Radiogenomics: what it is and why it is important

Maciej A. Mazurowski, Ph.D.

Department of Radiology, Duke University School of Medicine
Duke Cancer Institute
Duke Medical Physics Program, 
Duke University
2424 Erwin Rd, Suite 302 
Durham, NC 27705 
Tel: +1 919 684 1466 
e-mail: maciej.mazurowski@duke.edu

Abstract

In recent years a new direction in cancer research has emerged that focuses on the relationship between imaging phenotypes and genomics. This direction is referred to as radiogenomics or imaging genomics. The question that subsequently arises is: what is the practical significance of elucidating this relationship in regards to improving cancer patient outcomes? In this article I try to answer this question. While I discuss some limitations of the radiogenomic approach and describe scenarios in which radiogenomic analysis might not be the best choice, I also argue that radiogenomics will play a significant practical role in cancer research. Specifically, I argue that the significance of radiogenomics is largely related to practical limitations of currently available data that often lack complete characterization of the patients and poor integration of individual datasets. Radiogenomics offers a practical way to leverage limited and incomplete data to generate knowledge that might lead to improved decision making in oncology.
Introduction to ‘radiogenomics’

Increasingly, even casual readers of the scientific literature are encountering such terms as radiogenomics, imaging genomics, and radiomics. Because of their recent introduction into common usage, their usage and definitions are still in flux.

‘Radiogenomics,’ in particular, has been inconsistently used to refer to different endeavors or research topics, all of them related to cancer.

The most frequent use of the word ‘radiogenomics’ [1], [2] is to refer to the relationship between the imaging characteristics of a disease (a.k.a. imaging phenotype or radiophenotype) with its gene expression patterns, gene mutations, and other genome-related characteristics. As a simplification, I will refer to them collectively as “genomic characteristics” or simply “genomics”. A particular focus of radiogenomic analysis has been on the relationship between imaging phenotypes and gene expression patterns which include expressions of individual genes as well as measures that summarize expressions of specific gene subsets (e.g. tumor molecular subtype, or Oncotype DX). ‘Radiogenomics’ also refers to a research effort aimed at finding this relationship. Another term used to refer to this kind of research is imaging genomics. Another, also very common use of the term ‘radiogenomics’ is to refer to the analysis that looks for associations between patient genetics and his/her reaction to radiation therapy [3], with a focus on radiation toxicity. As opposed to an effort to match imaging phenotype and genomic characteristics, this genre of research focused on phenotypes representing radiation toxicity [3]. In 2009, the Radiogenomics Consortium was established in the United Kingdom [4] in relation to this research.

Finally, ‘radiogenomics’ has been equated with another approach called ‘radiomics’ [5], [6], [7]. However, rather than describing a particular relationship of interest, radiomics focuses on the methodology used in the analysis. Specifically, the radiomics paradigm proposes extraction of a large number of quantitative features from images using computer algorithms. The extracted features can then be related to other data of interest, including patient outcomes. These features can also be related to genomic characteristics and such a pursuit could be referred to as the ‘radiomics approach to radiogenomics.’

State of the art in brief

The body of literature on radiogenomics is still limited, but a rapidly increasing number of papers are appearing in relation to brain cancer with a focus on glioblastoma, breast cancer, lung cancer and others. Since the objective of this paper is mainly to discuss the issue of significance of radiogenomic research, this review of the literature will not be exhaustive but will rather allow the reader to judge the general level of advancement in radiogenomics research.

In glioblastoma (GBM), Zinn et al. [8] showed that an upregulated PERIOSITIN gene is associated with a high tumor volume in FLAIR MRI exams. Jamshidi et al [9] showed that specific molecular phenotypes
correlate with some imaging traits in GBM. Further evidence of the potential association between molecular phenotypes and imaging can be found in [10] and [11].

