FAIR quantitative imaging in oncology: how Semantic Web and Ontologies will support reproducible science

A. Traverso[1], Z. Shi[1], L. Wee[1], A. Dekker[1]

[1]Department of Radiation Oncology (MAASTRO), GROW – School for Oncology and Development Biology, Maastricht University Medical Center, Maastricht, the Netherlands

Abstract. The automated extraction of quantitative imaging biomarkers from patient's scans, could augment physician decision making in radiation oncology. Unfortunately, lack of reproducibility and robust methodology current limits this promising field to be applied in the clinic. In this paper, we state how the combination of quantitative medical imaging with Semantic Web and Ontologies techniques could speed up the role of quantitative imaging.

Keywords: Ontologies, Semantic Web, Quantitative Imaging, Radiation Oncology.

1 Introduction

1.1 A new era of medical imaging: from images to big data

Medical imaging has expanded its fundamental role in radiation oncology since the advent of the first Computed Tomography (CT) scans in the 70s, followed by PET (Positron Emission Tomography) and MRI (Magnetic Resonance Imaging). Radiological examination has moved from purely descriptive to semi-quantitative and fully automated analysis. In the recent years, the availability of enterprise digital imaging and the overflowing role of AI (Artificial Intelligence, like Machine Learning) domain (e.g. machine learning) led to the development of many quantitative imaging models aimed at assisting and augmenting physician decision-making. The term "radiomics" was first created in 2012 and it describes the process of advanced quantitative clinical imaging analysis in medicine. The hypothesis behind radiomics is that tumor biological properties, often obtained by invasive techniques such as tissue biopsies, can be measured in a non-invasive fashion via extracting image-based descriptors (referred as 'features') from medical images [1].
After 2012, the number of radiomics computational packages has increased [2]. However, no consensus has been reached: a) on the optimal configuration that should be used to extract these features for a problem; b) about the robustness of radiomics features when evaluated in different contexts. Therefore, most of the users simultaneously extract features using different parameters, leading to an increase of the number of features. Typical radiomic studies often extract from 500 to 10000 features while starting only from 100 unique features [3] . We are now facing the same

"data explosion" defined by Rubin about multi-detector row CT scanners. One main difference divides the two processes: if the CT data explosion was mainly driven by an advance in hardware development, producing more images faster than expected; the new quantitative imaging data explosion is driven by automated imaging analysis computational pipelines that produce a large amount of processed data (e.g. radiomic features) from medical images. This data seems mimicking all the attributes of big data: a) volume: the large amount of data to be processed and analyzed via machine learning requires now dedicated computational power and powerful machine learning able to deal with a large hyperspace of parameters; b) velocity: new data are generated faster as soon as new computational radiomics software become available, with a larger hyperspace of parameters that can be tuned for features extraction; c) variety: not only singe features should be stored in quantitative imaging, but also information about the original source (image, region of interest, computational details) making the data variety larger; d) veracity: in the hyperspace determined by features and associated metadata, some information could be redundant and only meaningful one should be extrapolated [3]. For all the above-mentioned reasons, quantitative imaging strictly connects to the world of big data. We believe that extending the usage of ontologies and Semantic Web technologies to quantitative imaging could help solving some of the issues that would be presented in the next paragraph and further speed up the adoption and acceptance of new image based quantitative biomarkers in the clinic.

1.2 Reproducibility crisis in quantitative imaging

Still a strong unbalance exists between published radiomics-based prediction models and their real usage as decision support systems in the clinic [4].
The lack of reproducibility and transparency in radiomics is the major slowdown of its applicability in the clinic [5]. The lack of reproducibility mainly relates to the fact that most radiomics-based models are built on limited-datasets and often validated in one single institution, with no guarantee of generalizability power when applied to multiple centers. This evidence also seems colliding with recommendations from the TRIPOD (Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis), suggesting and encouraging TRIPOD IV-type models, which are fully validated on completely independent external datasets [5]. TRIPOD IV models are based on the possibility for an external user to fully reproduce and validate a previously developed model. Unfortunately, this reproducibility crisis reflects not only on the difficulty for external users to fully reproduce a radiomics experiments developed in another institution, but also within the same institution.
This issue mainly connects to the previously mentioned concept of lack of transparency. In absence of a standardized and structured way of describing radiomics studies, most of them only report single feature names or values, with no further details on how the model was developed, how the features were computed and which where the computational parameters used (metadata). Even in presence of publications that made available software and datasets, re-usability and inter-operability remain issues. It is not unlikely

that two software could call a radiomic feature with the same name but meaning a totally different quantitative descriptor. On the other hand, two features could express the same quantitative descriptor but show different values when computed with different software. Without then associated metadata, it is impossible to find the reasons behind this discrepancy, which probably lie in a different choice of hyperparameters.
It becomes then clear that quantitative imaging is far behind the FAIR principles that are taking the scene in clinical data science as incentive for reproducible and transparent science [6]. However, the absence of FAIR guiding principles represents a unique opportunity for the imaging community to propose a new paradigm for a new era reproducible quantitative imaging. We believe that ontologies and Semantic Web techniques should guide this effort toward reproducible, transparent quantitative imaging. On the other side, the imaging community needs to accept the challenge to work closely with the data science community and re-use as much as possible available tools. A possible framework and the ongoing actions taken by our group are presented in the following paragraph.

2. Proposed solution

2.1 Ontologies for quantitative imaging: a dynamic body of knowledge to enhance consensus

Ontologies represent a formal specification of the terms related to a specific domain and the relations among them [7]. In this specific case, an ontology for quantitative imaging should mimic the workflow that happens during a radiomic study: from image pre-processing, region of interest definition, computational settings definition and finally features extraction, as presented in [6]. Therefore, the ontology not only should include the main radiomic features and their corresponding units, but also all the metadata that relate to the above-mentioned workflow. In this view, building this ontology is a joint exercise between imaging research groups to represent the state of the art of the knowledge related to the quantitative imaging domain. The ontology acts as harmonizer and standardizer, eliminating barriers related to different nomenclature or labels. In fact, each concept in the ontology is universally defined and the whole community agrees on its meaning. For example, the ontology universally defines the radiomics features by describing them and associating a unique identifier and their provenance. In this view, it enhances consensus and creates a shared knowledge domain. It represents a dynamic body of knowledge that can be expanded with new concepts as the quantitative imaging field evolved (for example by introducing and defining new imaging features or computational methods). Our group took the lead in developing an extensive radiomics ontology (RO), released on the BioPortal (https://bioportal.bioontology.org/ontologies/RO) as door-opener for FAIR quantitative imaging. Recently, we published a modular python tool for making radiomics computations FAIR [8]. Finally, ontologies express concepts in a machine-readable language and therefore, when data and metadata are transformed via the ontology, they can be automatically parsed by

machines. This becomes of fundamental utility when comparing results computed from different software or under different conditions. If each radiomics computational package is setup to produce ontologies-labelled data and metadata, then automated meta analyzes can be performed and this will open the path to data-driven standardization and harmonization. A summary of the concept behind the RO and possible applications is depicted in Figure 1.

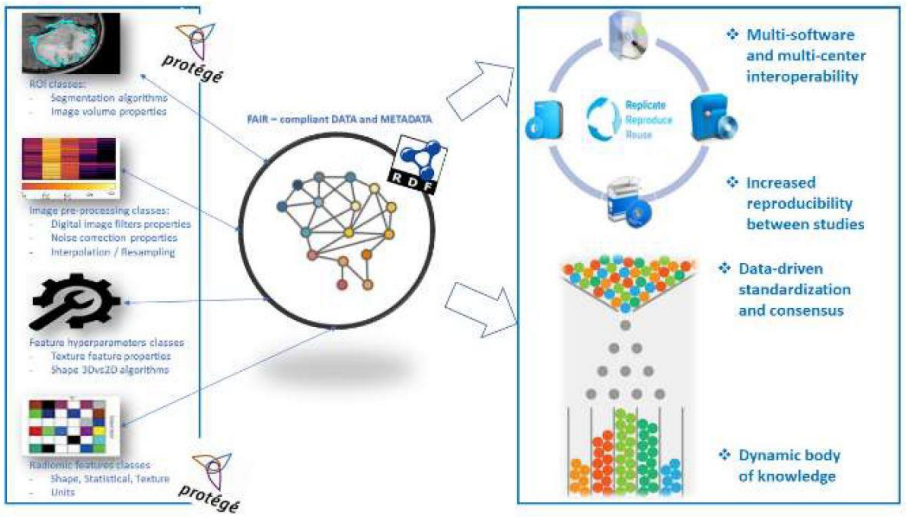

Figure1: the RO workflow. Not only standardization of the radiomics workflow is achieved, but the same instrument can be used to enhance the reproducibility and validation of radiomics-based prediction studies

2.1 Semantic Web: linking quantitative imaging with multiple domains

Semantic Web has the power to extract knowledge from data labelled via ontologies, using dedicated SPARQL language.
If radiomics data and metadata are transformed via the Radiomics Ontology and published on the Semantic Web, then they can be queried using the universal concepts defined by the ontology, without any prior knowledge on the original labels present in the original software. Also, the combination of ontologies and Semantic Web techniques allows parsing and joining data and metadata from multiple sources, such as different databases. For example, in a typical radiomics-based prediction study it could be interesting to query a) the value of a certain feature b) computed on an imaging modality c) referring to a patient with a certain disease; d) finding patients with similar feature values but different clinical outcomes for comparison. As it is clear from this example, that type of query requires merging radiomics data (a); DICOM metadata (b); clinical data (c), and data from other clinics (d). Sooner, additional sources of data

such as for example genomics data or pathology data, for better predictions and for exploring connections with medical images will be needed. Our group has developed a portfolio of ontologies for guaranteeing the road to FAIR compliant and transparent prediction models in radiation oncology: the ROO (Radiation Oncology Ontology) [7], the SEDI (Semantic DICOM Ontology) [8] and the presented RO.

We successfully showed how this workflow can be used in combination with Semantic Web for winning barriers related to data sharing and build more accurate models (distributed learning) [9]. For example, we successfully reproduced a classical centralized radiomics study [10] in a distributed fashion using the above-mentioned ontologies combined with Semantic Web [8]. By using only SPARQL queries we could retrieve the model and computational details of the model trained at one local institution and externally validated on the second one.

We believe the upcoming effort should focus on developing additional ontologies that could link the quantitative imaging domain with data from multiple sources presented above.

Finally, we state that ontologies and Semantic Web are the key for speeding up reproducible science. Therefore, the quantitative imaging community should work closely with experts from the semantics, FAIR and data science fields to provide a sustainable infrastructure for medical imaging and derived big data.

References

[1] R. J. Gillies, P. E. Kinahan, and H. Hricak, 'Radiomics: Images Are More than Pictures, They Are Data', *Radiology*, vol. 278, no. 2, pp. 563–577, Feb. 2016.

[2] L. E. Court, X. Fave, D. Mackin, J. Lee, J. Yang, and L. Zhang, 'Computational resources for radiomics', *Transl. Cancer Res.*, vol. 5, no. 4, pp. 340–348, Aug. 2016.

[3] V. Kumar *et al.*, 'Radiomics: the process and the challenges', *Magnetic Resonance Imaging*, vol. 30, no. 9, pp. 1234–1248, Nov. 2012.

[4] I. Buvat and F. Orlhac, 'The Dark Side of Radiomics: On the Paramount Importance of Publishing Negative Results', *J Nucl Med*, vol. 60, no. 11, pp. 1543–1544, Nov. 2019.

[5] A. Traverso, L. Wee, A. Dekker, and R. Gillies, 'Repeatability and Reproducibility of Radiomic Features: A Systematic Review', *International Journal of Radiation Oncology*Biology*Physics*, vol. 102, no. 4, pp. 1143–1158, Nov. 2018.

[6] M. D. Wilkinson *et al.*, 'The FAIR Guiding Principles for scientific data management and stewardship', *Scientific Data*, vol. 3, p. 160018, Mar. 2016.

[7] 'Ontologies', in *Ontology Learning and Population from Text*, Springer US, 2006, pp. 9–17.

[8] Z. Shi *et al.*, 'Distributed radiomics as a signature validation study using the Personal Health Train infrastructure', *Sci Data*, vol. 6, no. 1, p. 218, Dec. 2019.