designed and implemented in a web-accessible Virtual Research Environment (VRE). The VRE serves to extract the imaging biomarkers and further process them within prediction algorithms. In our work, we present the realization of the data models for the three cancer use-cases of the Project. First, we carried out an extensive requirements analysis to fulfill the necessities of the clinical partners involved in the Project. Then, we designed three separate data models utilizing entity-relationship diagrams. We found diagrams' modeling for colorectal and prostate cancers to be more straightforward, while gastric cancer required a higher level of complexity. Future developments of this work would include designing a common data model following the Observational Medical Outcomes Partnership Standards. Indeed, a common data model would standardize the logical infrastructure of data models and make the bio-bank easily interoperable with other bio-banks.**

***Index Terms*—Data models, entity-relationship diagram, imaging bio-banks**

I. Introduction

As the name implies, bio-banking refers to the practice of gathering and storing extensive collections of biological materials and their specific metadata for scientific research [1]. A bio-bank, or bio-repository, is indeed a valuable resource enabling researchers to access multiple types of medical data, such as samples of bodily fluids or tissues. The goal is usually to support research on personalized medicine on a *multiomic* perspective. Actually, multiomic sciences, such as genomics, proteomics, dosiomics, radiomics, radiogenomics, usually entail the production and analyses of enormous amounts of data [2], [3].

The availability of large volumes of high-quality data is essential in today's data-driven science. Following the evolution of digital innovation in healthcare, it is clear that data management requires advances able to facilitate the development of platforms for more effective usage of large volumes of data. Bio-banks play a central role in this scenario [4].

Besides the more common collections of body fluids and tissues, bio-banks are nowadays advancing to integrate also collections of medical imaging data. *Imaging bio-banks* are organized repositories of medical images, usually associated with imaging biomarkers. Most of the existing imaging bio-banks focus on cancer-related data and oncology imaging biomarkers collections [5], [6]. Their goal is to exploit the wealth of information hold in imaging data to discover novel diagnostic and prognostic biomarkers, especially when considering cancer phenotypes.

The NAVIGATOR Project, funded by the Tuscany Region, aims to establish the first regional imaging bio-bank, with the goal of boosting precision medicine in oncology. To do this, the Project plans to employ quantitative imaging and multi-omics analyses towards a better understanding of cancer biology, cancer care, and cancer risks. [7]. In this respect, the Project relies on two main tools: an open imaging bio-bank and an open-science-oriented Virtual Research Environment (VRE). The bio-bank is being designed to collect and preserve quality and curated imaging data as well as the

The research leading to the paper was carried out under the project NAVIGATOR, funded by the Tuscany Region under the PAR FAS BANDO RICERCA SALUTE 2014-2020

TABLE I
Healthcare Partners in NAVIGATOR Project and Neoplasia of Interest

Healthcare Partner	Neoplasia
University of Pisa, Pisa, Italy	Colorectal
Careggi University Hospital (AOUC), Florence, Italy Azienda USL Toscana Centro (AUSL-TC), Florence, Italy	Prostate
University Hospital of Siena (AOUS), Siena, Italy	Gastric

related omics data, in a secure and privacy-preserving model. Data include Computed Tomography (CT), and Magnetic Resonance Imaging (MRI) data for three relevant tumor cases (i.e., colorectal, prostate, and gastric cancer). Clinical data from regional healthcare services, molecular and liquid biopsy data are planned to be collected as well. The VRE serves as a web-accessible digital laboratory to process the multi-omics data for two primary purposes: (i) to extract gold-standard and novel imaging biomarkers based on radiomics analyses; and (ii) to create and test *digital patient models* based on cancer phenotype and risk stratification. Big data analytics and Artificial Intelligence are planned to be used in this respect.

The building block of the bio-bank design is the definition of the data model, namely an abstract model that organizes and standardizes the relationship between data elements and real-world entities' properties. Typically, the tool used to build a data model is the entity-relationship diagram (ERD), which is afterwards translated into a physical database.

In this paper, we present the first implementation of the data models and the corresponding ERDs for the three cancer use-cases of the NAVIGATOR Project. These models are serving as a basis for designing the physical database and repository connected with the VRE. The data models are also serving for mapping the entities to the OMOP Common Data Model and the Standardised Vocabularies related to it [8]. Though still a preliminary step, we consider that the data models themselves can be of interest for the research communities working in formalising clinical and radiological data.

II. Methods

A. Cancer use-cases

The NAVIGATOR Project involves four healthcare partners, and includes three different cancer use-cases. Each healthcare partner is in charge of gathering clinical and laboratory findings of the tumor and relating them to the radiological images, which will be filed and processed for biomarkers extraction. Table I shows the healthcare partners and the particular neoplasia on which they focus.

B. Requirements Analysis

As a first step in the life-cycle of our model, we carried out a requirements analysis with the clinical partners. Its prime purpose was to define all the clinical variables, their type, value range, and measurement units. In this process, it was possible to identify those variables with frequently missing values for retrospective cohorts of patients and, hence, exclude them. Additionally, in the continuous confrontations, the clinical experts described the clinical workflow patients typically undergo for each cancer use-case.

We concluded this phase by drafting a detailed requirements specification that synthesized all of the healthcare partners' needs. From that, significant differences emerged both in the patient protocol and the type of data collected for the three use-cases. For this reason, we found it difficult to conceive a transversely valid ERD. Indeed, that would have required strong generalization assumptions and too abstract entities in the diagram. Specifically, most of the entities in the ERD would have been parents of three children entities, one for each cancer use-case and with specific peculiarities. As a result, we would have obtained a remarkably complex diagram, probably undermining the efficiency of ERDs. Consequently, we built three distinct data models for the cancer use-cases of the Project, connected mainly based on patient's identification.

C. Entity-Relationship Diagram

We chose to adopt the popular ERD approach to develop these three data models. Indeed, ERDs are great tools that provide a clear database design visualization. An easy-to-interpret representation is fundamental in contexts that require a precise data structure. That also contributes to better communication between technical and clinical experts.

In the design process of ERDs, we first identified the entities with their attributes from the requirements specification. Then, we defined the relationship between different entities and added cardinalities to them, i.e., the amount of data that would come from each entity. Based on that, we drew the ERDs and asked the clinicians for a validation check. This process was iterative, with continuous interactions between the two counterparts, delivery of prototypes, and implementation of changes. In the end, we reached the final form of ERD for all the cancer use-cases.

III. Results

In this section, we describe the main features characterizing the three ERD diagrams. From here on, words in *italics* will represent the names of entities in the diagrams. For visualization clarity, we omitted from the diagram figures the attributes of each entity. Each model is presented in a narrative fashion to illustrate how the entities were identified based on the clinical workflows and to not exceed the page limit.

A. Colorectal Cancer Data Model

The resulting model for colorectal cancer is reported in Fig. 1. The diagram evolves from the *Patient* entity and takes into account data regarding all the stages the patients undergo during their diagnostic-therapeutic journey. First, data regarding patient's risk factors (e.g., familiarity, age ≥ 50, previous neoplasms) are collected in the *Anamnestic Data* entity. Then, patients are subjected to *Rectal Explorations* and colonoscopy with *Histological Specimen Data* for the initial diagnosis. Here, the clinical question is whether the mesorectal fascia is invaded or not, or whether the peritoneum itself is invaded (for high tumors) or whether the internal and external sphincter itself is invaded (for low tumors). Then, a local staging MRI is performed to determine how to treat the tumors. In this regard, the entity *MR Study* collects all the information regarding the date of exam, field strength, etc. All the DICOM tags are included to report the examination parameters. The MRI sequence of choice is T2-weighted, but Diffusion Weighted Imaging (DWI) is also used to better look at the lymph nodes. Since an MR study may comprise several specific acquisitions, the *Image Stack* entity includes their information. From the MR study, the radiologists extract a series of measurements, which are collected in the *Image-based Measurements* entity in the diagram. Those comprise information about the primary tumor, but also the presence of positive lymph nodes outside the mesorectal fascia and the vein invasion. Furthermore, to understand distant metastases for the complete staging, a chest-abdomen CT with a contrast medium can be performed (*TNM:M stadiation* entity). In the end, *Therapy* and *Outcome Evaluation* complete the diagram.

B. Prostate Cancer Data Model

When a patient presents symptoms related to prostate cancer, a rectal examination is performed. The resulting outcome can be collected in the ERD (Fig. 2) under *Symptoms & Examinations*. The haematic level of the Prostate-Specific Antigen (PSA) is also typically measured since a high level of that enzymatic protein in blood represents the first sign of prostate cancer. The values obtained are stored in *PSA Measurement*. If further examinations are needed, the patient undergoes an *MR Study*. The *Image-based Measurements* entity collects measurements resulting from the study, such as the prostate dimensions, its volume, and the presence of lymphadenopathy. Each study comprises different sequence modalities (T2-weighted, DWI, and Dynamic Contrast Enhanced) stored in the *Image Stack* entity. The diagram then allows for collecting information for each *Lesion* identified from those images. Suspicious lesions are then analyzed via histology, thus obtaining the corresponding *Histological Data*. Due to physical and protocol limitations, clinicians may perform a biopsy of a different *Zone* than the one suspected by imaging. The patient may then undergo further investigations through *Other Imaging Exams*, such as CT and PET, to check for metastasis presence. Based on the previous analyses, a *Treatment* is then assigned to the patient, which may include radiotherapy, radical prostatectomy, or minimally invasive surgery.

C. Gastric Cancer Data Model

The gastric neoplasm is usually assessed in n the so-called Staging phase, via imaging exam. After Staging, if a radical intervention is performed, the next diagnostic moment for the patient is called Follow-up, and recurrence/relapse is monitored. In this case, the tumor evaluation is assessed on the removed corpus (pathological evaluation). On the other hand, if after Staging the patient undergoes chemotherapy or a non-radical intervention (a residual disease is left), the following diagnostic moment for the patient is called Re-staging. These three phases were fundamental to develop the ERD for gastric cancer (Fig. 3), and therefore we took it into account in the *Stadiation Phase* entity. Our ERD models the clinical practice according to which, depending on the phase (*Staging, Re-staging, Follow-up*), the patient undergoes specific examinations. Our diagram also mimics the scenario where some examinations may be performed at different phases. For instance, *Surgery* is accessible in Re-staging/Follow-up, *Neoadjuvant Therapy* only concerns Re-staging, while *Endoscopy* can be performed at all phases. That is also true when a patient undergoes a *CT Study*: all patients have a Staging CT, then, based on the evolution of their phase, they can have possibly one to three *Re-staging CTs* and possibly a *Follow-up CT*.

IV. Discussion and Conclusion

In this work, we have presented our early realization of the data models for the three cancer use-cases of the NAVIGATOR Project. First, we performed an extensive requirements analysis to investigate the needs of the clinical partners. Based on that, given the peculiarities of each use case, we decided to implement three separate ERDs. As can be seen by comparing Figs. 1-3, the data models for colorectal and prostate cancers shared a similar structure. Instead, the design of the gastric cancer model showed a higher level of complexity.

This is the preliminary step necessary to develop the NAVIGATOR imaging bio-bank. The implementation of the corresponding database is currently running and is taking into account existing standardization initiatives, such as Molgenis and OMOP-CDM. Indeed, a common data model would standardize the logical infrastructure of data models and make the bio-bank easily interoperable with other bio-banks. Methods to access and operate the data would be shared so that all applications could use the same standardized procedures.

Acknowledgment

The authors would like to thank the whole NAVIGATOR consortium.

References

[1] M. N. Fransson, E. Rial-Sebbag, M. Brochhausen, and J.-E. Litton, “Toward a common language for biobanking,” *European Journal of Human Genetics*, vol. 23, no. 1, pp. 22–28, 2015.

[2] M. Asslaber and K. Zatloukal, “Biobanks: transnational, european and global networks,” *Briefings in functional genomics and proteomics*, vol. 6, no. 3, pp. 193–201, 2007.

[3] L. Coppola, A. Cianflone, A. M. Grimaldi, and M. Incoronato, “Biobanking in health care: evolution and future directions,” *Journal of translational medicine*, vol. 17, no. 1, pp. 1–18, 2019.

[4] Y. Kumar, K. Sood, S. Kaul, and R. Vasuja, “Big data analytics and its benefits in healthcare,” in *Big data analytics in healthc*. Springer, 2020, pp. 3–21.

[5] L. Martí-Bonmatí, E. Ruiz-Martínez, A. Ten, and A. Alberich-Bayarri, “How to integrate quantitative information into imaging reports for oncologic patients.” *Radiologia*, vol. 60, pp. 43–52, 2018.

[6] E. Neri and D. Regge, “Imaging biobanks in oncology: European perspective,” *Future Oncology*, vol. 13, no. 5, pp. 433–441, 2017.

[7] “Navigator: An imaging biobank to precisely prevent and predict cancer.” [Online]. Available: http://navigator.med.unip

[8] “Omop common data model.” [Online]. Available: https://www.ohdsi.org/data-standardization/the-common-data-model/

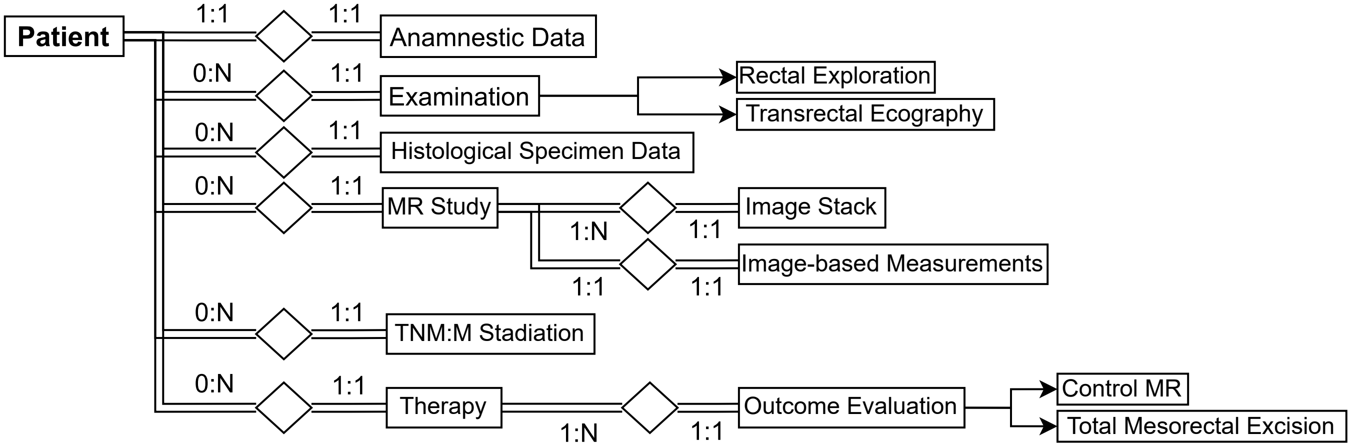

Fig. 1. Entity-Relationships Diagram for the Colorectal Cancer Data Model.

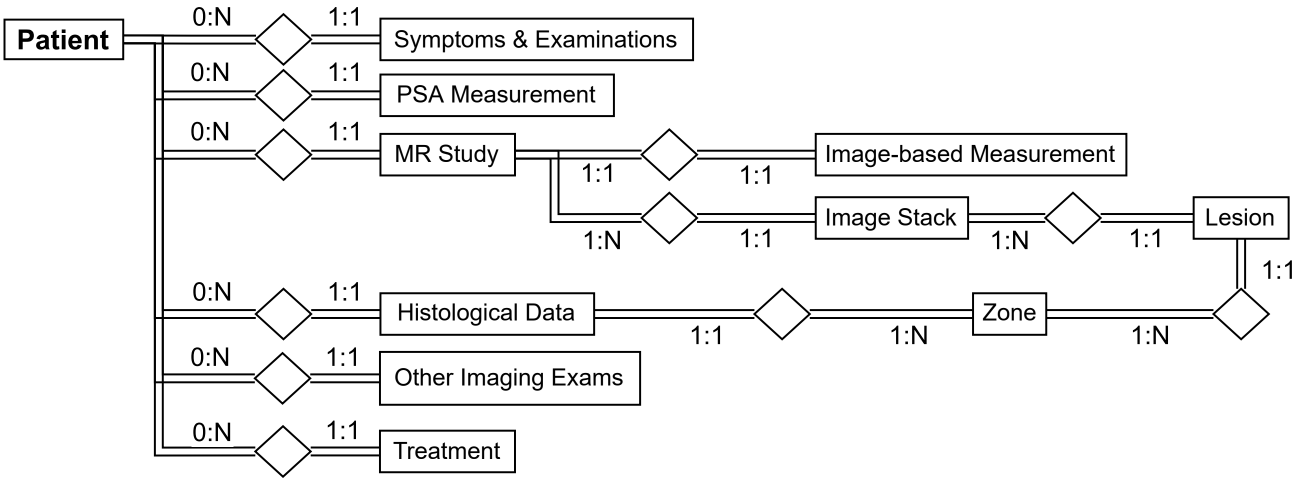

Fig. 2. Entity-Relationships Diagram for the Prostate Cancer Data Model.

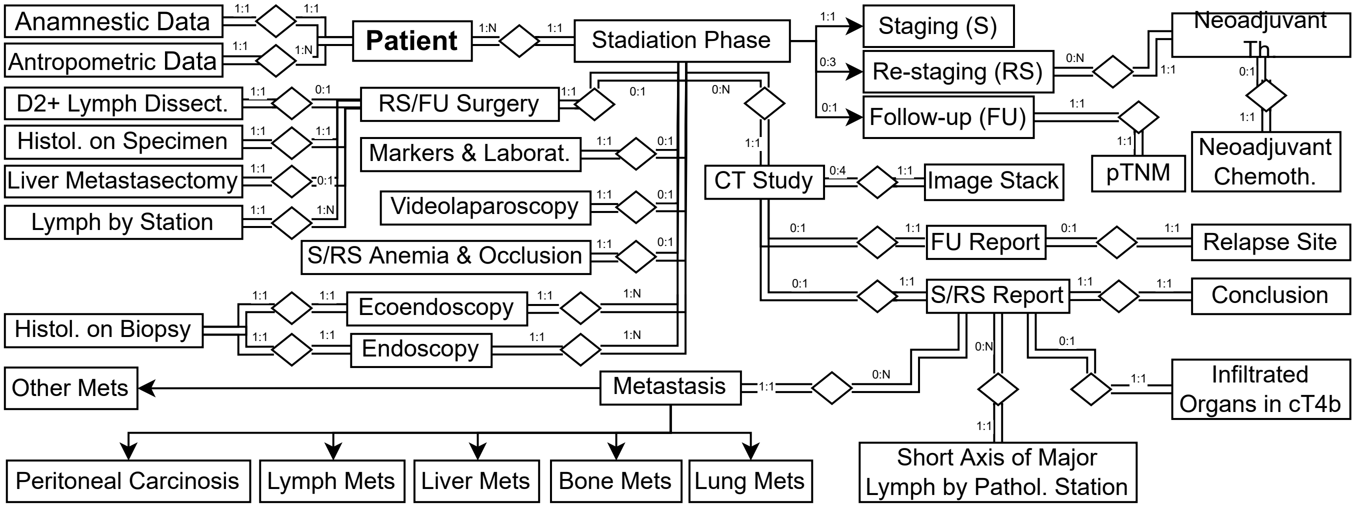

Fig. 3. Entity-Relationships Diagram for the Gastric Cancer Data Model.