In breast cancer, Yamamoto et al. [12] showed the potential for an association between imaging and genomics with a small sample of 10 patients. This was followed up by discovering a relationship between semi-automatically extracted imaging features describing MRI enhancement dynamics with Luminal A and Luminal B subtypes [13], [14] and Oncotype DX [15], [16]. Semi-automatic feature extraction involves both a human reader and a computer algorithm.

In lung cancer Gevaert et al. [17] showed a correlation between molecular phenotypes and some imaging traits in lung computed tomography (CT). Radiogenomic analysis has also been applied to hepatocellular carcinoma [18] and clear cell renal cell carcinoma [19].

The typical research methodology of the studies above involves manual or semi-automatic assessment of imaging features and their correlation with individual gene expressions, combined gene expression patterns such as previously defined genomic subtypes, or other molecular phenotypes. The currently available studies are typically characterized by smaller sample sizes (most often < 100) which limit the conclusions that can be drawn.

**Why should we care?**

Radiogenomics attempts to establish and examine the relationship between tumor genomic characteristics and their radiological appearance. While there is certainly a lot to learn about these relationships, one could ask: what is the practical significance of radiogenomic discoveries? From the perspective of the patients, cancer patients in this case, it is their outcomes that are of interest, such as survival, time to recurrence, or response to a particular treatment. A question appears: If imaging data and particularly specific features extracted from the images are available along with the outcome of interest, why not simply build a model that relates the imaging features to the outcomes directly? Relation between some imaging features and outcomes is already established and utilized in treatment planning. What is the benefit of including genomics in the mix?

One could even argue that using genomic characteristics as the intermediate step in the analysis could damage the potential that imaging has for predicting patient outcomes. However, current models relating molecular phenotypes to outcomes and usage of different therapeutic regimens are highly imperfect. These models often show only minor differences in outcomes for different molecular phenotypes (e.g. for different molecular subtypes) and therefore provide limited prognostic/predictive values. Radiogenomic models relating imaging data to genomics, especially now, in their early days, are also capturing fairly weak or noisy relationships. When the tenuous imaging to genomics and genomics to outcomes relationships are combined to establish an imaging to outcomes relationship, the resulting link might be very weak or nonexistent. The second reason for relating imaging features directly to outcomes instead of including genomic characteristics in the middle is that imaging phenotypes potentially contain information that is not available in genomics data. For example, gene expression
patterns are typically assessed based on a relatively small tumor tissue sample or are “averaged” for tissue samples from multiple regions in the tumor and therefore might not reflect the usual heterogeneity of cancerous tumors [20]. Imaging on the other hand can potentially capture this heterogeneity [21]. Constructing a radiogenomic model first and then applying it to predict outcomes without training it using the imaging-outcomes data limits the information from imaging available to predict outcomes to what is already contained within tumor genomics. The imaging information that complements genomics is not used in such a scenario. To utilize such complementary information, a model directly relating imaging to outcomes is needed.

These are limitations of radiogenomics. However, this does not at all mean that radiogenomic analysis is without use. I will argue that the significance of discoveries in radiogenomics is largely related to a very practical aspect of science: availability of data and availability of knowledge.

Well organized and publically available data repositories containing molecular data are commonly available. More recently, in support of developing radiogenomics, efforts have been undertaken to assemble large cancer imaging datasets (e.g. The Cancer Imaging Archive [22]). On the other hand, routine imaging data, unlike molecular phenotype data, is readily available in large quantities in patient records and can be relatively easily and cheaply collected retrospectively by investigators working at large clinical institutions (although sharing such data poses difficulties). Finally, while collecting patient outcomes data is relatively easy, in order to determine clinical significance, very long follow up periods are required. This means that usable outcomes data might not be available. I.

As a result of prior and current data collection efforts, various datasets, private and public [22,23], are available containing different combinations of imaging, genomics, and outcomes data (often just one or two components). The quality of data in each component might also differ dramatically between different datasets.

Radiogenomics allows leveraging imperfect data along with prior knowledge of the relationship between outcomes with either imaging or genomics to draw new conclusions. One example: Prior research demonstrates that a specific genomic characteristic related to poorer survival. A research group has a dataset available that contains genome-wide gene expression data and imaging data but no outcomes data. The research group conducts an analysis that establishes the relationship between some imaging features and the gene expression data by building a model that represents the specific genomic characteristic (model output) in terms of imaging features (model input). This analysis also determines which imaging features are the most predictive of the genomic characteristic. This is significant because it allows for identifying imaging features (or even defining new features) correlated with the genomic characteristic previously shown to be related to outcomes. These individual imaging features are likely to be correlated with outcomes too and they are good candidates for further analysis. Moreover, the constructed model that predicts the genomic characteristic using multiple imaging features, can be directly applied to predict outcomes. Inversely, when a relationship between a particular imaging signature and outcomes is known, finding a correlation between that signature and genomics may identify specific genomic characteristics that are associated with outcomes.
One example of this approach is [24] where the authors apply radiogenomic analysis to clear renal cell carcinoma. In this paper, the authors identified a number of imaging features which they correlated to mutations in VHS, PBRM1, SETD2, KDM5C, and BAP1 genes that were previously indicated in relation to clinically significant factors of advanced grade, stage, and diminished survival prognosis. The authors found correlations between some of the imaging features and the gene mutations and through this discovery identified imaging features that are potentially predictive of outcomes.

Another example is our own studies [13,14], where we found a correlation between computer-extracted imaging features and intrinsic tumor subtypes in breast cancer. Tumor molecular subtypes are defined using a set of tumor gene expression data and include Luminal A, Luminal B, Basal, and HER2 type. We were able to demonstrate that computer-assessed features describing tumor enhancement dynamics can distinguish Luminal B from other subtypes. The potential practical application of this result is the identification of Luminal A or Luminal B patients using the imaging surrogate and adapting therapy accordingly. Furthermore, this result suggests that tumor enhancement dynamics in MRI might be predictive of outcomes.

Another novel way of using radiogenomic analysis to address the issue of limited data can be found in a recent publication by Gevaert et al. [25]. In this paper, the authors had two non-small cell lung cancer datasets available: dataset 1 with imaging and gene expression data (no outcomes) and dataset 2 with genomic and outcomes data (no imaging). The authors constructed models that map specific genomic characteristics (model input) into imaging features (model output). Multiple models were constructed for the imaging features. Then, using dataset 2, the authors tested the predictions of imaging features based on specific genomic characteristics (i.e. surrogate imaging features) in terms of how well they predict patient survival. The goal of the analysis, as stated by the authors, was to identify imaging features that are predictive of survival by evaluating their gene-based surrogates. They showed that the proposed approach identified imaging features that were related to outcomes.

Finally, beyond filling the gaps in knowledge, radiogenomics discoveries have a more basic significance of building a better understanding of the imaging representations of various molecular phenotypes, uncovering biological processes that are underlying phenotypes seen in imaging [26] (i.e. causal relationship between the two), which could drive future discoveries in cancer research.

**How to extract imaging features?**

In radiogenomic analysis, investigators assess the association between imaging and genomic data. In order to conduct the analysis, specific features have to be extracted from the images. This can be done by radiologists (or other qualified individuals) or with the assistance of a computer program.

In early research, the imaging features are typically extracted by human readers, sometimes with some minor assistance from a computer program [11], [24]. However, the need for reproducibility of the results beyond a single experiment, depends the use of well defined lexicons to guide extraction. A lexicon contains a set of terms (features) that describe the tumor and its surroundings along with
An established example of a radiology lexicon is the Breast Imaging-Reporting and Data System (BI-RADS) by the Academic College of Radiology. For example, the lexicon contains descriptions of mammographic masses such as “mass margin” with five possible values of circumscribed, microlobulated, obscured, indistinct, and spiculated. Limiting the readers to choosing one of the previously specified answers allows comparability of the features among different readers and different cases. Another example of a radiology lexicon, developed specifically in the context of radiogenomics and imaging biomarker development is the VASARI (http://cabig.cancer.gov/action/collaborations/vasari/) lexicon for the annotation of brain tumors. Specific terms include but are not limited to major axis length, proportion of the enhancing tumor, deep white matter invasion. The development of lexicons in radiology has recently gained notable interest and significant work in the discipline of ontology [27]. The goal is to develop lexicons like RadLex [28] that will lend themselves naturally to radiogenomic analysis.

While radiology lexicons facilitate the extraction of easily usable features for radiogenomic analysis, manual analysis of images has some disadvantages, most significantly, interobserver variability. While the lexicon terms can be well defined, there are still often notable differences between the assessments of individual features by different radiologists. This introduces noise to radiogenomic analysis or simply means that a particular correlation of imaging features with genomic characteristics does not hold when different radiologists annotate images. Related to interobserver variability is the lack of precision when human observers conduct quantitative assessment of features such as mass volume, or volume of a specific part of a tumor. Another disadvantage is the significant time commitment required to assess individual cases. For example, the VASARI lexicon contains more than 20 features, many of which require inspection of more than one MRI sequence. This means that a skilled individual will have to expand a good deal of time analyzing an imaging examination.

Automatic or semi-automatic feature extraction alleviates some of the issues listed above. In automatic segmentation, computer vision algorithms are used to automatically segment the abnormality and extract a variety of features. In semi-automatic segmentation, a large part of the segmentation and feature extraction is still carried out by a computer, but a human reader is involved in the process. An example is when a reader indicates roughly the boundaries of the tumor and a computer algorithm completes segmentation and feature extraction. Involvement of computer algorithms in the process allows for a more precise, more consistent assessment of quantitative features and extraction of a larger number of features, some of which might not be easily perceptible by a human eye. While embraced in the field of radiomics, this approach to extraction of imaging features for radiogenomic analysis has not reached its full potential.

Conclusions and discussion

The main limitation of radiogenomics is related to the fact that if strong imaging and outcomes data are available, and the prediction of outcomes is the primary goal, the radiogenomic analysis may not bring an additional contribution to the analysis. In such cases, imaging features can be directly correlated with the outcome of interest. However, in the most common situation, where the data is limited,
radiogenomic analysis can be used, along with previously generated knowledge to identify imaging features that might be related to outcomes. Furthermore, elucidating the relationship between imaging and genomics will lead to a better understanding of cancer in general and may lead to improved treatment protocols. Finally, radiogenomic findings could guide the collection of future datasets by identifying imaging or genomics with a high potential to predict outcomes.

In this article I focused on radiogenomic analysis directed at finding a relationship between genomics and imaging. However, a broader understanding of the term is sometimes applied to include any analysis that involves genomics and imaging. Various research directions within this broader theme and related to it are emerging and gaining popularity, some of which with very high significance. A very promising research direction is using computer- and radiologist-extracted features to directly predict patient outcomes [11], [29], [30]. Another involves combining imaging and genomic markers. Still another is traditional imaging-based surveillance of patients with higher disease risk determined by their genotype, for example higher breast cancer risk determined by BRCA1 or BRCA2 mutations.

A significant opportunity related to radiogenomics lies in the development of new statistical methodologies for the construction and evaluation of radiogenomic models. New methodology may instruct us on how to better leverage radiogenomic models in the development of clinically useful tools. As the discipline of radiogenomics matures, the significance of radiogenomic analysis will become clearer and new applications will undoubtedly emerge.

Conclusions in bullet points:

- Radiogenomics investigates the relationship between disease genomic characteristics and its radiology phenotypes
- In some scenarios, direct prediction of outcomes using imaging features might be preferable to radiogenomic analysis
- Radiogenomics will play a significant role in cancer research as it creates a new avenue of generating important knowledge from limited data

Acknowledgments

The author would like to acknowledge Dr. David Haynor for a helpful discussion, Drs. Bruce Hillman and Lars Grimm for their help in editing this manuscript as well as two anonymous reviewers for their helpful suggestions.
REFERENCES:


