Abstract

This chapter introduces themes related to modular software design, bioimage processing, Computer Vision, and image-annotation, that we will return to in subsequent chapters. Our goal here is to conduct a fairly broad-ranging survey of image-processing and radiomics technologies in contexts such as cardiac care and immuno-oncology, so as to present an empirical background as a precursor to discussing image-annotations and radiomics in a more theoretical vein. We also consider a few examples of computational simulations applied to disease and prognostic models (such as tumor growth) and outline how cellular or tumor-scale simulations can be double-checked via cross-reference against image and clinical data. In general, we explore the digital logistics of integrating bioimages, simulations, and clinical data, examining software-development practices in fields such as Systems Biology which often unify these distinct forms of data into multi-disciplinary models. We conclude by arguing for a specific form of modular software design as a useful paradigm for implementing computer code appropriate for these sorts of interdisciplinary syntheses.

Chapter 4: Modular Design, Image Biomarkers, and Radiomics

1 Introduction

In this chapter, and again in Chapter 8, we will consider image-annotations as a case-study in the methodology we will describe as "multi-aspect modular" design. Modular design in general attempts to realize larger-scale projects by subdividing them into relatively autonomous smaller parts. We propose the term "multi-aspect modules" to describe software components which are not monolithic applications — they are intended to be mixed with other modules to create full applications — but which offer integrated functionality in a self-contained fashion. For example, each module would autonomously implement GUI classes, database access, data serialization, and scripting/runtime-reflection features.

Image-annotations furnish a good example of how concerns related to GUIs, serialization, data persistence, and procedural models are inter-connected. Bioimage annotations are also a useful starting-point from which to consider how multi-aspect design might be adopted for other branches of bioinformatics and biomedical software engineering, because bioimaging data is, intrinsically, conceptually linked to other kinds of clinical/diagnostic information. We assume, then, that hypothetical image-annotation modules are deployed in conjunction with different bioinformatic modules as part of an overarching clinical, diagnostic, or research application. For example, operations associated with those other modules could potentially be requested from within GUI objects managed by the annotation module.

2 Image Biomarkers (and Others) for Cardiac and Oncology Diagnostics

From one perspective, there is nothing special about image-annotations such that GUI objects within that domain would be a distinguished starting point for bioinformatic operations in general — in principle any window in a biomedical application could potentially be the starting-point for user actions that lead to capabilities realized across two or more modules. On the other hand, bioimaging is an important case-study for medical *software* because in this context the entire data-acquisition chain occurs "*in silico*." Unlike tests based on lab equipment, for example (e.g., colorimetric assays) or clinical evaluations which rely on practitioners' subjective judgments, every step in the image-analysis process is performed via computer code that itself can be evaluated, and all data generated during diagnostic/predictive processes can (at least in principle) be preserved (whereas in a typical blood workup, say, samples are discarded once the experiment is concluded). Moreover, analyses can be exactly repeated by running the same computer code against the same images.

It is also true that because image analyses are already in digital form, there is no extra data-entry step needed, nor ambiguities arising from imposing numerical measures on subjective evaluations (cf. the Medical Research Council scale for muscle-strength we mentioned in Chapter 3). Applying Computer Vision algorithms can yield large quantities of numeric data from image-analysis, and while not all feature-signatures will have obvious biomedical interpretations, the sheer volume of raw data can be beneficial for Machine Learning, which will sometimes discern statistical correlations that impugn diagnostic or prognostic value to mathematical radiomic entities which may be too subtle for the naked eye. The value of such *a priori* quantification is advanced further, too, by conventions which codify imaging results, such as **RadLex** (for radiology) [41], [61] or, more recently, IBSI, the Image Biomarker Standardization Initiative (centered on radiomics) [93], [94], [3], [37], [27].

These are among the factors that drive the search for reliable image biomarkers, envisioned as essential component in biomedical research/treatment ecosystems. Other analogous factors include how image-based diagnostics can be quicker and cheaper than lab-based alternatives. More broadly, imaging technlogies preserve a more holistic data space than (say) biopsies, which by definition excise biologic material from its *in vivo* context. Images preserve a record of spatial relationships (or spatio-temporal relationships, if we consider 4D media such as Flow MRIs) which is lacking from data obtained via gene sequences, blood, or tissue samples. As expressed in an "Assessment of Imaging Informatics for Precision Medicine in Cancer," for instance:

> Educating clinicians on the benefits of imaging methods in clinical practice is key to their adoption [because] in many cases ... genomic analysis alone is sometimes inadequate. Genomic analysis will not reveal carcinoma versus benign growth and mutations analyses alone cannot provide a specific diagnosis. [I]n the case of Leiomyoma (benign disease) vs. Leiomyosarcoma (cancer), the genetic mutation is the same, but human cognition and the use of microscopes are required to accurately diagnose cancer versus benign growth ... Spatial phenotypic heterogeneity is not captured by genomic data. There is no way of understanding interactions between the various cell types in a tumor microenvironment (TME). If the cell composition is the same, but the interactions are different, in two different TMEs, genomics cannot tell them apart. Hence, the study of images, and their spatial data, is crucial. [10, pages 6-7]

This assessment also stresses the importance of sharing image-based data and observations in a consistent manner:

> Integrating radiology, pathology, clinical, and -omics data requires that image annotations be stored in a standardized and interoperable manner. ... Frequently, the annotations, created on commercial image viewing workstations, are collected and stored in

> either proprietary formats or as DICOM presentation state objects, which are like graphical overlay objects. This enables rendering the information visually, but does not support search of, and access to the annotations, nor any computation on them... Consequently, ... there is no interoperability of image annotations across platforms and applications. To realize the potential value of integrative radiology-pathology-omics, it is vital that image annotations be stored in standardized interoperable formats such as the Annotation and Image Markup (AIM) standard or DICOM ... DICOM Working Group 8 is working to harmonize and unify the AIM and DICOM standards and create a DICOM Structured Reporting object to store AIM image annotations [and] provide a standardized interoperable format for image annotations (page 5).

In short, bioimaging (including radiomics) is not likely to entirely replace diagnostics/prognostics via other means, but image biomarkers may well substitute for other kinds of biomarkers in some contexts (as a more cost-effective option, say) or may serve to reinforce or confirm other analyses. In an overview of "radiogenomics," for example, [55] expresses the value of radiomic markers as follows:

> [Whereas] "Radiomics" refers to the high-throughput extraction of quantitative features from images, i.e., conversion of images to mineable data, and subsequently using these data for decision support, including patient outcome ... "Radiogenomics" or "imaging genomics" refers to the study of the associations between radiomic data (imaging features) and genomic patterns. ... Such imaging data and associated radiomics may serve as a "virtual biopsy," which is non-invasive, includes the entire tumor, and is repeatable ... and may yield a quantitative predictive signature for advancing precision medicine. (page 7)

See also [64], which stresses the importance of quantitative imaging as a "decision support" tool for

> this era of personalized medicine in oncology [where] we have a responsibility to collect as much meaningful information from different modalities as possible, which can help to make better informed decisions. ... Quantitative imaging is able to contribute significantly to decision support for 3 major reasons: 1) virtually every patient with cancer is imaged with CT, MRI, and/or PET; 2) these images are obtained from the entire tumor, along with metastases, and thus can be used to describe and classify heterogeneity; and 3) these images can be obtained routinely longitudinally to monitor responses and to guide specific therapies. (page 12)

For these sorts of reasons, and also because image-annotation marks a good case-study for modular design, it is reasonable to consider scenarios where user requests within clinical software originate from a bioimaging context. That is, we will focus on patterns where the specific software-operational sequences under consideration begin with users working within a bioimaging/image-annotation module, and may then proceed to examine other (related) content.

Certainly there are numerous pathways wherein an annotation-related GUI object could be the origin point for actions leading elsewhere. For example, many ground-images presented in the context of bioimage annotations would presumably be associated with an image series taken in a diagnostic context and/or from a specific patient; therefore, consistent with principles of responsive User Interface design, users could plausibly be given the opportunity to request more information about the image series, the diagnostic context, or the relevant patient, from annotation-related visual objects (e.g., context menus activated relative to the ground image). This could potentially lead users to any genre of data related to the patient or diagnosis — information that might in turn encompass a wide range of data types, some of which we will review here.

2.1 Image Registration and Radiomics for Cardiac Care

Our first overview will provide one example of how image analysis generalizes to other bioinformatic domains: we will specifically look at cardiac diagnosis, and in particular image feature-extraction using techniques associated with radiomics, which has been more widely applied to cancer/oncology. The term "radiomics" overlaps with what in more general contexts (not restricted to bioimages) one might call feature-extraction (adopting the "-omics" suffix to suggest parallels with genomics, proteomics, transcriptomics, and so forth [33, page 4]), although *radiomics* specifically tends to be applied in contexts where large feature-vectors are extracted and then statistically analyzed to find diagnostic correlations. In short, radiomic methods are not contingent on *a priori* anticipations that any given image feature would have an established biomedical interpretation (the way that ground-glass opacity on lung scans, for example, is clearly associated with Covid, via well-understood biologic mechanisms).

In the cardiac case, several studies have appeared in recent years which attempted to disentangle correlations between image features and disease expressions, without those correlations being known ahead of time [7], [60], [8], [39]. Concrete results in the cardiac-radiomics literature have tended to focus on image-textures indicating cardiac lesions and scarring (associated with greater risk of adverse events such as heart attack and strokes), but researchers have also mined hundreds of image-features for evidence of a select few that seem definitively correlated with heart disease.

In one analysis based on over 5000 UK Biobank patients, for example, [8] point to "grey level heterogeneity" as a feature associated with diabetes and smoking, and possibly other sources of cardiac damage (page 9). "Median" inten-

sity (i.e., overall image-brightness) was also elevated in diabetic patients, from a statistical point of view. These results suggest that diabetes (and possibly other cardiac risk facts) "leads to a global alteration of the myocardial tissue and thus of the overall myocardial appearance in CMR images" such that bioimages of these tissues register as brighter and less uniform than corresponding images for healthy cohorts. This is an example of imaging patterns for which we can provide a plausible clinical explanation.

Multiple studies have attempted to identify radiomic features that appear to be diagnostically correlated with cardiac damage along these lines. For example, [5] found two signals that were especially strong indicators of myocardial scarring, one involving "autoregression models" for image textures, and one which was histogram-based (we will consider these metrics in slightly more detail momentarily). The UK Biobank analysis also found strong correlations vis-à-vis certain morphological and "morphometric" (as compared to textural) features; for example, healthy cardiac muscle is apparently correlated with the Left Ventricle (LV) taking on a visible elliptical shape. Heart *disease*, conversely, is indicated by "spherical disproportion (i.e., the inverse of sphericity) of the myocardium at end-diastole" (page 8), meaning that the LV being more spherical just before heart-contraction is a sign of tissue damage. Indeed, the authors also report that the Left Ventricle has (according to image-based calculations) relatively less surface area relative to its volume in the presence of decreased cardiac functioning; this intuitively fits the pattern of sphericity, because spheres minimize surface area for their corresponding volume. As the authors suggest, the biologic mechanisms underlying these morphological observations appear to be related to LV hypertrophy: the ventricle becoming enlarged due to exerting greater effort. In general, [8] endorse combining morphological and textural features to develop hybrid image biomarkers strongly predictive of heart risks.

With respect to textural features, several studies have noted the statistical significance of Autoregression (AR) Models. In the context of image-segmentation, these models are based on the technique of calculating pixel-intensities as weighted sums of the intensity of neighboring pixels. In effect, rather than defining pixel color as a free combination of red/green/blue scalars (or those of some other color basis), each pixel's color (or often just its intensity) is determined as a linear combination of surrounding pixels. In the same way that two different relatively monochrome regions will tend to have similar color-vectors for pixels inside the region — whereas comparing sample pixels from each region yields vectors which are far apart in color-space — two dissimilar textures within an image will tend to have distinct patterns of linear weights, so that these patterns can partition the image into different regions (analogous to segmentation based on color). These principles give rise to autoregression-based segmentation methods.

Radiomic studies suggest that mathematical descriptions of linear weight-patterns along these lines can also be used — apart from image-segmentation — to quantify characteristics of textures for comparison across images, and therefore potentially as a classificatory tool. We mentioned [5]'s analysis which found that an AR-based feature was one of two most strongly diagnostically indicative (the other was a first-percentile histogram, effectively delineating the lowest intensity threshold where a region is separated from its background). In [5] three other parameters also showed noteworthy diagnostic correlations, albeit less consistently than the two just mentioned, so that [5] proposes in effect a five-part radiomic signature that can be derived from patients' Cine MRI videos. However, as a rule, extraction of radiomic signals does not always point toward scientific *reasons*, or "biologic correlation" [85], for why some imaging patterns and not others tend to track disease conditions (some analyses attempt to bridge this gap; *see* [35], [75], [68, page 6], etc.).

It is not always obvious how to interpret such image-feature diagnostic correlations biologically, because the kind of work we have just summarized tends to seek statistical patterns in large numbers of radiomic signals, without anticipating *a priori* which features are likely to be presented differently in diseased tissues or organs than healthy ones. Some correlations are intuitively plausible. For example, the five most-indicative parameters in [5] just mentioned include intensity histogram variance, which measures the degree to which brightness levels vary from place to place within a Region of Interest (RoI). The authors also found noticeable signals related to "wavelets," which generally measure the degree of homogeneity or heterogeneity in an RoI, taking into consideration pattern-(dis)similarity in different directions reinforces homogeneity (or the lack thereof). Intuitively, healthy tissues in many contexts may be more homogeneous than damaged/diseased tissue, or vice-versa. In that sense one might expect that there would be consistent variation in radiomic features derived from pictures of healthy and diseased tissue, respectively, insofar as those features are affected by how textural heterogeneity presents itself visually.

Other discriminative signals have less obvious interpretations. For example, [8]'s findings with respect to autoregression and intensity histograms found particularly strong signals within several specific parameters that are part of larger parameter groups — e.g., the first percentile was calculated to be statistically more pronounced as a potential biomarker than alternatives such as the 25th, 50th, 90th, or 99th (it is not clear why a 1st-percentile intensity threshold should be singled out in this context). Also, in the case of autoregression, their analysis implies that patterns in the weight through which pixel-color is influenced by neighboring pixel-color was most pronounced in just one specific direction. It is not clear what geometric phenomenon in cardiac muscle could account for one autoregression direction being more significant than others, without the use of contextualizing techniques such as those introduced by [17].[1] On the face

[1] Based on [8]'s description of methods it seems most likely that they used AR models built in to MaZda, their analytic software which is also the

of it, the fact that "first theta" parameters in an AR model would form much stronger biomarkers than features from other theta-directions seems hard to account for.

Other studies which similarly look for diagnostically significant image-features have highlighted different parameters, so there does not yet appear to be a scientific consensus on which sorts of feature-vectors provide bonafide biomarkers for heart disease. One factor which likely contributes to this problem is that image quality and metadata vary from one dataset to another (*see* [53, page 3], for example, for an overview of variance based on equipment and/or image-registration techniques, or [84] for assessment of quality-control methods). According to [39]:

> [R]econstructed images can vary markedly not only in image quality but also in how the heart is presented on an image, including changes in orientation of the heart and differences in the plane of imaging, signal intensity of pixels, and degree of artifacts present on the image. Artifacts or poor-quality imaging can degrade radiomic image analysis. The two image quality factors with the greatest impact on texture analysis (TA) — the most common type of radiomic analysis performed in cardiac MRI — are spatial resolution and signal-to-noise ratio .... and numerous additional ones, including MRI field strength and image slice thickness ... [L]ittle to no study has been done to discern how these factors specifically affect cardiac MRI radiomics. These image acquisition-related factors are a potential source of error in published studies. (page 2)

These limitations, however, do not prevent us from considering how the goals of radiomics affect software design, with respect either to applications used for initiating radiomic analyses or those dedicated to showing their results. Cardiac feature-extraction requires a multi-stage image-processing workflow, which would have to be designed in a standardized (and at least semi-automated) fashion for large-scale deployment of cardiac radiomics. Cardiac imaging is usually carried out by recording full 4D pictures of the heart in action, so a preliminary step is always to select particular 2D frames from a full 4D series.[2] Each 2D image accordingly is associated with data concerning how it is oriented within a 4D context. This orientation is moreover defined in terms of recurring patterns in the hearts' rhythms, as well as the hearts' own 3D morphology and positioning vis-à-vis the human body.

Terms of anatomical orientation, such as "saggital," "transverse," and "coronal" (corresponding to yz, xy, and xz planes if we consider the x and y axes to extend left/right toward the arms and front/back respectively, and z to measure height off the ground) are relevant for contextualizing bioimages when the anatomic positioning of the organs or tissues visible in bioimages is consequential to their functioning, which of course applies to the heart. The heart's morphological details — divided into left and right halves with distinct shapes, and with the bulk of mass concentrated in the myocardial musculature enclosing the ventricles — are also of significance insofar as these details guide segmentation and registration algorithms applied to cardiac images.

When multiple images are jointly utilized for a diagnostic investigation, image-registration sets up correspondences between points in one image and the "same" points in a second image, the equivalence between them defined in terms of their underlying anatomical locations. Some registration methods in the cardiac context specifically are based on "control points" (which can be manually or algorithmically identified) defined in terms of the relatively fixed morphology of the heart [32, page 120], [67, page 2], [56, page 14], [9, page 8], etc. Image-registration is often needed in the context of 2D freeze-frames from a single 4D cardiac image because the heart's motion has the effect of shifting the reference frame oriented to cardiac anatomical features against the axes produced by the imaging device [59, page 1012], [40, page 3].

Also, some analyses of single-patient data employ registration algorithms to coordinate image series acquired via different imaging devices, on the premise that distinct image-acquisition methods are more accurate for specific analytic goals: "As each imaging modality provides unique information and overcomes only certain challenges in cardiac imaging, the physician usually prescribes more than one imaging procedure to gather as much information of the heart's condition before making a treatment decision." [51, page 1]. Registration is also used to normalize images obtained from *different* patients so as to "normalize a population of hearts into a common heart template space." [66, page 31]

Considering these registration and orientation requirements, then, any 2D cardiac image has a fairly detailed anatomic context which is established, or must in part be calculated, prior to methods such as texture analysis being applied. Each image is oriented from the spatial/geometric point of view against our planar model of the human torso (in the sense that these details define how the 2D region sits within its enclosing 3D space) and against morphological landmarks in cardiac anatomy. Each image may be oriented to other images either in the same series (showing different time or planar slices) or to other heart-images entirely (with the goal of normalizing the image to a generic heart-model or "cardiac atlas" [26], [92], [21], [31], [72], [30]). This registration-related aspect of orientation produces metadata reflecting how control points, deformation, or axis transforms map the current image onto a different target image. Images are also oriented temporally against the structured sequence of cardiac rhythm. The totality of these aspects of orientation can potentially be aggregated into data structures characterizing the image *context* which is logically anterior to image *features* obtained via radiomic analysis.

basis of numerous other cardiac-imaging studies, which establishes "theta" parameters according to directions that remain constant across the RoI.

[2] Of course analyses can be performed in three or four dimensions directly, but much of the existing literature is devoted to feature-extraction in two dimensions only, so that time and plane slices of the full 4D data need to be computed ahead of time.

At the other end of the radiomic pipeline, meanwhile, one obtains feature vectors such as the five-parameter aggregate identified by [5] as strong diagnostic/classificatory signals. We therefore have two varieties of data structures which need to be joined to particular images: *context* data defining the image's situation in a 4D cardiac representation (which is largely *prior to* analysis logically speaking); and *feature* data derived from morphological and textural analysis (thus largely *after* analysis logically speaking). Connecting these two is the radiomic workflow itself: performing analyses which take the image's context as parameters and compute radiomic features against that background.

Consider these data structures from the point of view of software implementations. The *context* and *feature* data packages bookend the radiomic analysis, and would presumably be relevant to users of the software whether initiating the analysis in the first place or viewing the results. Of course, the (full) contextual data may itself be available only after a complex process with its own workflow, e.g. image-registration possibly paired with manual "control point" annotations. We'll set this detail aside for discussion and just consider the context data as an overall package: a user reviewing the image context would want to obtain information about how the image is oriented temporally and anatomically vis-à-vis the beating heart, and could benefit from seeing control-points or annotations summarizing image-features employed during the registration process. Orientation relative to the saggital, transverse, and coronal planes can sometimes be visualized via a 3D diagram of a schematic torso; the **3DimViewer** application, for example, which constructs 3D models from 2D image-series, uses a three-frame viewport rendering an image for each of the three anatomical planes and using a torso-figure to track the position of the planes relative to one another as users scroll on those frames (*see* [16, page 87], for example). These orientations may also be presented numerically.

From a modular point of view, a reasonable software design might stipulate that *context* data is presented in a secondary window which could have some multi-media content, e.g., a torso-diagram showing planar orientation and perhaps a wave-illustration for temporal anchoring in the heart rhythm, as well as rendering of numeric values for these and other contextualization parameters in key-value form. If context-data involves image annotations (such as control points) these could be shown on the primary image-view, but with the user having the option of hiding those annotations (perhaps the context-data annotations could be overlaid semi-transparently when the context-data window is visible).

Similar principles could apply to *feature* data. A secondary "feature" window might display key-value data for radiomic parameters while also showing radiomic features in visual form when appropriate, e.g. via image-intensity histograms for every Region of Interest. Moreover, annotations on the main image (such as overlays demarcating texturally segmented RoIs) could be switched "on" or "off" (and perhaps rendered semi-transparent when the feature-data window is visible). Context menus and drag-and-drop handlers (where warranted) could interconnect all three of these relevant windows (main image, context data, and feature data).

One goal when designing an image-annotation module would be to furnish a common pattern for how the module's windows are organized, and its functionality accessed, which could be reused by multiple applications. When the same module is found in multiple places, those use-points gain the benefit of sharing common interaction-patterns which may be helpful to users (easing transition between applications, for instance), particularly if the module is well-designed and user-friendly. Indeed, one reason to describe the *data* that would need to be handled by an image-annotation module is so developers can get a handle on what users will want to see when they interact with windows provided by the module. As an example, we have offered a very summarial sketch of "context" and "feature" data that would tend to accompany cardiac-imaging use-cases.

In our basic outline, a primary image-window would be supplemented with secondary windows rendering *context* and *feature* data when appropriate. This setup could then be extended to other secondary data profiles. When using the module to *initiate* radiomic workflows, one window could provide a visual summary of the workflow, and/or even a text editor where the user may compose scripts defining the workflow operationally; this window could then be a starting-point for workflow runs. Of course, normally the actual workflow implementation would depend on other modules, so the annotation module would need to orchestrate data-export and cross-module procedure protocols in coordination with other modules (those data export sites and formats would then be modeled jointly within the module's data model and procedural model). Systematic description of GUI requirements helps translate data and procedural models into user-friendly modular designs, because (ideally) procedural capabilities are exposed to end-users in a consistent manner, one which presents the user with similar experiences across different applications and which is optimized for the specific needs of the module's domain. For example, the part of image-annotation data models related to image-registration (exporting data to registration pipelines) should be formalized in conjunction with specifying how registration data should be visualized (e.g., via control points as annotations).

In addition to workflow definitions, secondary data for image-annotation modules might step outside the imaging context entirely. Recent research, for example, has attempted to refine cardiac image-registration methods by consulting simulation and mathematical models of the heart's mechanics. Mathematical descriptions of cardiac rhythms — and of the associated structural changes to the heart's shape during different phases of the heart-beat — can identify constraints which would also be apparent (projected onto two dimensions, if working in a 2D context) in cardiac images. The "unique combination of ... b-splines in the Fourier do-

main ... (BSF)" introduced in [89], for instance, "aims to improve the tracking accuracy of myocardial motion by an add-on regularisation layer of pairwise image registrations. The proposed framework is designed to enforce spatio-temporal smoothness, cyclic-nature of cardiac motion, and temporal consistency" that is "an 'add-on' regularisation framework ... usable on any ... registration algorithm" (page 2). Virtual Reality is likewise adopted in [1] "to dynamically interrogate biophysical and biochemical events in the 4-D domain" (page 2) yielding statistical signature of cardiac motion-patterns (page 6).

These are examples of cardiac motion simulations yielding data which can improve the accuracy of cardiac image-analysis by defining mathematical constraints or statistical properties of cardiac motion, and the heart's geometry at different points in the beat-cycle. Other simulations ground mathematical cardiac models at smaller molecular/cellular scales ([77], [20], or [71], for example). As supplemental data complementing (and potentially guiding analysis of) image data, such simulations are analogous to computational tumor models that we discussed in Chapter 1. In general, an image-annotation module might need to establish a protocol for viewing information about such simulations as a peer data package (comparable to context and feature data) and/or to interface with a simulation-implementation (analogous to serving as an entry-point for a radiomic workflow). We will return to the *oncology* simulation case later in the chapter.

Before bringing the discussion back to oncology, however, note one further detail in the cardiac context: imaging data may sometimes be integrated with more conventional biomarkers drawn from biopsies, tissue samples, or clinical health records. In [43], for example, the use of image processing to evaluate myocardial fibrosis is double-checked against direct examination of heart tissue (sampled from explanted hearts, obtained after heart-transplant surgery). Programs such as Canada's "HELP" (Human Explanted Heart Program) [90] and the UK Biobank (referenced above; and *see* [74]) encourage researchers to cross-reference cardiac radiomics with other sorts of biomarkers. In [4], data analyzed from the UK Biobank reveals correlations between genetic factors influencing details of cardiac anatomy, such as left-ventricular traits (page 1326), with image-derived phenotypes (page 1320). Genetic factors, specifically MicroRNAs, were likewise correlated with both image and tissue data in [23, see page 9]. In short, some research projects which include cardiac imaging also require analyzing tissue samples, biopsies, genetic data, and other non-image biomarkers obtained from patients in conjunction with image-acquisition.

From the software-engineering point of view, this external data should be linked to images when warranted based on how study-designs provide context for image-acquisition, even if managing that data is not the direct responsibility of image-annotation modules. This raises the question of how clinical, genomic, or histological data should be integrated with different kinds of bioinformatic modules — how should this more general data be packaged so that it may be presented to the user in multiple application contexts, since that data will be relevant in multiple contexts? How should modules request and render data which lies outside their scope (e.g., genetic data in an imaging context)? These questions, which we have noted in terms of cardiac imaging, are also quite relevant to oncology.

2.2 From Image-Annotations to Image Biomarkers

Some use-cases for diagnostic imaging require only relatively low-level image scanning (by a person or computer), such as visually confirming the presence of a tumor or, say, bone fracture, or calculating a tumor's width (or a fracture's degree of displacement). Modern image-processing and Computer Vision applications, however, allow for much more detailed algorithms to integrate image data within informations systems designed for predictive analytics and precision medicine. Textural analysis of tumors or lesions, for example, can yield fine-grained classifications of different patient's particular cancers, which may partition cancer patients into more rigorous groups as criteria for selecting treatment plans, or predicting patient outcomes in light of different possible therapeutic interventions.

The question for image-annotation is how to describe the relationships between image data proper and the textural patterns or image features which are interpreted through the lens of these fine-grained biomedical details. Annotation in the case of simpler, visually evident image-patterns (such as a tumor visible as a darker region against a light background) need only be visually marked or circled to call attention to the Region of Interest, whose biomedical significance is assumed to be evident to a qualified diagnostician who inspects the image. Biomarkers derived from more complex statistical processing of image-data, however, can only be fully described by representing the mathematical results which result from sophisticated image-processing algorithms. The domain of image *annotation*, then, tends to merge with that of feature vectors and/or image-processing pipelines.

For concrete examples of these issues, consider cases such as tumor-microenvironment (TME) research. Radiomics can be used to decode signals latent in tumor imaging which indirectly describe how tumors are biologically interacting with surrounding tissue, measured in terms of parameters or processes such as hypoxia (a situation where a tumor lacks oxygen and tends to respond by more aggressively expanding into surrounding tissue), angiogensis and vascularization (where tumors try to coopt blood supply by spawning new blood vessels) and heterogeneity (reflecting different genetic or morphological patterns in different parts of a tumor, which can potentially make the tumor more resistant to therapy).

One challenge when using image biomarkers in the context of predictive/precision medicine is that of reducing potentially multivariate feature-vectors into signals of just one or two dimensions, which can facilitate grouping patients

into clusters of similar diagnostic profiles. For example, [47] discuss hierarchical image segmentation (with specific applications to diagnosing cervical cancer) where contrasts between each region and its enclosing "parent" region provide additional data points (complementing those derived from regions individually). Some regions are composed of smaller ones which have some level of differentiation, and therefore are relatively heterogeneous, whereas other regions are more homogeneous because their subregions are similar to one another. Measuring heterogeneity and homogeneity across hierarchy-levels allows algorithms to isolate regions which are large enough to be biologically meaningful (smoothing out over-sensitive segmentations that perceive large numbers of small regions due to image "noise").

In particular, important regions tend to be homogeneous at their level and so on down the hierarchy but to be children of noticeably more heterogeneous regions at the next higher level. These considerations give rise to a scalar "homogeneity measure" which can be provided via a single formula. In [47] this measure is paired with a metric of region shape based on the eccentricity of ellipses which best approximate each region. The authors call this measure "circularity," which is larger for shapes similar to a circle and smaller for shapes more like a straight line. The homogeneity and circularity measures are single scalar values applicable to each distinguished region. In [47], RoIs are selected as regions which are large in both homogeneity and circularity, followed by a step where RoIs are further classified (using other statistical parameters) as corresponding to cell nuclei or cytoplasm. In short, homogeneity-plus-circularity forms a compact two-valued signature which serves both as an analytic tool and a summarizing device for cellular-scale image segmentation.

Once nuclei are isolated, cancer cells are indicated by nuclei which are enlarged and have irregular boundaries [79, page 4]. This phenomenon applies to many sorts of cancer, although the diagnostic importance of nuclear morphology is more pronounced in cervical cancer than elsewhere because cervical cancer is commonly diagnosed via blood samples (rather than via imaging solid tumors, for example). Different algorithms can be employed to quantify nuclei deformity, but the common theme is to quantify the deviation of the nuclear membrane from a smooth curve which encloses a similar region [82, pages 4 and 7].[3] The end result is a single scalar estimate of nuclear morphology, which can be applied to all nuclei identified by the prior segmentation. The presence of measurably irregular nuclei correlates with a likelihood of cancerous or pre-cancerous cells, so that these measurements serve as an image biomarker extracted via this form of Computer Vision pipeline.

This review presents merely one simplified account of a full analytic pipeline. There are many different segmentation algorithms which can be employed to isolate nuclei and cytoplasm: [42, page 2] in a recent (2021) study cite 15 papers describing image-processing methods specific to cervical cytology, and three others for nuclear segmentation more broadly. Image segmentation and then classification of nucleus (and cytoplasm) regions are two separate analyses where different methods for each step can be combined independently.[4] Our main point for the moment is that a key step in these analyses is to convert numeric data whose significance is confined to the intermediate stages of image-processing into a small group of numbers that can serve as image biomarkers, ultimately integrated into bioinformatic contexts which combine image biomarkers with other kinds of data (genomic, biochemical, histological, and so forth). The analyses we summarized here yield (first) "homogeneity" and "circularity" metrics for each screened RoI and (second) "irregularity" metrics for regions classified as nuclei. This relatively simple system of three parameters encapsulates image-processing routines which could generate thousands of (intermediate) data points during the course of the pipeline.

Although the goal of image-analysis is usually to reduce complex analytic data into simpler, biologically meaningful metrics, there are many different kinds of derived quantities which can be computed as consequential image-features. In [42] criteria such as contour size, average intensity, "solidity" (defined here as a quotient of contour-area against convex-hull area), and "inertia ratio" (essentially the inverted aspect ratio of an approximating ellipse) are identified as visually distinctive qualities of nuclei (page 3) and used for nucleus classification. In another recent review of extant work, [81] identifies several dozen feature varieties:

> Some authors analyzed four parameters: area, integrated optical density (IOD), eccentricity, and Fourier coefficients. Other authors used 16 features: area of nucleus, area of cytoplasm, nuclear gray level, cytoplasm's gray level, and so forth. Some authors acquired nine parameters: mean intensity, variance, number of concave points, area, area ratio, perimeter, roundness, entropy, and intensity ratio. Finally, some other authors used 27 parameters, which included contrast, energy, correlation, and homogeneity. ... It remains to be studied which parameters are more appropriate for cell classification.

Any of these parameters could potentially be employed as image-features that have some diagnostic/predictive significance, which can result in a given image yielding a diversity of realistic biomarkers, even if most such features only have biological interpretations in specific contexts. Moreover, [81]'s list of parameters are centered largely on those characterizing region *morphology*; different metrics can likewise be obtained for describing image *textures* which are evident inside regions, and which also may have biomedical interpretations (e.g., for tumor-microenvironment investiga-

[3] Informally speaking, techniques can start with a complex contour — viz., the outer boundary enclosing a region — and simplify it to a smooth curve, measuring how much the original contour changes in the process; or, one can proceed in the opposite direction, starting with the curvature one would expect to find on a smooth contour, and measuring how much the actual boundary deviates from these expectation in the neighborhood of individual points.

[4] We are not aware if the algorithms described in the specific papers we cited to summarize examples of the segmentation process and then the classification process have in fact been used together, but they illustrate the kind of workflow endemic to cytological image analysis.

tions as cited above vis-à-vis hypoxia, heterogeneity, angiogenesis, and vascularization). In the context of Covid-19 radiology, for example, [18] describes an algorithm for assessing the probability of SARS-CoV-2 infection from chest CT scans, where hypernodes represent high-dimensional vectors (191 dimensions overall) and hyperedges represent k-nearest-neighbors; here each hypernode represents an entire image, mapped to a 191-dimensional feature-vector.

Some research can potentially close the gap between statistically discerned image-features and biological interpretations/explanations or "biologic correlates" [58, see esp. page 1491*ff*] for their statistical significance by simulating cellular-scale or tissue-scale processes which produce patterns latent in an image. In [29], for example, a theory of "habitat characterization" (page 13) yields a scaffolding for image-feature signatures and (incidentally) an account of biodynamic mechanisms causing radiomic patterns:

> We contend that [image] subregions represent distinct habitats within the tumor, each with a distinct set of environmental selection forces. These observations, along with the recent identification of regional variations in the genetic properties of tumor cells, indicate the need to abandon the conceptual model of cancers as bounded organlike structures. Rather than a single self-organized system, cancers represent a patchwork of habitats, each with a unique set of environmental selection forces and cellular evolution strategies. For example, regions of the tumor that are poorly perfused can be populated by only those cells that are well adapted to low-oxygen, low- glucose, and high-acid environmental conditions. Such adaptive responses to regional heterogeneity result in microenvironmental selection and hence, emergence of genetic variations within tumors. The concept of adaptive response is an important departure from the traditional view that genetic heterogeneity is the product of increased random mutations, which implies that molecular heterogeneity is fundamentally unpredictable and, therefore, chaotic. (page 12)

It is worth noting that variable parameters which govern how image-processing workflows are executed can also serve as biomarkers, or at least as metadata informing how biomarkers should be interpreted (and as such information that should be included in biomarker packages). In the case of [42], the authors outline seven "tunable parameters" (page 7) which their computer code takes into effect, and which can determine the outcome of the segmentation-and-classification pipeline. Since the image-features which emerge from that workflow are dependent on the initial vector of seven predefined parameters, those parameters are also an intrinsic part of the analytic data, and should be recorded when integrating this data into larger clinical contexts.

Our last few paragraphs, then, have presented concrete examples of how radiomic pipelines convert image data (which has mathematical significance only in the narrow context of image statistics) into meaningful biomarkers. To the degree that image *annotations* are used to represent these biomarkers, the annotations proper must be connected with data structures summarizing extracted image-features. Annotations can interact with image-data on several levels. Since features are often correlated with specific image segments or RoIs, the demarcation of the RoI itself constitutes an annotation which serves as a ground for defining feature data. Also, many RoIs are computed by reference to simpler shapes that are defined in their neighborhood, such as ellipses approximating a region's extent, or polygons forming their convex hull. These simplified shapes can be directly encoded as annotations using conventional geometric descriptions. Finally, feature vectors might be encoded as data structures associated with an annotation in the sense that the annotation describes the region to which the feature-vector applies.

The fact that image-features are usually associated with *regions* (not the entire image) points to a close association between feature-vectors and annotations. On the other hand, the scope of (even a general-purpose) annotation framework does not necessarily extend to thorough descriptions of image-features in general, which may require an entirely different set of mathematical and bioinformatic concepts. As the above discussion has hopefully pointed out, there are literally dozens (if not hundreds) of features that might be quantitatively extracted from an image, and it would be difficult to schematically define all such parameters *a priori*. Moreover, different kinds of image-features — and by extension different image-processing techniques — tend to be associated with different biomedical domains. Segmentation of cellular-scale images for the purpose of nucleus and cytoplasm classification is diagnostically important for some clinical contexts, such as cervical cancer. Other kinds of processing, which may involve very different algorithms, have different clinical rationales. For example, tumor-microenvironment research is focused on images at a different scale (e.g., solid tumors rather than individual cells) and is oriented toward texture analysis more than region-morphology.

We can see the overall field of bioimaging as partitioned into multiple (relatively autonomous) contexts, where the image-acquisition modalities, the applicable forms of Computer Vision, the interpretations of image-features as biomarkers, and the prototypical processing pipelines can all vary from one context to another. As such, it is premature to schematically outline the domain of biomedical image processing as a whole, rather than modeling such different technological contexts individually. Moreover, these contexts are not static; new imaging technologies as well as new software/computational methods can improve upon existing radiomic techniques while also consolidating algorithms and workflows which are characteristic of specific diagnostic and investigative domains. For example, enhanced segmentation capabilities emerging over the last decade have apparently consolidated the nuclear-classification pipeline we reviewed

in this section as a canonical methodology for cervical cancer detection. Much of the terminology and numerical details intrinsic to this use-case would be less applicable to, say, tumor microenvironments.

These comments imply that the connections between image-annotations and image-features should be left open-ended, with detailed models of how annotations integrate with biomarkers left to be represented more broadly within computational environments where multiple varieties of biomarkers are juxtaposed. The challenge for modular implementations is to provide enough structure for an image-annotation module and radiomics/Computer Vision modules to interoperate, but simultaneously to avoid pre-emptively restricting the kinds of data and procedures which each module can take on within its own domain. Finding the proper balance between data and procedural expressiveness/open-endedness, on one hand, and rigorous interoperating protocols, on the other, is a crucial artistry within modular design. We will examine this issue in more detail in later chapters.

2.3 Tumor Histopathology and Simulations

As mentioned above, one line of cardiac research has connected cardiac imaging to simulations and mathematical models of heart-beat rhythms and biomechanics. A similar development may be observed in oncology with respect to computational models of tumor growth and evolution. Tumor imaging and simulations can be mutually reinforcing, in that simulations can help explain how biologic mechanisms within the Tumor Microenvironment (TME) engender patterns of tissues, vascularization, and tumor growth that can be viewed on radiographic images, while image analysis can at the same time double-check simulations' accuracy. Simulating tumor microenvironments (and other pathological or histological processes) helps expose the causative factors underlying cancer observations, including those warranted by biomarkers. One of the clinical payoffs is more refined precision/personalized medicine: multiscale biological models may clarify the factors driving categorizations such as benign/malignant and between tumors which do or do not respond to radiation therapy, potentially improving automated classifications in ways that are clinically significant — ideally, selecting probabilistically advantageous treatment plans.

When constructing biological models, many tumor simulations combine models of biological processes (at the histological, cellular, and molecular levels, often integrating models at different physical scales) with spatial and geometric simulations of tumor growth and evolution. Such spatial simulations yield geometric prototypes that can be cross-referenced against image biomarkers. That is, accurate tumor simulations may yield predictions about how tumors with specific properties (e.g. specific degrees/regions of hypoxia) would appear observationally in the context of different imaging modalities, such as conventional radiography or newer whole-slide imaging or nano-radiomics: "modeling has provided mechanistic understanding of phenomenological observations based on physical principles and helped establish important quantitative relationships ... spanning several biologically relevant scales in time and space." [19, page 2] To the degree that such models are correct, the predicted observable patterns could then be considered as biomarkers for tumors having their simulated properties. For example, if computational models suggest that tumors in certain circumstances will acquire unevenly-distributed but consequential hypoxia (sufficient to diminish the effectiveness of conventional therapies) and predict that tumors in this context will exhibit prototypical textural patterns, then image-processing tools which detect such patterns can be deemed accurate in identifying which tumors have levels and distributions of hypoxia that should be factored in to treatment plans.[5]

As this example suggests, there is often potential for identifying correlations between simulations and image biomarkers: simulations help us to understand *why* biomarkers actually signal the conditions which they do, which in turn can help us improve biomarkers' accuracy. The synergy between simulations and image biomarkers is accelerated further by the inherently spatial and geometric nature of many algorithms and modeling primitives employed in both areas. For instance, in "unstructured lattice" simulations individual cells are modeled within a lattice grid and allowed to move independently, subject to system constraints reflecting biological processes (such as cells' access to nutrients, oxygen, and blood flow; *see* [12] for instance) and geometric constraints (such as proper spacing between cells) [86]. Simulations in [78] present an example of lattice-based models generalized to three dimensions. A structurally different lattice-based method, one which permits (rather than forbids) multiple cells on one lattice site, is developed in [24], providing a case-study in how related simulations may present different modeling parameters, assumptions, and structural frameworks that need to be properly documented when comparing simulation results and conclusions. The structure of the underlying grid, along with the specific evolutionary constraints recognized for a given simulation, provide geometric and data-field primitives which algorithmically generate the overall model. These examples illustrate how biophysical and systems-biological models are often based on shared (or at least analogous/contrastable) geometric primitives which capture how large collections of smaller-scale units (such as cells and proteins) aggregate into structures evincing holistic patterns (such as tissues and ECM) in the presence of local forces and generative rules.[6]

When comparing simulations and radiomics, it is also worth considering how models are codified and documented. Alongside the code that executes computational-biology simulations — or, correlatively, image-processing workflows — bioinformaticians have also specified formats for representing models' parameters, assumptions, and investigative purpose. Popular model-description languages in on-

[5] *See* [13] for image-processing algorithms targeting irregular hypoxia patterns.

[6] Similar comments could be made for other mathematical tissue-models, e.g. for surgical simulations; a good overview is [91], or [83], [73], [45], etc.

cology and immunotherapy include SBML (Systems Biology Markup Language), BNGL (the custom BioNetGen language), BioPAX (Biological Pathway Exchange), TumorML, CellML, FieldML, ISML (Insilico Modeling Language), MIRIAM (Minimum Information Required In The Annotation of Models) and MIASE (Minimum Information about a Simulation Experiment). In addition to special description languages, Object-Oriented methods for simulating tumor histology and microenvironments via general-purpose programming languages (such as C++) are analyzed in [15] (*see also* [46, page 138]). A project such as the Digital Model Repository which "allows a model to be executed as an independent computer application" and will "in the future ... enable seamless integration of different computational modules based on the use of various accepted ontologies and semantically annotated objects/parameters exchanged between applications" [88, page 9] employs Semantic Web tools for data *representation* but envisions (pages 7-8)

> having the models and data available in standardized formats with clearly stated dependencies [to] facilitate the creation of workflows that can generate model results to compare model predictions with experimental data, all in an automated fashion. This task can be facilitated by collaborating with other groups developing standard formats for model exchange at different biological scales, such as the Systems Biology Markup Language (SBML) CellML, BioPAX, and FieldML.

which implies a modular design integrating parsers, simulators, and analyzers cross-referenced with empirical data sources — complex software systems that would almost certainly be engineered on an Object-Oriented foundation. Object-Oriented models of cancer-related simulations (tumor formation, growth, morphology, microenvironment, angiogenesis, vasculogenesis, histological properties, and so forth) — and concurrently of radiomic biomarkers (and biomarker-extraction techniques) — can therefore play two complementary roles: to serve as a basis for model descriptions via languages such as SBML, TumorML, etc.; and to implement algorithms for computational-biology simulations and/or image processing. In short, expressing modeling constraints and parameters via Object-Oriented classes allows model description and algorithm implementation to be tied together, which can then streamline the inclusion of bioimaging content since the vast majority of image-processing libraries are based on (Object-Oriented) C++.

These considerations suggest a programming strategy wherein — staying with systems biology as a case-study — the building blocks of systems-biology/immunotherapy-related data sets or research code libraries would be C++ classes which simultaneously provide model descriptions (that may be formalized via SBML and related languages, potentially generating descriptive markup code automatically via metatype reflection) and, via class methods, provide algorithmic capabilities. These building blocks could then be composed and aggregated into heterogeneous data sets and cross-disciplinary research programs, insofar as most immunotherapy research combines multiple forms of biomarkers, data sources, and/or investigative procedures.

So long as each component part of such hybrid models are rooted in a coding method that integrates model-description and algorithm-implementation, complex hybrid models will have on aggregate a consistent interface for descriptive modeling and for algorithmic logistics (the testing, programming interface, data-acquisition logic, and similar requirements for using implemented algorithms scientifically). If consistently adopted, such Object-Oriented architecture would yield more consistently designed and reusable Research Objects (arguably more so than current paradigms, where meta-models and operational requirements are expressed and tested via a diverse and disconnected array of incompatible tools and languages).

As a concrete example, the study of tumor hypoxia (decreased oxygen within tumor tissue, a condition which generally makes solid-tumor cancers more dangerous and resistant to conventional therapies) draws together bioimage markers which can estimate hypoxia via image textural analysis, mathematical and cellular/histological simulations of tumor growth to advance our understanding of how hypoxia emerges in heterogeneous and anisotropic tumor microenvironments, genomic and proteomic data relevant to proteins which influence tumor hypoxia, and empirical clinical or lab data (including tissue samples and disease progression to correlate with bioimages). Most studies of tumor hypoxia combine three or more of these distinct data profiles.

Given that this is the case, it would be helpful to employ a common programming framework to manage all of the data acquisition, modeling, and analysis requirements which are distributed across these distinct domains. For instance, assuming the underlying programming environment is based on C++, all the pertinent data structures integrated in a given tumor-hypoxia research project could be expressed as C++ objects. These data structures might include image annotations identifying regions of interest and quantifiable features (vector fields, diffusion measures, etc.) in tumor images; simulation kernels and evolution parameters for tumor growth models which predict how hypoxia manifests in tumors; clinical records for comparing predictive modeling with actual disease progression; molecular models for proteins influencing tumor hypoxia; and so forth. By structuring all of these research parameters as C++ objects, scientists would then have a centralized paradigm for describing all relevant observational or computational details contributing to theoretical models and research findings, as compared to a diffuse assembly of narrower models expressed in terms of data-acquisition methods rather than data models themselves (e.g., XML file types, table formats, and so on).

In the case of tumor hypoxia, constructions such as textural image biomarkers and tumorogenesis simulation parameters are theoretical posits that can be concretized in Object-Oriented programming idioms, providing a reusable and interactive coding platform which is arguably more con-

venient and pedagogically valuable — more conducive to experimentation — than static figure illustrations or opaque mathematical equations. These points are well illustrated by several existing publications and code repositories which simulate tumor hypoxia via reusable code, built on top of projects such as **PhysiCell**. The same points similarly apply to other TME factors such as angiogenesis and cellular density. In short, a unified programming platform providing classes supporting different aspects of tumor microenvironment research — from image analysis and historical simulations to empirical clinical records and proteomic or genomic database queries — would serve both to facilitate implementation of research code and to document research methods, which in turn would add rigor to publications and data sets summarizing the research project.

It is self-evident of course that "in silico" simulations require computer code. More to the point, simulations (at least in contexts such as cardiology or oncology) are not performed in a vacuum, but rather connected to clinical or pathological/diagnostic data either to provide initial parameters to a simulation, or to double-check its results (or both). We'll mention several interesting examples: [62] describes the process of constructing a "virtual patient" cohort by mining real-world clinical data and then using this collection of virtual-patient profiles as the basis for a Quantitative Systems Pharmacology (QSP) model simulating a tumor microenvironment via biochemical equations, a simulation which propagates to a model of tumor evolution that can be used to investigate immunotherapy mechanisms. A multipart research workflow as in [44] might combine genomic data (specifically, Cancer Genomic Atlas sequences) with cellular data (from the Broad Institute Cancer Cell Line Encyclopedia), with protein abundance metrics derived via mass cytometry, and with simulations of intracellular signaling and of therapy regimens targeting these signaling mechanisms. The *brain* is used as a "control" standing for "normal" tissue in a simulation comparing tumor angiogensis in a hypoxic environment against blood vessel emergence in non-cancerous organs [54, page 4]. An "effort to integrate mathematical models of cancer with real data in an attempt to develop quantitative, predictive models" [70] is combined with multiscale mathematical modeling (this article is a good reference overview on mathematical simulations of hypoxia, invasiveness, and other cancer details as well as documenting new multiscale techniques; *see also* [38] which applies these techniques in a context that also emulates real-life patient "subpopulations," in the sense of cohorts within a clinically and sociodemographically diverse human community). In [52], hypergraph-analysis methods are applied to several datasets and mathematical models relevant to systems biology, such as genetic regulatory networks, human-disease networks, and protein complexes; for example, "node statistics or motif detection" may be analyzed on hypergraphs representing interactions between sets of related genes and of related diseases (page 5). Statistical analyses reported by [49] introduce a rigorous quantitative definition of angiogenic "hot spots" (fluctuations in the density of vascular growth at different regions in a tumor) and demonstrate that such fluctuations are likely to manifest underlying biologic processes rather than result from chance. Next-Generation RNA sequencing is combined with an analysis of tumor samples in [25] and [11] to calculate "a gene-specific model by fitting a smoothing spline with four degrees of freedom to transform RNA-seq data ... into 'microarrays-like' data"; from that transformed data the authors investigate tumor-infiltrating immune cells via microarray-based algorithms, research ultimately targeted at leveraging anti-tumor immune reactions to contain cancers and/or amplify the benefits of cancer treatments [25, page 1037]. Empirical models of "hallmarks" — which are patterns in gene-expression explaining genetic contributions to cancer development that have been incorporated into databases for cancer research — are merged into simulations of cancer development and progression in [63]. Finally, [34] employs decentralized Object-Oriented coding techniques for "a computational multiscale agent-based model capturing spatially explicit dynamics of tumour development in the presence of adaptive immune response" (page 2).

Characteristic of these various multi-methodological studies is the recurrence of biomarker or prognostic factors in different investigative modalities, such as Microvessel Density (MVD) in [49], which can be expressed in results from histological assays based on immunostaining [49, page 19163], in Whole Slide Imaging applied to tumor cross-sections, and also as a simulated emergent pattern in (e.g.) [48]. The former paper moreover presents a method for merging and cross-referencing bioimaging on two different scales (tumor radiography and tumor histopathologic Whole Slide Imaging), allowing MVD image biomarkers to be derived via two different workflows. In [87], a proteomic biomarker (associated with Ki-67, a protein whose levels are correlated with cell division, and in particular with tumors' aggressiveness) and several other Immunohistochemical (IHC) indicators (obtained via tissue reception), as well as basic clinical information (e.g. patient age and "Menopause status"), are juxtaposed with ADC-based (Apparent Diffusion Coefficient) image features, seeking a noninvasive tumor-growth prognosis which would statistically mimic the Ki-67 proliferation index in the context of breast cancer. In [28], histopathologic image features are correlated with genomic, transcriptomic and survival data to derive a classification system for cancer types. The common theme of these research projects is that underlying genetic, biochemical, or biomechanical processes can yield multiple biomarkers expressed in different modes of observation (e.g. tissue examination via lab assays versus non-invasive diagnostic imaging).

In the case of Ki-67, levels of the protein in tumor cells signal the rate that these cells are primed to subdivide (and therefore the cancer's invasiveness). As such, Ki-67 serves as a molecular indicator of tumor characteristics which has prognostic significance. An analogous molecular expression in the context of tumor angiogenesis derives from the "ED-B" isoform (variant) of the protein **fibronectin**, which in general

is associated with such functions as cell growth and vascular development. The ED-B form is particularly involved in embryonic development [69, page 1] but is less present in adult tissue (in the absence of cancer), so that antibodies against ED-B can serve as a measure of tumor angiogenesis [2, page 6]. Proteins can of course be indirectly signaled by antibodies against them as well as by genetic factors, insofar as many proteins depend on specific genes to produce them. Personalized immunotherapy often depends on evaluating genomic signatures marking the characteristics of patients' cancers at a granular level, allowing targeted interventional therapies (as we discussed summarily in the previous two chapters).

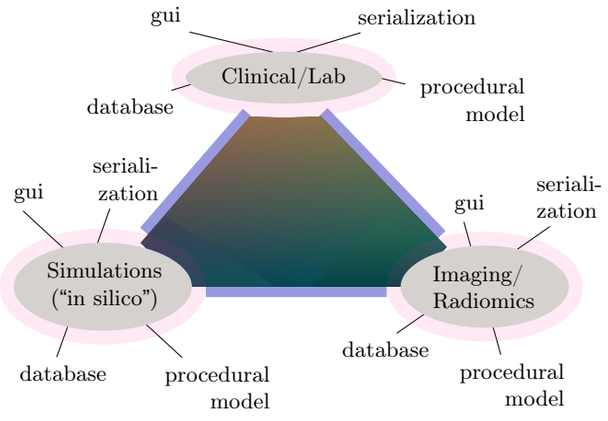
Transcriptomic indicators (i.e., those based on RNA) have also been widely studied for oncology, yielding for instance biomarkers for hypoxia in the context of cervical cancer, for example [65], as well as hypoxia as a complicating factors in radiation and immunotherapy for many kinds of cancer in general [80, page 6]. In sum, proteins, genes, and antibodies can all serve as indicators allowing diagnosticians/pathologists to infer details about cancer growth and the tumor microenvironment, as well as genomic or molecular factors in the cancer which may have implications for treatment plans and prognoses.

Often similar details may also potentially be gleaned from image-analysis at one or more scales (e.g. radiographic scans of tumors or microscopy images of cancerous tissue), so that image-biomarkers could either reinforce or take the place of lab-based blood work or tissue analysis.[7] Sources of information about cancer properties and TME may therefore involve some combination of (1) non-invasive imaging; (2) tissue/blood lab analysis (proteins, genes, antibodies), which in turn may entertain (2a) cytometry, or lab assays (whose diagnostic mechanisms would comprise, say, color-tests and test-tubes), (2b) image-analysis on tissues/cells (e.g., Whole Slide Imaging), or (2c) genomic/transcriptomic sequencing; and (3) biochemical or biomechanical simulations of tumor growth and evolution. From the computational point of view, these diverse observational modalities give rise to a distinction between software components focused on simulations, imaging, and clinical/lab data management, respectively (see Figure 1). Many research projects combine two or three of these modalities, so that this figure (which we will discuss further in later chapters) presents component-types as (vaguely defined) domains that tend to integrate and interoperate in different combinations.

Thus far in this chapter we have discussed the combination and cross-referencing of image biomarkers with other clinical and molecular disease-indicators. The various examples and case-studies that we have mentioned are of course only a sampling of volumous research literature that could be summarized; a review of the breadth of data-acquisition procedures, observational modalities and research data sources for cardiology or oncology is well beyond the scope of one chapter. Our overview and examples however are hopefully sufficient to picture the information landscape so as to present relevant software-design issues. As an organizing device, we have focused on the theme of bioimaging, which translates over to design considerations for software components presenting images (and their annotations) and potentially interfacing with image-processing workflows. Given the correspondences between different forms of biomarkers sketched out via Figure 1, the analogous software-engineering concerns derive from the challenge of integrating components devoted to the three "vertices" of the triangle (imaging, simulations, and clinical/lab data), which we will address in the rest of this chapter.

[7] Of course, image-analysis can replace lab methods in some context but reinforce them otherwise, so that non-invasive imaging can be used for diagnostic/prognostic purposes but direct lab analysis used as well when patients need to undergo invasive procedures anyhow; and research comparing results from imaging and from more invasive methods can also be done to establish a baseline for how (and how well) imaging results correspond to lab-based results and vice-versa.

Figure 1: Triangular Relationship Between Simulations, Imaging, and Clinical/Pathology/Diagnostic Data



3 Multi-Aspect Modular Design in a Heterogeneous Data Space

In this section we explore the idea of "multi-aspect" modular design, which adds to the underlying principle of modularity the notion of individual modules having multi-faceting software-engineering responsibilities, including (for example) GUI rendering, data persistence/serialization, and runtime reflection or remote procedure capabilities, e.g., exposing a scripting interface for modules' functionality. The above example of a bioimaging/annotation module interfacing with modules in parallel domains (genomics, histology) serves as a plausible illustration of why multiple-aspect design can be implementationally beneficial — in our running example, an imaging module may at some point in time present the user with data which invites follow-up through other modules, but it is outside the scope of the initial (imaging) module to know how such external data should actually be presented.

An image currently being viewed could link to other data structures simply by virtue of its biomedical representatum — cardiac images to echocardiogram readings or heart tissue biopsies, for example; tumor images to genetic tests for

proclivities to a cancer variety or tests for the patient's anti-tumor immunological responses. A bioimaging module, e.g., might acquire notifications to the effect that such links are available (as part of the image metadata) but would not know how to display a GUI containing genetic, cytological, or histological data, for example, which is why the actual presentation of that extra-modular data would be deferred to the proper modules. At that point, then, those modules would be presented with contextual information (perhaps an identifier for the current image series, i.e., the on-view image's container, which would likewise indicate the relevant patient) and would have to identify the relevant data within their own scope applicable to that context (they may need to pull relevant info from a data "lake," for example).

Insofar as each module has multi-faceted and mostly self-contained functionality, one module could provide a concise entry-point through which a peer module would be able to derive a more comprehensive package of information to present to the user. Starting from a medical image, say, the user could potentially wish to see a detailed overview of related cytological, histological, or genomic data. Insofar as such data is presented via their own modules, each of these may (from the entry-point provided the current viewed image's metadata) reconstruct a holistic data package and present this via their own GUI components, so that users would experience the overall progression as switching between (or juxtaposing) image-based views with views characteristic of a separate domain (histopathology, and so forth).

The idea behind multi-aspect modular design is to orchestrate flows of user activity along these lines in a relatively decentralized manner. Modules which are "multi-aspect" could independently promote a holistic User Experience given a concise entry point — e.g., an image series and patient identifier leading to presentation of genomic or histological data in an integrated GUI component. Modules which were narrower in their range of engineering concerns would require greater central control to permit the "switch" in user focus across domains. If GUI and database access capabilities were separated between two different modules, say, then transitioning from an image-view to (for example) a histopathology view would require a central controller which *first* loads the relevant histopathology data and *then* passes this data to a histopathology-GUI component. Having the GUI and database-query functionality merged into a single histopathology module would eliminate most requirements for centralized control, allowing the imaging and histopathology modules to interact on their own.[8]

3.1 The Overlap Between Research and Clinical Data

We have thereby set forward a case for modular design — where modules are interoperable, but relatively self-contained (in the "multi-aspect" sense that they package features related to GUI design, data persistence/serialization, and so forth), and at the same time prioritize minimizing external dependencies and adapting to diverse computational environments, rather than being usable only in a select range of environments where computational performance can be optimized.[9] Interrelated clusters of code libraries linked to different sub-specialties are provided by some software projects (**BioConductor** is a good example), but these components often do not span multiple aspect and feature-sets with the breadth that we propose via "multi-aspect" design.

As a rule, for example, **BioConductor** packages do not provide built-in database integration or GUI features (except those inherited indirectly from the R environment (which provides a GUI context for evaluating R code in general but does not natively provide tools for packages to customize the GUI to their needs, without a foreign-language bridge such as Qt bindings). Reviewing prominent projects for open-source biomedical components, there are libraries which emphasize analytics, UI development (e.g. **OncoJS**), data-set preparation (as with the Research Collaboratory for Structural Bioinformatics Protein Data Bank), APIs for genetic/proteogenomic data access (**EPICO**, GDC, etc.), simulations (**PhysiCell**, **Amber**, **Chaste**, **MMBios**), and so forth, but the components within these projects do not internally address the multiple aspects of (in particular) GUI, database access, and data serialization all together.

Likewise, the three APOLLO ("Applied Proteogenomics OrganizationaL Learning and Outcomes") networks — GDC, TCIA, and CPTAC (the Proteomics Data Portal) — all provide some tools to help researchers submit and acquire data sets. Most of these tools, however, are exposed as web services rather than as code libraries that could be bundled into scientific applications. For example, in the case of CPTAC, the UI code is designed to provide data visualization capabilities in conjunction with downloaded CPTAC data sets. Specifically, downloads acquired through the Proteomic Data Portal will include HTML files providing interactive plots and other figures summarizing information encoded in the rest of the data set. If CPTAC support is encapsulated in a module embedded in scientific/biomedical applications, similar visualization files may be targeted at the host application, allowing users to visualize the CPTAC data summaries directly rather than opening a separate web browser to examine the included HTML files.

With respect to CPTAC and TCIA, both of these networks require a multi-step data-acquisition process which could be streamlined with the help of dedicated software components. The TCIA API is split between two interfaces, and while one

[8] Not to imply that central monitoring would be fully ruled out, because one may still intend certain cross-modular interactions to be validated. For example, patient privacy concerns might imply that a technician granted access to one part of a patient's data would not necessarily have authority to view other parts. The point however is that central control is not *implementationally* necessary; modules are *able* to orchestrate cross-modular functionality on their own, even if programmers may want to use central monitoring to override potential traffic when appropriate.

[9] Publishers and hospitals, we would argue, have a shared interest in curating software-development tools and partial code libraries that could be leveraged to build modules in both publishing/scientific data-hosting and clinical data management contexts, and in particular modules focused on different biomedical subdisciplines (radiomics/bioimaging, cellular systems, cytology, histology, genomics, and so forth.

of these interfaces may be accessed via a client library provided in Java and Python, these tools have limited value for standalone biomedical applications (*see* [50], for instance). Similarly, the UI tools provided with CPTAC could be generalized to an integrated proteogenomic toolkit combining the CPTAC and GDC UI components. The combined code base would then be available as a suite of GUI classes suitable for embedding in scientific/biomedical applications.

In the context of **MMBioS**, a C++ API — identified as a project-aim in conjunction with curating "a database of molecules, rules, and models that can be used for comparative analysis of existing models and development of new models"[10] — would permit **BioNetGen** "actions" to be called directly rather than through a command line, though perhaps the API could be designed (as is a popular pattern in scientific applications) to support workflows that can be manifest either through command-line actions or directly in code sequences (as well as indirectly via remote procedure calls). In addition to generating command-line invocations, the **BNGAction** Perl Module (used to access **BioNetGen** functions for data management and analytics) does rather detailed preparatory checks in some cases, so equivalent functionality would have to be implemented as an API layer. In other words, a C++ API would presumably have several stages, with preliminary logic at one stage yielding data structures to a second API layer that communicates directly with underlying C++ code: C++ data structures would be generated in lieu of command-line invocations.[11] There are numerous options for modeling distributed/asynchronous procedure requests (e.g. whether the response requires a separate parsing/value-extraction step, whether response callbacks carry state, and the specific reactive/function-object mechanisms are used to supply callback procedures); a rigorous API should allow each of these options to be employed when appropriate — itemization of the various cases could be part of (what we call) a "procedural-exposure" model — and should clearly define the protocols and requirements in each case.[12]

Finally, in the case of GDC, a property-graph based data model integrates molecular, clinical, and genomic data according to predefined data types and interconnections. This model is instantiated within computations used by the GDC to validate and harmonize submitted data sets prior to their being made publicly available (data sets may be submitted to GDC either through the GDC API or via an online portal). However, although the GDC data model is clearly documented on the GDC website, the GDC does not provide software tools or code libraries to facilitate the implementation of computer applications which would curate data submitted to and/or acquired from GDC so that researchers could perform their own validation steps in preparation for the GDC analysis.[13] Similarly, the GDC User Interface components — which are intended for software providing visualizers for GDC data — are based on **JavaScript**, thereby requiring a **JavaScript** programming environment and/or an HTML rendering engine. In short, the GDC data model, API client, and UI toolkit are each targeted at different programming environments, and the GDC itself does not provide a unified framework which integrates these areas of functionality into a common platform. The GDC's published code and web services document the requirements for applications seeking to interface with GDC data submission, validation, and acquisition protocols, but it is left to third-party software to unify these capabilities into a single platform.

In short, failure to promote software development environments which facilitate the implementation of relatively broad-featured and "multi-aspect" modules contributes to either a paired-down modular design, which (we contend) inhibits data integration, or to the consolidation of software ecosystems whose building blocks are monolithic applications more than inter-combinable modules, which contributes to ecosystem fragmentation.

There are six or seven facets of data management which come to the fore with database systems in the context of application integration, APIs, scientific data curation, and publication/dataset management — parsers for special-purpose languages; domain-specific query evaluators; custom GUI components; interacting with analytic capabilities both in-process and out-of-process; metatype systems allowing software components to model scientific processes/phenomena; application integration via type-level serialization and persistence; and dataset/publication integration. This set of concerns reappears in numerous scientific-computing contexts (one can identify similar patterns in bioimage processing, for instance), which suggests that they may serve as a general architectural framework for developing scientific-computing "modules."

Requirements such as GUI design, serialization, and database interop reappear in different contexts in different ways — if we consider research projects that involve some combination of clinical/lab data, bioimaging, and simulations, these concerns will be present in each of these areas (we have tried to connote this visually by inserting concerns diagrammed by a "saltire" as part of the triangular outline in Figure 1 — this depiction will be clarified in Chapter 8). Because these different research areas tend to be combined and integrated in different ways — scientists' flexibility to piece together different research projects and paradigms in various combinations is potentially an important source of new discoveries — one could argue that embodying research data and protocols in multi-featured (acting as relatively self-contained mini-applications) but inter-combinable

[10] https://mmbios.pitt.edu/research/technology-research-and-development/network-modeling

[11] Though at this step it would make sense to allow such data structures to be serialized into workflow descriptions, executed indirectly via command line or RPC and other deferred/distributed methods, etc., as well as being executed directly.

[12] It also appears that much of the data visualization associated with **BioNetGen** runs through a similar Perl/command-line pipeline as BNG Actions, so presumably the API could support visualization, perhaps expanding the range of visualization outputs (maybe full-fledged C++ GUI components instead of text formats such as GML).

[13] Although the official GDC API client is a guide for programmers who wish to generate requests against GDC endpoints, it cannot be utilized directly to access GDC data unless applications embed a Python interpreter.

modules is a better software-architectural match to the contemporary scientific landscape than either single monolithic applications or narrower "single-aspect" code libraries which need more centralized oversight to interoperate.

3.2 The Problem of Software Ecosystem Fragmentation

It is often via an evolutionary and decentralized process that software components, data formats, and analytic methods get consolidated into digital "ecosystems" with their own paradigms and conventions. Digital image-processing is a good example; bioimaging (in particular) tends to gravitate toward certain canonical image-acquisition formats (e.g., DICOM), file types (e.g., TIFF or PNG) and analytic libraries (ITK, OpenCV, etc.). These components fit well-defined roles, allowing multi-component workflows to arise organically even if they are not formally proscribed, rehearsed, or crafted *a priori*. That is to say, software use-cases often embody "informal" workflows, which amount to recurring patterns in how different components' functionality are pieced together in order to implement what is needed for research projects. Such intuitive patterns are structurally quite similar to formal workflows, even they are not standardized or officially notated as such. Moreoever, software components tend to cluster together based on their coexisting within the scope of informal workflows along these lines.

The converse can also be true: informal usage-patterns might become entrenched into distinct ecosystems even if there is some overlap within their domains and methodology. Perhaps IHC assays evince an example of this phenomenon, being designed to yield visually obvious diagnostic markers, in contrast to sophisticated image-processing methods that detect subtle signals in (say) radiographic images. Because the images generated by laboratory assays such as IHC can be affected by how technicians prepare the imaged tissue samples, these assays are designed to yield signals which are as "unsubtle" as possible; refining the laboratory methods and equipment involves making the experiment more accurate by amplifying the desired observable effect, such as the pattern of staining evident in certain parts of a tissue sample in contrast to the background. For those sorts of reasons, detailed image-analysis is not an intrinsic part of the (informal) workflow usually associated with techniques such as IHC, in contrast to scenarios such as radiographic scans where researchers have limited control over how tissue images are visualized *except through* automated image processing.

As a result, the software ecosystem centered on assays such as IHC appears to be somewhat isolated from contexts which rely more on intensive bioimage processing. We have not done a rigorous analysis of usage-patterns (to the degree that such an analysis would even be feasible), so these comments should be considered impressionistic and observational, considering existing literature related to entrenched protocols such as IHC. Yet, as some supporting evidence, attempts such as [57] to refine IHC quantification point to the tendency of IHC to rely on image-viewing software rather than image-processing libraries for deriving summarial results. More generally, software ecosystems are more likely to become fragmented when the principal components of those systems are *applications*, rather than (in particular) code libraries, or also (say) plugins which can extend applications' functionality. Imaging applications such as **ImageJ** or MaZda provide some image-processing capabilities (e.g. "lasso" tools to grab region contours) and are popular in some scientific contexts (including IHC, judging by literature frequently mentioning **ImageJ** and its peers in discussions of research protocols), but workflows which rely on users manually interacting with applications are less robust and extensible than software ecosystems which can pass data to specialized domain-specific code libraries.

Commercial software products also predominate in many workflows involving special data-acquisition devices, such as biosensors. Flow Cytometry, MFP probes, and SPR equipment (just to mention technologies identified in this chapter or Chapter 8) are powered by machines that provide their own software (which has to be custom-implemented given the unique physical mechanisms and configuration options of the machines involved). Flow Cytometry is a good case-in-point: commercial vendors of FCM instruments tend also to provide software applications accessing the data which their equipment generates. While it would be theoretically possible for researchers to export FCS (Flow Cytometry Standard) files from the instruments' software and write their own analytic code, most scientists appear to be more comfortable working within the confines of existing FCM applications. This situation is roughly analogous to using image viewers or DICOM consoles for image-evaluation, rather than hand-coded image-processing algorithms. By way of comparison, research projects built around intensive image-processing are more likely to feature a custom code base, with different components sharing data via (at least potentially) automated pipelines.

We alluded several paragraphs ago to bioimage workflows organized around widely-used libraries such as DCMTK (for managing DICOM data) and ITK or OpenCV (for image processing). These kinds of components are typically joined together within an overarching code model; for example, a research project may develop a code library which links against both DCMTK and OpenCV, implementing procedures which perform the steps need to carry data resulting from DCMTK processing over to analytic code featuring OpenCV. Alternatively, DCMTK and OpenCV might form the core of two separate modules which would interoperate via a command-line interface, but in this case the individual modules would still be designed with the understanding that their respective functionalities need to be synthesized into an overarching workflow. Individual components, that is to say, intrinsically support functionality allowing them to be used in a multi-modular context, such as initializing their working environment via a Command-Line interface and exporting data according to shared protocols.

In effect, each individual component supplies the requisite capabilities allowing different components to be pieced together, and moreover with sufficient preparatory code these workflows can fully or partially automated. In these sorts of contexts an important step in research design is to develop a code base which supports automated workflows along these lines; the research code orchestrating the flow of data between workflow components and the sequence wherein operations exposed by each component are triggered.[14]

Workflows achieved through custom programming as just outlined can be "informal" in the sense that they are not explicitly described (or conceived as "workflows" per se), but instead simply follow common usage-patterns, where various code libraries have evolved to play specific roles (and to expose functionality and data formats conducive to multi-component interoperation). Nevertheless, these workflows acquire a certain rigor because they are made possible by specific kinds of functionality being provided within each component, notably code for importing/exporting data according to specific formats, and one or more "entry points" (or what we will term in Chapter 6 "meta-procedures") which can be initialized with parameters specifying how the components' specific contribution to the larger workflow should proceed. These sorts of workflows can be relatively flexible, with clearly delimited prerequisites for how they may expand in scope — either by existing components recognizing a wider range of data formats, or exposing new functionality, or via separate components encapsulating extended functionality being designed in a manner that permits their integration into the existing workflow patterns.

Fragmentation of software ecosystems into isolated clusters of components typically used together may still be a problem, but at least there is a technical foundation for considering the proper extent of workflows' scope — one can identify the particular elements within each component which support their interoperation. These elements are the specific sites in the components' code which would be affected if researchers or programmers were to adopt usage patterns that effectively widen the scope of existing (maybe informal) workflows; one could ask, for example, how difficult it would be to implement support for new data formats (in terms of parsers and runtime representations of new kinds of data given components' existing parsing and representational procedures). Likewise one could consider whether it is practical to implement functionality needed to manage new kinds of data (e.g. the structural or mathematical operations endemic to new data profiles which depart to some degree from those the components has previously targeted) given components' existing architecture and capabilities.

Issues of ecosystem fragmentation are more likely to become entrenched in the context of informal worflows which are, so to speak, "application-driven" in the sense that major components are distinct *applications*, often provided by commercial vendors, rather than (one could say) "code-driven" (where components are code libraries). Insofar as informal workflows take the form of common usage-patterns for distinct applications, the logistics of sharing data and orchestrating the proper sequence of operations tends to rely on human users manually interacting with the applications. Compared with code libraries — which are explicitly designed to be integrated into larger programming environments — monolithic applications generally have fewer features enabling functionality available within the application to be exposed for automated workflows. Moreover, full-fledged scientific applications tend to be difficult to extend, even if they are open-source projects with no commercial impediments to customization.[15]

In short, monolithic applications (compared with more open-ended code libraries) tend to be resistant to context-specific modifications which could allow applications to participate in multiple workflow-like environments. One consequence of these limitations is that informal workflows centered on *applications* rather than on *code libraries* tend to be more rigid, foreclosing the possibility of workflows expanding in scope, which in turn drives and reinforces (what we are calling) "fragmentation."

For a concrete example, we will note in Chapter 8 that Flow Cytometry gating and data visualization, and also viewing data generated by image-processing pipelines (such as identified Regions of Interest) have many parallels with image-annotation. These overlaps suggest that a single suite of GUI components could potentially be used to cover both image-annotations and feature-visualization, and moreover extended to other modes of data acquisition such as Flow Cytometry when they engender image-like presentations. One benefit of unifying distinct concerns along these lines is code-reuse. As long as the domains involved are not prohibitively divergent, broadening the scope of existing GUI environments to accommodate a wider range of use-cases can be more efficient than recreating entire GUI tools *ab initio*, even if wider-ranging code bases could become more complex as they support special-purpose data formats, use-cases and functionality.

While it is worthwhile to analyze such trade-offs between code reusability and complexity, for now we simply note that questions about the proper scope for code components tend to dovetail with data-integration concerns. Consider a scenario where image-annotation, image-feature visualization, and Flow Cytometry GUIs are confined to distinct code libraries. One consequence of this separation is that the respective components will likely employ somewhat different

[14] It may be imprecise to describe such workflows as "automated" because custom programming is needed to implement the code which acts as a framework for workflows to be executed (we are not referring in this context to workflows visually designed through a workflow-management application rather than by programming them directly). Once the overarching code is implemented, however, each particular iteration of the workflow sequence can typically be performed without human intervention midstream.

[15] For example, complex applications are often difficult to compile from source (as compared to installing a prebuilt binary package) which precludes extending the application by modifying the source code directly — this is especially true for software intended to be run on relatively high-powered computers found in research settings, which may have extensive external dependencies that would be impractical to reproduce on more pedestrian hardware. These kinds of scenarios could stymie a graduate student, let's say, trying to work on some specific extension to the application code on a generic laptop computer.

representations for annotation (and gating) geometry, image dimensional data, data set provenance, and similar artifacts which structurally overlap across all three domains. Consider the data generated when a user alters the geometry or visual style of an annotation or Flow Cytometry gate, or an image region segmented/demarcated via tunable processing parameters. The domains of image annotations, features, and FCM are arguably similar enough that representations of user actions along these lines will be strongly correlated, and could be expressed via a common description language. Standardizing the respective components' representation of user actions would be beneficial in contexts where data from two or three of these modalities are integrated, so that projects could maintain a holistic record of users' actions during the course of a research cycle.

Common representations of similar kinds of data are more likely when the components that generate such data are implemented as distinct pieces of an overarching environment (e.g., distinct GUI components within a larger GUI toolkit), or at least are self-consciously designed to be interoperable. Fragmentation of software ecosystems can have the effect of obscuring possibilities for the synchronization of data-representations along these lines. More to the point, the presence of multiple data sources which evince similar data profiles, but express information via structurally discordant data models, impedes the process of data integration because such mismatches end up requiring extra "bridge" code. Granted that standardization efforts try to promote interoperability between components engineered by different teams and companies — as we will argue at the end of this chapter, external interop initiatives can have only limited success when they get layered on a code base retroactively, rather than emerging organically from components adapting from the design phase onward to a modular/interoperative environment.

4 Data-Integration via Multi-Aspect Modules

A reasonable overview of oncology or cardiology research — how clinical and diagnostic results are obtained; how scientists explain the biological mechanisms behind cancer or heart disease and extrapolate disease signatures and prognostic indicators from those explanations — suggests networks of interconnected but narrowly focused research programs. Methods and terminology can vary substantially, depending on how researchers target different scales — for example, analyzing precancerous lesions or cardiac tissue versus larger-frame analysis of heart movements or solid-tumor morphology — and also whether the focus is on *in vitro* or *in silico* experimentation on disease processes in controlled environments, or evaluations of real patients (for diagnosis, prognosis, or selecting among treatment options).

Similar dispersion may be found in the software which threads through these research agendas; clusters of similarly-focused research work tend to converge on a particular set of software applications, code libraries, or algorithmic conventions. Here we refer to this clustering-effect in terms of "ecosystem fragmentation," or the tendency of research communities to evolve distinctive patterns of software use and computational paradigms, which may have merit in that they encapsulate "best practices" discerned over time, but can also be somewhat inflexible and isolated.

We argue here for a more accommodating methodology which allows modules encapsulating numerous distinct biomedical disciplines to be connected and combined in flexible combinations. We also advocate for modules which are adaptive to different computing environments, without being laden with complex dependencies or locked in to exceptionally high-powered technologies.

To be fair, many research/diagnostic methods in (say) cardiology and immuno-oncology are quiet subtle, relying on precise computational treatments to derive biologically meaningful findings from faint statistical patterns. As such, software which is adequate for these sensitive computational tasks should be fine-tuned for (metaphorically) amplifying faint signals. Under these circumstances, one might reasonably question whether "fragmentation" is a bad thing; perhaps instead this is the only way for researchers to consistently use methods which yield reliable results.

More generally, one can question the extent to which disparate research projects are truly "integrated" in the substantive body of their work mid-stream, as compared to the handful of general principles, clinical indices, diagnostic protocols, and other practical results which hopefully come into focus as research progresses. Obviously, an important research phase is translational — distilling the science into a relatively simple explanatory or evaluative framework that can fit into existing clinical knowledge and methodology. This may take the form of a few prognostic indicators, or a probability distribution estimating the favorability of different personalized treatment plans. Such data points or recommendations would then enter the clinical record and could be a factor in how physicians proceed, with the overall lab or diagnostic process serving as an integral but self-contained interlude in the larger trajectory of patient care. Ultimately, research is most relevant when it yields protocols that slot in to clinical practice in this manner, but the profound details of the research — the complex science behind single biomarkers or indices — can be largely self-contained within the research work or, to the degree that it has practical applications, to the work localized within a given lab or diagnostic center, whereas mostly just summarial findings become integrated into the larger course of treatment.

If complex research data does not in and of itself tend to flow into the clinical mainstream, and if a given body of research has progressed to the point where established protocols exist to yield relatively simple diagnostic findings and recommendations which are clinically relevant, then it is reasonable to ask why any importance should be attached to the

insular or conventionalized nature of computational environments which drive research. It is true that many research projects dip into multiple biomedical subfields at once, and therefore make use of computational resources from distinct "ecosystems," but these are often separated as different stages or facets of the overall research. Well-organized research papers usually provide adequate detail describing methods and instrumentation, including description of software protocols (sometimes including published source code) applicable to distinct phases of the research. This means that competent research work is transparent about its methods at each stage, and claims about how the various smaller parts of the research may fit together, to create a larger scientific theory, can be evaluated conceptually.

Much biomedical research is interdisciplinary — whether the juxtaposition of different methods and perspectives exists within a single paper or more within the interplay between multiple research projects which take the same problem from different angles — but integrating diverse disciplinary perspectives is often only possible by accepting certain empirical or theoretical results localized within one disciplinary area as a starting point for further integration. Establishing the foundation of theories or data involves work narrowed to that specific area of research; to the degree that scientists feel confident about component results, the goal of cross-disciplinary integration is one of unifying models accepted as provisionally validated by encapsulating their contributions into a few most important details, poised to be conceptually and operationally merged with contributions from other directions. Interdisciplinary collaboration does not necessarily entail low-level engagement of different scientists on the evidential minutiae which need to be curated by individual research programmes to the degree that they can flow into larger multi-disciplinary paradigms.

Such an account of low-level details — "encapsulated" by research programmes that become, in effect, subtheories in a space of integrative scientific practice — would seem to argue for low-level details being the concern only of small groups of researchers (or, upon translation to operational practice, to technicians who have the isolated responsibility of providing their own well-defined step in a clinical pipeline). This is a plausible picture which might accurately describe an ideal of clinical "division of labor," but it is incomplete (we claim) in several contexts, which we will analyze to conclude this chapter. Specifically, we find this an oversimplified account when considering, *first*, large-scale biomedical data management; and, *second*, the publication, dissemination, and reproduction of research work. We will elaborate on these claims over the next several subsections.

4.1 Research Dissemination and Incremental Replicability

So, why can't we consider low-level research details as fully encapsulated within narrowly delineated research projects or clinical practice, which would legitimize what we have called "ecosystem fragmentation"? Here we will address two issues.

First, consider the question of reproducing research. While the term "replication crisis" may be exaggerated, it is true that scientists in numerous fields — especially medicine — have increasingly prioritized structuring research in ways that promote replicability, and that this tendency is driven by scientists' frustration at failures to replicate prior research [76], [6], [36]. If not a "crisis" (a term which might hyperbolically imply that large swaths of medical knowledge could be discredited) then such failures are surely at least a "problem." Assuming that research work is done rigorously, this is not necessarily a problem which researchers can solve individually — the best any individual scientist can do is pursue progress with the most current theories and research tools/equipment possible, and if new science (e.g., new data-acquisition equipment) disconfirms earlier results, the overall trajectory would still be one of science being continuously refined. What researchers *can* do is structure their methodology to prioritize transparency and reduce the difficulty of recreating the research work as much as possible.

While laudable in theory, replicability can be complex in practice, since quality research by definition will often involve cutting-edge ideas and/or material — the physical accoutrements of empirical research — such that duplicating the scientific environment is impractical for logistical (even if not conceptual) reasons. As we stressed above, a lot of research — especially in the context of bioimaging — has to tease out subtle patterns from complex image and/or statistical data, often depending on specialized image-acquisition tools, and such methods may depend on sophisticated investigative equipment and/or powerful computer systems which cannot readily be mass-produced. In these situations researchers can still promote replication by transparently describing their techniques and providing future scientists with guidelines on how to reconstruct their work *in those contexts where* requisite hardware and/or software tools are available, but actually following through on such reproductions may involve logistical and financial hurdles. In that sense, even conscientious research may be difficult to reproduce in practice. Since researchers should *not*, in truth, be discouraged from leveraging advanced scientific tools (since these may be the engines driving new discoveries) — however scarce access may be to them among their peers and their field's community writ large — it would be counter-productive to fret over obstacles to replication to such an extent as to obscure the value of original research to begin with.

Nevertheless, science as a whole can still take steps to minimize impediments to replication. The central dynamic of replicability as a *problem* is that sophisticated research may be hard to recreate because only select groups of scientists have access to the materials which would enable such replication. Here we can set aside other (in theory more corrigible) source of replication problems, such as poor research design or execution *ab initio*, or failure to properly document methods and assumptions. As scientists become more

cognizant of replication issues, it is reasonable to hope that these more superficial hurdles could gradually dissipate over time.[16] The more intractable problem is that breakthrough research may be difficult to emulate precisely because research novelty often requires investigative modalities that are not widely available, or on innovative conceptual or mathematical frameworks that will take time to be digested by the larger community.

Against this background, optimal strategies for mitigating replication issues would not necessarily start from the goal of reproducing entire research projects *tout court*. However, replicability is not an all-or-nothing proposition, where scientists need to re-enact every aspect of a research project in order to certify a replication success. Instead, it is helpful to think of replicability as *incremental*; as a matter of degree. We should be able to reproduce parts of a multi-faceted research agenda even if it is difficult to redo every piece of the original puzzle. Moreover, prioritizing replicability can also be seen to include facilitating future researchers' ability to identify what would be *involved* in replication to varying levels of detail or thoroughness. The depth and breadth of a replication endeavor should be something that scientists can fine-tune based on available resources and equipment.

Consider a complex, multi-faceted research project where logistical barriers (financial requirements, equipment availability, and so forth) would may it difficult to reconstruct the project in its entirety. Future scientists can still approach replication in a couple of different ways. First, they can assess the feasibility of reconstructing the whole project, or at least substantial portions thereof — separate and apart from full-fledged replication there is the question of planning and preparing for a replication effort, or estimating what would be required for such an effort to take place. Second, scientists can decide to re-evaluate some portion of a multi-faceted project. They could attempt to confirm the methodology or validate the data involved in some parts of the project, or to recreate all or part of the project to a limited extent (which may involve less comprehensive data sets, less precise equipment, and so forth, perhaps not equaling the standards of the original project, but still contributing some information to an overall assessment of the initial research work).

These possibilities raise several questions *which can be anticipated by the original research framework*: even if this research is carried out using relatively scarce equipment and (e.g., computational) resources, are there paths to reproduce the work (albeit imperfectly) in a less stringent environment? Can some component parts of the research be re-evaluated and (hopefully) re-confirmed even if redoing the whole project is impractical? Enabling some level of *incremental* replication can therefore be considered an indicator of quality research design from the outset.

[16] Research trends toward greater quality and precision, and one manifestation of such progress is better research design (enforced by review boards and funding sources, hopefully) and how scientific writing clarifies methods, data sources/availability, or formal protocol descriptions (through projects such as FAIRSharing, "Research Objects," and MIBBI, or Minimum Information for Biological and Biomedical Investigations), enforced (hopefully) by publishers.

The idea of "incremental" replication has several consequences from a software point of view. Even if a research project uses very large data sets (and this scale is intrinsic to the investigation's merit) there may still be value in approximating the research methods in the context of "smaller" data. Modestly-sized data sets could be employed, for example, to estimate or prototype strategies for reproducing the research as a whole. Small data sets could also double-check the accuracy or programming logic of algorithms and/or implementations featured in the original research. Likewise, scientists might at least examine the feasiblity of reconstructing prior research on less advanced but more widely available equipment. Would confirmation of the original work (or, for that matter, contra-indicative results) reinforce (or, respectively, challenge) the original work, or is the original methodology tightly bound to the sophistication of its specific materiel?

These issues point to the domain of replicability being more general than just the full-scale reenactment of prior research work. The larger scope of replication bleeds into metatheoretic framing and pedagogical dissemination of a research programme. Consider scientists evaluating what would be entailed in attempt a relatively broad restaging of some complex research. Should we characterize such pre-replication study as a pedagogical happenstance (the scientists trying to arrive at a detailed familiarity with the prior work so as to estimate the scope of a potential replication) or as a conceptual analysis of the original work? The line between pedagogy and meta-methodology may be hard to define in practice.

In short, the community which might be involved in the full scope of replication could be much larger than just those who specifically take on the role of actually carrying out large-scale reproductions. Aside from explicit and relatively full-fledged replication efforts we have to consider planning *for* replication, assessing how replication projects may be carried out, replications of smaller parts of multi-faceted work, and so forth. Organizing a research programme so as to facilitate subsequent follow-up, then, is not only a matter of streamlining the process of recreating the original working environment in its totality. It is also a matter of enabling scientists to reproduce part of the research setting, or to recreate the environment in a partial or limited manner, so as to *prepare* a more comprehensive reenactment or to replicate just one *part* of the prior work.

Projected onto the specific domain of software development, these concepts imply that the software ecosystem through which new research is disseminated and assessed would, in concord with broad-based replication paradigms, be considered more widely than just in terms of the logistics of full-scale research re-enactments. Replication involves more than just software, of course. It may require the correct genre of equipment, procurement of tissue or other biologic samples (for *in vitro* or *in vivo* studies), access to data for reuse (or functionally analogous data), and so on. Nonetheless, software in particular is well poised to serve as a case-study

for (what we have termed) "incremental" replication. The software ecosystem through which research may be *incrementally* reproduced or re-evaluated need not duplicate the software through which the original work was carried out (or even the computational resources which would be needed to fully recreate the original context).

Consider scientists investigating what would be *required* for replication, performing a kind of pre-replication prototype of the original project. They would not necessarily need to work on data with the same scale, or computational resources with the same power, as necessary to rederive the original findings. The goal of such preliminary replication is not to actually recreate the original work, but rather to prototype the environment within which that work *could* be reproduced. In addition, such "pre-replication" should be deemed an intrinsic aspect of replication in general. The consequence of this idea for presentations of scientific findings is that respecting replicability involves more than transparently reporting on methods and protocols *for the benefit of* scientists who might engaging in replication full-court. Replicability also entails anticipating the needs of scientists who may simulate the original research environment in a simplified, more modest, or prototyped fashion *as preparation for* potential replication in the more exacting sense.

The possibility of "incremental" replication is one element which complicates, we believe, the problems associated with (for example) what we have called "ecosystem fragmentation." The arguments for research communities narrowing in on relatively isolated and "insular" clusters of computational paradigms and software applications tend to focus on the needs of reproducing research with a level of detail and sophistication commensurate with the original. Indeed, we concede that it is reasonable for scientists to develop research programmes which narrowly focus on certain specific software and computational resources if these are the only options for subsequent researchers who want to build on or re-examine the original work in an environment on par with that original. However, *incremental* replication complicates this picture. Full-fledged replication does not come out of thin air: it depends on scientists intellectually mastering the original work to a degree necessary to play the part of the original researchers in re-enacting their work; and on some anticipation or prototyping of the environment where the replication will be carried out.

In short, researchers should assume that "replication" does not just mean a small handful of follow-up project structurally analogous to the original work. More broadly, replication also involves partial, simplified, or prototype-like simulations of the original research environment and methods toward pedagogical and preparatory ends. Replicability is facilitated by the relevant research community grasping some of the conceptual and operational details of what full-fledged replication would entail, even if many of those researchers are not in fact in a position to launch a replication project of their own. For this dissemination to work most effectively, the larger community which assesses research should ideally have access to at least a simulacrum of the original scientists' research environment.

Moreover, an ecosystem designed for "pedagogical" recreation of the original environment can be more open-ended and less exacting than the original environment itself, because the purpose of such an incremental ecosystem is to sharpen scientists' understanding of research methods as much as to produce new data. A widely-accessible "incremental replication" ecosystem could be built around low-cost materials, open-access (and not prohibitively large) data sets, open-source code, and software components which do not have intractable dependency chains or hardware requirements and which are not locked in to extra-ordinary computing frameworks (e.g., some version of the research protocol should be enactable on ordinary desktop computers).

In short, support for incremental replication along these lines changes the design requirements for software components which are intended to be part of a research ecosystem. The goal of software in this "incremental" context is not to maximize computational performance, or to achieve scientific breakthroughs by leveraging raw computing power. Instead, software in the milieu of incremental replicability serves the primary goal of providing an accessible and educational window onto the protocols and computational patterns intrinsic to a given research programme. Such an ecosystem should seek to disseminate an operational and granular understanding of a particular research project within the relevant scientific community as a *precursor* to more thorough reproduction efforts. The software which a community uses to initially conceptualize and model original research need not be the same software as that which powers actual full-fledged replications, but we should on principle consider that broader community-scale understanding to be a prerequisite for the planning and prototyping of replication efforts when they do get put into practice — especially when the original research involves rarified methods or equipment that takes effort to reincarnate in a new setting.

4.2 Heterogeneous Health Data and Curation

The issue of "incremental replication" which we just discussed points to how original research is disseminated in a larger community than just the set of scientists who may in fact be in a position to reproduce prior work with some level of completeness, especially if that work involves materials that are not accessible to many scientists. Even if only a small subset of the relevant scientific community is in position to feasibly contemplate comprehensive research-reproduction, the process of preparing for such replication — and for the replication effort itself to pay due dividends in terms of the larger community's appreciating its results vis-à-vis the original work — would seem to depend on the larger community itself, overall, having some operational and granular understanding of the original work. That goal in turn may best be achieved by presenting the community with concrete tools

to reenact the original work in a partial and approximative manner, for pedagogical as well as empirical reasons. Software components encapsulating research work, in these kinds of context, may be addressed to a larger group of scientists than just those who intend to reconstruct the original research in a computational environment on par with the original. In the context of incremental replication, software components also or primarily serve goals related to the conceptual and pedagogical dissemination of the original research frameworks and protocols.

Analogous roles might be played by certain software components in the context of multi-domain bioinformatic/clinical data curation. For sake of discussion, we will consider biomedical data in the context of relatively large-scale and heterogeneous information-spaces such as a data "lake," where the goal is to deposit as much information as is practically storable, without constraining the data to fit predetermined schema or representations. Hospitals and other medical institutions have increasingly turned to some form of data lake along these lines, although we will speak here in terms of generic paradigms rather than specific technologies (we are imaging hypothetical information spaces which instantiate the conceptual notion of data lakes or their peers).

Consider the trajectory of a patient's care between hospital admittance and discharge. It is reasonable to assume that the hospital will accumulate a significant corpus of information about that patient, from a mixture of real-time data collected from devices monitoring the patient's condition, to test results and doctor's observations vis-à-vis the patient's evolving condition. Some of this data will be registered on patients' Electronic Health Records, but other information may simple be discarded, or might be siloed into different contexts. For example, intermediate data used to generate lab or diagnostic results may be retained by the clinical entities (inside or outside the hospital) which contribute findings to the official health record, but such more detailed data (presumably more granular but also less summarial and reusable) might not be computationally accessible within the hospital's digital ecosystem.

This book has focused on bioimaging as a source for case-studies exemplifying issues related to biomedical data in general, so, continuing that (expository) device, consider the specific situation of hospitals outsourcing diagnostic or prognostic investigations to external imaging centers. Data communications between the two entities (the hospital and the imaging center) would presumably be governed by standards such as DICOM, which should clarify how relevant clinical data would be presented to the imaging center and how summarial results would be sent back. In addition to structured diagnostic reports or treatment recommendations, the information sent back to the hospital might include some images, although not necessarily the full image series relevant to that specific patient and/or study (potentially only the most diagnostically pertinent images might be shared) or the full data generated during image-processing (for example, the imaging results could include, so as to filter out midstream calculations, a more compact radiomic "signature" quantitatively merging extracted image-features which prior research suggests are correlated with the patient's specific medical condition). A larger quantity of information may be retained by the imaging center, e.g., on a DICOM database.[17] There would typically, however, be no automated network connection which would allow the hospital receiving the imaging results to access the center's associated PACS archive, should more granular imaging data be required on their end.

The data "lake" paradigm generally takes a more liberal view of sharing and storing data. In this hypothetical imaging context — again, we are speaking in general/conceptual terms, not analyzing specific implementations — we can envision the relevant hospital and imaging center adopting a more data-hungry protocol wherein a relatively larger volume of image assets and/or metadata is shared between the two entities.[18] The hospital might maintain a system functionally analogous to PACS which could retain multiple images for each study (and therefore each patient) as well as a reasonably thorough database of radiomic and radiographic image-features extracted from those images.[19] The net result is that within a hospital network itself anyone with proper access rights would be able to pull up patient images, along with annotations and/or image feature-vectors, in conjunction with other branches of patient data (clinical reports as well as diagnostic results in other media, such as blood or tissue samples or genetic sequencing).

There are various scenarios where greater breadth in retaining patient data may be useful. One would be revisiting earlier findings in light of new information: imagine a patient who returns to a hospital some months or years after a prior visit; doctors could find it relevant to look at images taken during that earlier stretch of care. Or, test results from non-imaging modalities might cause doctors to reconsider how the images are to be interpreted. Similar re-evaluation can be warranted if a patient does not respond to intervention in ways which accord with their prognostic cohort, as established by an initial image-based assessment. Also, re-analyzing the original images via different software or different radiomic/radiological methods might yield variant results. The possibility of gleaning new results from re-examining prior data should not be foreclosed due to lack of data availability. Finally, there are always possibilities of patient cases being material for subsequent research; image data (and other patient records) may be analyzed in light of the eventual patient outcomes. Were a patient to be entered into a clinical trial, image data may become relevant insofar

[17] All PACS (Picture Archiving and Communication System) workstations are typically programmed to maintain a database of images viewed through the system, aggregated in conjunction with patient and study metadata.

[18] We write here in terms of *hospitals* but similar comments would apply to smaller medical institutions, such as "Urgent Care" centers or even private doctors, assuming that the relevant technology can be down-scaled to the computer systems typical of smaller offices and that we envision many private systems interconnected into a sort of "virtual" Data Lake.

[19] The details of how such data would be communicated alongside the images themselves would have to be established by the protocol connecting the hospital's and imaging center's computer networks, presumably based on image-annotations.

as such data is pooled from the cohort of trial participants for statistical analysis or data mining.

These comments for bioimaging would also apply to other biomarkers/indicators, such as those derived from blood or tissue samples, genetic tests, functional examinations (e.g., cardiac stress tests or assessments of cognitive functioning), or observational clinical data. Assuming a single software system is available to access the full spectrum of data thereby curated, that one system would accordingly be an entry point to information instantiating a broad range of data profiles. In such an environment users of the associated software would have opportunities to explore the data sets along numerous paths or directions — images of a patient's heart, for example, might lead towards results for patients' exercise tests, blood work, clinical observations (e.g., tracking hypertension levels over time) and genetic data that might have implications for heart disease. Analogously, Covid-19 patients' lung scans could be linked to SARS-CoV-2 antibody tests, assessments of cognitive functioning, contact-tracing data, and so forth. In effect, the software-design issues here are not only those of storing and maintaining large and heterogeneous information spaces, but also structuring the interface for accessing that data in a flexible manner, giving users freedom to traverse the space in multiple directions and according to multiple criteria.

It would be difficult to implement such a system effectively without a rigorous modular design, because the range of data profiles would exhibit significant heterogeneity. Consider the case of a GUI window showing cardiac or tumor scans and annotations, which is then linked to a separate window showing histopathology results and a further window showing genomic information. The data structures and computational protocols associated with these three domains — radiomics, genomic, and histopathology — are sufficiently different that it would be difficult for a single GUI "template" to effectively handle all three cases. Modular implementations would allow components to be focused on specific areas: modules for image and image-annotation rendering would be separate than those presenting genomic data and from those devoted to histopathology, for example. Such modules might be implemented by different teams, based on rigorous understanding of the science underlying the forms of data they emphasize, rather than trying to fit into a predetermined mold (for data representation, GUI layout, task organization, functional design, and so forth).[20]

Assume, then, that a hospital system employs a *modular* form of "data lake" software system where the common heterogeneous data source is accessed by collections of discrete modules, each optimized for specific scientific areas: bioimaging, genomics, cytology, histopathology, hematology, neurocognitive informatics [22], epidemiology, and so forth. One issue is then protocols for interoperation between modules; support for flexible usage-patterns implies that users should be able to switch between (or visually juxtapose) views provided by different modules. Continuing the above example, cardiac or tumor images might be linked to modules showing genetic data and results on tissue sample-tests, respectively. The imaging module would therefore need to know first which other modules are potentially linked to the current patient data which the user is viewing, and second how to describe this current data so that those peer modules would present information which is relevant to that current context — for example, histological analysis linked to the tumor investigated by the current image series.

[20]This is not to imply that modularity is necessarily embraced within EMR software — for example, commercial clinical data management vendors appear to develop software in a more centralized manner, but such systems are also often criticized by practitioners for being inflexible and difficult to use.

4.3 Modularity and the Clinical/Research Overlap

One goal in this chapter is to present modular design as a solution to problems arising from "software ecosystem fragmentation," the idea being that encapsulating functionality in *modules* (which can be flexibly combined and modified) as opposed to relatively monolithic and isolated *applications* would counter the tendency of usage-patterns consolidating into disconnected paradigms. Issues of research replication add further considerations insofar as clusters of entrenched usage-patterns can hinder replicability, even if the impetus toward quality research is precisely the force which carves out ecosystem boundaries in the first place. That is to say, researchers may repeat similar usage-patterns because those are paradigms which are optimal to their research work, and would presumably be so for re-enactments of that work in many cases. In short, *full-scale* replication would seem often to entail future scientists converging on the same computational ecosystem (or at least a functionally similar one) as that surrounding the original work.

As we have argued, the problem with this picture is failing to take *incremental* replication into account. Suppose we grant that for some research work a comprehensive reproduction would (to achieve the best results) use the same or similar software as that providing ambient capabilities for the original. We suggested earlier that, such requirements notwithstanding, full-scale replication may only be feasible (and only maximally useful to the relevant scientific community) to the extent in which researchers have operational understanding of what replication entails; can maybe perform some partial replication in miniature or as a simulacrum of the original; and insofar as prototypes for the replication are discussed and modeled as part of the research process. In short, "incremental" replication can serve as a precursor to full-scale replication, as a tool for building a deeper conceptual and logistical understanding of the applicable research protocols, and as a pedagogical prompt helping the larger community understand the research with greater depth. Replication reinforces the scientific parameters of a research project — instead of a one-off operation, a project reproduced one or more times becomes abstracted from its precise institutional context, it becomes a pattern that can be restaged with some level of variation from place

to place. But grasping research projects as a "gestalt" in this sense is easier if scientists in the larger research community engage with the research work at least to some degree in an active, experimental fashion — not just reading an article but carrying out their own mini-replications, for example, or exploring the software which supports that research gestalt.

Even if a certain narrowly-circumscribed software ecosystem is necessary for holistic reconstruction of a research project, then, this is only one facet of replication — a further dimension is the broader dissemination of exploratory familiarity with the research methods, a collective intuition of the research challenges from an operational point of view that serves as a precursor to potential replication; and the software appropriate for this pedagogical and exploratory stage may be different from what is prerequisite for the technical management of replications comparable in scope to the original. Against this background, we would argue that similar architectural models appertain to information systems such as clinical "data lakes." In the same way that the community of scientists who may be engaged with a research programme at some interactive/operational level is broader than just those with the resources to contemplate holistic replications, likewise a clinical data space spanned by heterogeneous domain-specific modules will be visited by specialists in many areas, and modules should anticipate being using by a diverse array of practitioners, not only those with technical skills finely tuned to the module's core domain.

In the case of radiomics, for example, complex image-processing software and/or code-libraries may be needed to extract feature-vectors with sufficient parametric diversity to support radiomic "signatures" and Machine Learning algorithms endemic to contemporary immuno-oncology or (say) cardiac genomics. These tools, then, for understandable reasons, tend to from a kind of "sub-ecosystem" understood by experts well-versed in the science and mathematics of statistical image-processing. Part of the role of software serving these technical communities is to permit effective data sharing and communication: consider the case of an image series being re-evaluated by a different practitioner, or two different image series (perhaps testing disease/treatment progression) being compared. The interplay between two different practitioners or teams in these contexts is analogous to the relation between an original research group and the consort which replicates their work — these teams may converge on similar software patterns because they are simply guided by desire for the most accurate results. Carrying the analogy over to an imaginary hospital context, research-replication might be compared to a hospital prescribing diagnostic imaging and then sending the resulting image-series to a second lab for re-evaluation, and/or evaluating a second study later in the course of care ("replicating" the original analysis, so to speak). Because the two practitioners/teams (or the same practitioners at two different times) are performing two different analyses, there may not be actual data sharing involved (a replication project does not typically reuse the original data, but rather generates a new data set), but the teams may use similar software in each case, software which is functionally targeted toward image-analysis in particular and is disconnected from the hospital's own data space.

Conversely, consider a modular data lake where at least some radiomics-based capabilities and information models are integrated with the hospital's own data management system. In that case, images and radiomic features would be data aggregates accessible within the total package of patient data, and may potentially be consulted by specialists in different medical areas. Instead of radiomic signatures being curated once in a bioimaging laboratory and then only revisited, if at all, by peer practitioners in a a similar technical setting, the radiomic features of a bioimaging module could potentially be explored by a broader community of users navigating through the overall clinical information system. This broader "community of users" is therefore analogous to the community of researchers who may be involved with "incremental replication" of a research project, a larger group than the scientists who might engage in a replication study in detail. Consider the case of interacting with a radiomics modules and then switching to a histopathology module accessing tissue data drawn from the same cancer patient. An analogous transition in the publishing context might be evaluating a data set used for studying radiomic signatures for tumor vascularization and then switching to a module providing simulations of blood vessel formation in the tumor microenvironment *in silico*.

Whereas a radiomics *application* could serve the needs of isolated laboratories sharing data with referring hospitals in only limited, predefined patterns, a radiomics *module* would be designed for a broader use-base, something that might be embedded in a heterogeneous data-management system and interoperate with other modules. A radiomics *module* therefore could not assume that it resides in a computational environment tailored to Computer Vision specifically; ideally such modules would be adaptable in the sense that more esoteric dependencies are optional, e.g., that the code is not tied to exceptionally recent compiler versions or library prerequisites (even if swapping in alternatives yields a degradation in performance). The purpose of a radiomics module would not be instantiating the most powerful radiomic capabilities which science can offer at the moment (as compared to bioimaging software that may be deployed in a diagnostic lab) but rather to expose radiomic capabilities to a larger community, insofar as radiomic data is intrinsically interconnected to information keyed to other biomedical domains (which in turn would occupy other modules). Analogously, too, this situation is comparable to software enabling "incremental replication" having different and (with respect to user base) broader priorities than replication *tout court*.

The analogy between modules targeting a "data lake" and modular design for "incremental replication" has another angle — if we consider *publishers* as akin to *hospitals* (or other institutions curating heterogeneous clinical data). In a decentralized sense a publisher's digital platform is indeed

roughly analogous to a data lake: assets on such a platform include publications themselves, of course, but also bibliometric data such as influence factor (e.g., references into and out a publication), and, increasingly, digital resources such as nultimedia content, data sets, and computer code. Taken in totality, such assets collectively function as an information space whose scope, diversity, and architectural challenges are comparable to biomedical records of a large health-care system. Data sets accompanying publications would be analogous in turn to domain-specific biomedical records which a hospital (say) might store alongside more generic clinical data — e.g., image-processing annotation and feature vectors reported by an external diagnostic-imaging lab.

In some respects this is more than just an analogy; after all, the kinds of data which may be included in a sufficiently broad-based EHR system (radiomic, genomic, histopathological, etc.) are also curated by data-hosting platforms and dataset-archives associated with research publications. Some of these are taken directly from clinical/diagnostic practice — consider projects such as the Genomic Data Commons (GDC), the Oncology Research Information Exchange Network (ORIEN), Human Protein Resource Database (HPRD), The Human Protein Atlas[21], BioGPS[22], The Cancer Imaging Archive (TCIA), the Clinical Proteomic Tunor Analysis Consortium (CPTAC), the International Human Epigenome Consortium (IHEC)[23], or the APOLLO network, the GDC, TCIA, and CPTAC, which we considered earlier.[24]

The term "data lake" is encountered more often in the context of clinical records than scientific publishing. However, scientific publishing platforms could generally match the conceptual underpinnings of "data lakes," particularly if we include research data sets (in many cases the organizations hosting scientific data repositories are not the same as publishers hosting collections of books and articles, but there are some companies which play both roles; and, conceptually as well as looking toward the future, we can envision the two technologies merging, so that publishing platforms of the next generation could well feature publications, multimedia, and data sets coexisting, in a cross-referenced and integrated fashion). We can envision modular design playing a role vis-à-vis publication portals analogous to our proposals for bioinformatic data "lakes." In the biomedical contexts, modular design is appropriate because a data lake will contain information with diverse profiles, some involving idiomatic data structures with specialized algorithmic, persistence, GUI, or user-interaction requirements. The same could be said for data sets associated with (and hosted via) a publishing platform. Different modules could be used to render data sets depending on their underlying scientific field, just as different modules would be selected to display data packages in a clinical system depend on the data's disciplinary provenance (radiomics, histology, etc.).

More to the point, in relatively open-ended contexts such as data lakes or publishing portals, software components providing such specialized features will often be better designed as *modules* than as *applications*. Whereas monolithic applications may have the requisite power to work with data to optimal degrees (e.g., diagnostic labs re-evaluating findings presented in a clinical data lake, or scientific teams replicating published research projects in detail), many users would have less stringent requirements, because modules can be interconnected in a manner that invites users to navigate between them. Furthermore, modules need to allow this free-form navigation; being able to co-exist and interopate with other modules in a dynamic fashion can be more important, from the perspective of modular design, than achieving maximal computational performance.

We will argue in later chapters that these ideas have interesting consequences for such issues as data-sharing protocols and information metamodels. One rationale for emphasizing common data *representations* is that information must often be exchanged between monolithic software components engineered in relatively "closed" environments. In that context the basic units of interoperability are often common data models, and secondarily common behavioral contracts for working with shared data.

Consider, however, a more open-ended development ecosystem where autonomous parties may contribute functional pieces to large-scale bioinformatics platforms in a modular fashion. Such modules are not data *standards*, but rather fully-implemented components that work on data directly — achieving standardization through shared implementation, rather than simply through normative mandates — and can be inserted into multiple applications. The module in one application responsible for some specific information-domain would be sharing data with *the same* module in a different application, not just a component which adheres to the same behavioral constraints. Simply using the same code in two different applications obviates the need to document how disparate components should be behaviorally aligned.[25] And, insofar as an application may not want to be forced to adopt a single code base to provide some functionality, at least alternative libraries could seek alignment with their peers by emulating those peers at a relatively low-level code/procedural level. Open-source code allows for components to be synchronized by studying each other's implementations, rather than through oblique standard-definitions.

In short, if the basic mechanisms of behavioral alignment are *code reuse* or, to similar effect, "implementational alignment," constructions such as data models and meta-representations can be built around concrete procedural models rather than abstract logical summaries of desired behavior, which has interesting consequences for theories of information representation (including ones mediated by Conceptual Space paradigms). We will return to this discussion in Chapter 9.

[21] https://www.proteinatlas.org/about
[22] http://biogps.org/dataset
[23] https://epigenomesportal.ca/ihec/index.html
[24] Not to mention biobanks, which curate not only patient *data*, but patients' actual tissue samples (for research rather than just therapeutic/diagnostic purposes) or cell lines.

[25] All code by definition is behaviorally aligned with itself!

References

[1] Arash Abiri, *et al.*, "Simulating Developmental Cardiac Morphology in Virtual Reality Using a Deformable Image Registration Approach". *Annals of Biomedical Engineering*, Volume 46 (2018), pages 2177-2188. https://link.springer.com/article/10.1007/s10439-018-02113-z

[2] Artor Niccoli Asabella, *et al.*, "Multimodality Imaging in Tumor Angiogenesis: Present Status and Perspectives". *International Journal of Molecular Sciences*, Volume 18, Number 9 (2017). https://www.mdpi.com/1422-0067/18/9/1864

[3] Saeed Ashrafinia, "Quantitative Nuclear Medicine Imaging Using Advanced Image Reconstruction and Radiomics". Dissertation, Johns Hopkins University, 2019. https://jscholarship.library.jhu.edu/handle/1774.2/61551

[4] Nay Aung, *et al.*, "Genome-Wide Analysis of Left Ventricular Image-Derived Phenotypes Identifies Fourteen Loci Associated With Cardiac Morphogenesis and Heart Failure Development". *Circulation*, Volume 140, Number 16 (2019), pages 1318-1330. https://pubmed.ncbi.nlm.nih.gov/31554410

[5] Bettina Baessler, *et al.*, "Subacute and Chronic Left Ventricular Myocardial Scar: Accuracy of Texture Analysis on Nonenhanced Cine MR Images". *Radiology*, Volume 286 (2018), pages 108-112. https://pubmed.ncbi.nlm.nih.gov/28836886/

[6] Stefano Canali, "Towards a Contextual Approach to Data Quality". *Data*, Volume 5, Number 4 (2020). https://www.mdpi.com/2306-5729/5/4/90

[7] Irem Cetin, *et al.*, "A Radiomics Approach to Computer-Aided Diagnosis with Cardiac Cine-MRI". https://arxiv.org/abs/1909.11854

[8] Irem Cetin, *et al.*, "Radiomics Signatures of Cardiovascular Risk Factors in Cardiac MRI: Results From the UK Biobank". *Frontiers in Cardiovascular Medicine*, 2020. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7667130/

[9] Xiang Chen, *et al.*, "Deep Learning in Medical Image Registration". *Progress in Biomedical Engineering*, Volume 3 (2021). https://iopscience.iop.org/article/10.1088/2516-1091/abd37c/pdf

[10] Chakra Chennubhotla, *et al.*, "An Assessment of Imaging Informatics for Precision Medicine in Cancer". *Yearbook of Medical Informatics*, Volume 26, Number 1 (2017), pages 110-119. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6250996

[11] Pornpimol Charoentong, *et al.*, "Pan-cancer Immunogenomic Analyses Reveal Genotype-Immunophenotype Relationships and Predictors of Response to Checkpoint Blockade". *Cell Reports*, Volume 18, Number 1 (2017), https://pubmed.ncbi.nlm.nih.gov/28052254/

[12] Yong Chen, *et al.*, "Simulation of Avascular Tumor Growth by Agent-Based Game Model Involving Phenotype- Phenotype Interactions". *Scientific Reports*, Volume 5 (2016). https://www.nature.com/articles/srep17992

[13] Dmitry Cherezov, *et al.*, "Revealing Tumor Habitats from Texture Heterogeneity Analysis for Classification of Lung Cancer Malignancy and Aggressiveness". *Scientific Reports*, Volume 9 (2019). https://www.nature.com/articles/s41598-019-38831-0

[14] David A. Clunie, "DICOM Structured Reporting". http://www.dclunie.com/pixelmed/DICOMSR.book.pdf

[15] Anthony J. Connor, *et al.*, "Object-Oriented Paradigms for Modelling Vascular Tumour Growth: A Case Study". In *Fourth International Conference on Advances in System Simulation*, Proceedings (2012). https://people.maths.ox.ac.uk/maini/PKM%20publications/354.pdf

[16] Jingjing Deng, "Adaptive Learning for Segmentation and Detection". Dissertation, University of Swansea, 2017. https://cronfa.swan.ac.uk/Record/cronfa36297

[17] Shihong Deng, *et al.*, "Autoregressive Image Interpolation via Context Modeling and Multiplanar Constraint". In *2016 Visual Communications and Image Processing* (VCIP), Proceedings, pages 1-4. https://ieeexplore.ieee.org/document/7805448

[18] Donglin Di, *et al.*, "Hypergraph Learning for Identification of COVID-19 with CT Imaging". *Medical Image Analysis*, Volume 68 (2021). https://pubmed.ncbi.nlm.nih.gov/33285483

[19] Prashant Dogra, *et al.*, "Mathematical Modeling in Cancer Nanomedicine: A review".*Biomedical Microdevices*, Volume 21 (2019). https://link.springer.com/article/10.1007/s10544-019-0380-2

[20] Nancy K. Drew, *et al.*, "Multiscale Characterization of Engineered Cardiac Tissue Architecture". *Journal of Biomedical Engineering*, Volume 138 (2016). https://pubmed.ncbi.nlm.nih.gov/27617880

[21] Frances Duane, *et al.*, "A Cardiac Contouring Atlas for Radiotherapy". *Radiotherapy and Oncology*, Volume 122 (2017), pages 416-422. https://cdn.mednet.co.il/2019/06/A-cardiac-contouring-atlas-for-radiotherapy.pdf

[22] Włodzisław Duch, "Neurocognitive Informatics Manifesto". *Cognitive Sciences ePrint Archive*, 2009. https://arxiv.org/abs/2101.03609

[23] Iacopo Fabiani, *et al.*, "Micro-RNA-21 (Biomarker) and Global Longitudinal Strain (Functional Marker) in Detection of Myocardial Fibrotic Burden in Severe Aortic Valve Stenosis: A pilot study". *Journal of Translational Medicine*, Volume 14 (2016). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5002330

[24] Grazziela P. Figueredo, *et al.*, "On-Lattice Agent-Based Simulation of Populations of Cells Within the Open-Source Chaste Framework". *Interface Focus*, Volume 3 (2013). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3638480

[25] Francesca Finotello and Zlatko Trajanos, "Quantifying Tumor-Infiltrating Immune Cells from Transcriptomics Data". *Cancer Immunology, Immunotherapy*, Volume 67 (2018), pages 1031-1040. https://pubmed.ncbi.nlm.nih.gov/29541787

[26] Carissa G. Fonseca, *et al.*, "The Cardiac Atlas Project -- An imaging database for computational modeling and statistical atlases of the heart". *Bioinformatics*, Volume 27, Number 16 (2011). https://academic.oup.com/bioinformatics/article/27/16/2288/253995

[27] Reza Forghani, *et al.*, "Radiomics and Artificial Intelligence for Biomarker and Prediction Model Development in Oncology". *Computational and Structural Biotechnology Journal*, Volume 17 (2019), pages 995-1008. https://www.sciencedirect.com/science/article/pii/S2001037019301382

[28] Yu Fu, *et al.*, "Pan-Cancer Computational Histopathology Reveals Mutations, Tumor Composition and Prognosis". *Nature Cancer*, Volume 1 (2020), pages 800-810. https://www.nature.com/articles/s43018-020-0085-8

[29] Robert A. Gatenby, *et al.*, "Quantitative Imaging in Cancer Evolution and Ecology". *Radiology*, Volume 269, Number 1 (2013), pages 8-15. https://pubmed.ncbi.nlm.nih.gov/24062559

[30] Tarun Kanti Ghosh, *et al.*, "Multi-class Probabilistic Atlas-Based Whole Heart Segmentation Method in Cardiac CT and MRI". *IEEE Access*, 2021. https://arxiv.org/pdf/2102.01822.pdf

[31] Kathleen Gilbert, *et al.*, "Independent Left Ventricular Morphometric Atlases Show Consistent Relationships with Cardiovascular Risk Factors: A UK Biobank Study". *Scientific Reports*, 2019. https://eprints.whiterose.ac.uk/157310/1/s41598-018-37916-6.pdf

[32] Xavier Gilbert Serra, "Anomaly Detection in Noisy Images". Dissertation, University of Maryland, 2015. https://drum.lib.umd.edu/handle/1903/18119

[33] Robert J. Gillies, *et al.*, "Radiomics: Images Are More than Pictures, They Are Data". *Radiology*, Volume 278, Number 2 (2016). https://pubmed.ncbi.nlm.nih.gov/26579733/

[34] Chang Gong, *et al.*, "A Computational Multiscale Agent-Based Model for Simulating Spatio-Temporal Tumour Immune Response to PD1 and PDL1 Inhibition". *The Journal of the Royal Society Interface*, Volume 14 (2017). https://pubmed.ncbi.nlm.nih.gov/28931635

[35] Patrick Grossmann, *et al.*, "Defining the Biological Basis of Radiomic Phenotypes in Lung Cancer". *eLife*, Volume 6 (2017). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5590809/

[36] Stephan Güttinger, "The Limits of Replicability". *European Journal for Philosophy of Science*, Volume 10, Number 2 (2020). https://core.ac.uk/download/pdf/237399625.pdf

[37] Akifumi Hagiwara, *et al.*, "Variability and Standardization of Quantitative Imaging: Monoparametric to Multiparametric Quantification, Radiomics, and Artificial Intelligence". *Investigative Radiology*, Volume 25, Number 9 (2020). https://journals.lww.com/investigativeradiology/Fulltext/2020/09000/Variability_and_Standardization_of_Quantitative.11.aspx

[38] Sara Hamis, *et al.*, "Blackboard to Bedside: A Mathematical Modeling Bottom-Up Approach Toward Personalized Cancer Treatments". *JCO Clinical Cancer Informatics*, Volume 3 (2019). https://ascopubs.org/doi/10.1200/CCI.18.00068

[39] Cameron Hassani, *et al.*, "Myocardial Radiomics in Cardiac MRI". *American Journal of Roentgenology*, Volume 214, Number 3 (2020). https://www.ajronline.org/doi/pdfplus/10.2214/AJR.19.21986

[40] Alessa Hering, *et al.*, "Enhancing Label-Driven Deep Deformable Image Registration with Local Distance Metrics for State-of-the-Art Cardiac Motion Tracking". https://arxiv.org/pdf/1812.01859.pdf

[41] Yi Hong, *et al.*, "Application of Standardized Biomedical Terminologies in Radiology Reporting Templates". *Information Services & Use*, Volume 33 (2013), pages 309-323. https://content.iospress.com/download/information-services-and-use/isu708?id=information-services-and-use/isu708

[42] Iram Tazim Hoque, *et al.*, "A Contour Property Based Approach to Segment Nuclei in Cervical Cytology Images". *BMC Medical Imaging*, Volume 21 (2021). https://bmcmedimaging.biomedcentral.com/track/pdf/10.1186/s12880-020-00533-9.pdf

[43] Leah M. Iles, *et al.*, "Histological Validation of Cardiac Magnetic Resonance Analysis of Regional and Diffuse Interstitial Myocardial Fibrosis". *European Heart Journal - Cardiovascular Imaging*, 2015, pages 14-22. https://academic.oup.com/ehjcimaging/article/16/1/14/2403445

[44] Mohammad Jafarnejad, *et al.*, "Mechanistically detailed systems biology modeling of the HGF/Met pathway in hepatocellular carcinoma". *npj Systems Biology and Applications*, 2016. https://www.nature.com/articles/s41540-019-0107-2

[45] Xia Jin, *et al.*, "Meshless Algorithm for Soft Tissue Cutting in Surgical Simulation". *Computer Methods in Biomechanics and Biomedical Engineering*, Volume 17 (2014). https://pubmed.ncbi.nlm.nih.gov/22974246

[46] David Johnson, *et al.*, "Semantically Linking In Silico Cancer Models". *Cancer Informatics*, 2014. https://journals.sagepub.com/doi/10.4137/CIN.S13895

[47] Aslı Kale and Selim Aksoy, "Segmentation of Cervical Cell Images". In *2010 20th International Conference on Pattern Recognition*, Proceedings. https://ieeexplore.ieee.org/document/5595797

[48] Yildirim Karslioğlu, *et al.*, "Chalkley Method in the Angiogenesis Research and its Automation via Computer Simulation". *Pathology - Research and Practice*, Volume 210, Number 3 (2014), pages 161-168. https://pubmed.ncbi.nlm.nih.gov/24359720

[49] Jakob Nikolas Kather, *et al.*, "Continuous Representation of Tumor Microvessel Density and Detection of Angiogenic Hotspots in Histological Whole-Slide Images". *Oncotarget*, Volume 6, Number 22 (2015), pages 19163-19176. https://pubmed.ncbi.nlm.nih.gov/26061817

[50] Pradeeban Kathiravelu and Ashish Sharma, "MEDIator: A Data Sharing Synchronization Platform for Heterogeneous Medical Image Archives". *SIGKDD 2015 BigCHat Workshop*. https://zenodo.org/record/844842#.YPoMs1NKiis

[51] Azira Khalil, *et al.*, "An Overview on Image Registration Techniques for Cardiac Diagnosis and Treatment". *Cardiology Research and Practice*, 2018. https://www.hindawi.com/journals/crp/2018/1437125/

[52] Steffen Klamt, *et al.*, "Hypergraphs and Cellular Networks". *PLoS Computational Biology*, 2009. https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1000385

[53] Márton Kolossváry, *et al.*, "Cardiac Computed Tomography Radiomics: A Comprehensive Review on Radiomic Techniques". *Journal of Thorac Imaging*, Volume 33, Number 1 (2018), pages 26-34. https://pubmed.ncbi.nlm.nih.gov/28346329/

[54] Chung-Wein Lee and Keith M. Stantz, "Development of a Mathematical Model to Estimate Intra-Tumor Oxygen Concentrations through Multi-Parametric Imaging".*BioMedical Engineering OnLine*, Volume 15 (2014). https://biomedical-engineering-online.biomedcentral.com/articles/10.1186/s12938-016-0235-5

[55] Hui Li, *et al.*, "Quantitative MRI Radiomics in the Prediction of Molecular Classifications of Breast Cancer Subtypes in the TCGA/TCIA Data Set". *NPJ Breast Cancer*, Volume 2016, Number 2. https://www.nature.com/articles/npjbcancer201612.pdf

[56] Lok Wan Lorraine Ma, "Mathematical Methods for 2D-3D Cardiac Image Registration". Dissertation, University of Ontario Institute of Technology, 2016 https://ir.library.dc-uoit.ca/xmlui/bitstream/handle/10155/756/Ma_Lok_Wan_Lorraine.pdf

[57] Robert D. Lovchik, *et al.*, "Rapid Micro-Immunohistochemistry". *Microsystems & Nanoengineering*, Volume 6 (2020). https://www.nature.com/articles/s41378-020-00205-2.pdf

[58] Meghan G. Lubner, *et al.*, "CT Texture Analysis: Definitions, Applications, Biologic Correlates, and Challenges". *RadioGraphics*, Volume 37, Number 5 (2017). https://pubs.rsna.org/doi/full/10.1148/rg.2017170056

[59] TimoMäkelä, *et al.*, "A Review of Cardiac Image Registration Methods". *IEEE Transactions on Medical Imaging*, Volume 21, Number 9 (2002), pages 1011-1021. https://pubmed.ncbi.nlm.nih.gov/12564869/

[60] Carlos Martin-Isla, *et al.*, "Image-Based Cardiac Diagnosis With Machine Learning: A Review". *Frontiers in Cardiovascular Medicine*, 2020. https://www.frontiersin.org/articles/10.3389/fcvm.2020.00001/full

[61] Jose L.V. Mejino, Jr, *et al.*, "FMA-RadLex: An Application Ontology of Radiological Anatomy derived from the Foundational Model of Anatomy Reference Ontology". *AMIA Symposium*, 2008, pages 465-469. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2656009

[62] Oleg Milberg, *et al.*, "A QSP Model for Predicting Clinical Responses to Monotherapy, Combination and Sequential Therapy Following CTLA-4, PD-1, and PD-L1 Checkpoint Blockade". *Frontiers in Cardiovascular Medicine*, 2020. https://www.frontiersin.org/articles/10.3389/fcvm.2020.00001/full

[63] Iurii S. Nagornov and Mamoru Kato, "tugHall: A simulator of cancer-cell evolution based on the hallmarks of cancer and tumor-related genes". *Bioinformatics*, Volume 36, Number 11 (2020), pages 3597-3599. https://pubmed.ncbi.nlm.nih.gov/32170925

[64] Sandy Napel, *et al.*, "Quantitative Imaging of Cancer in the Postgenomic Era: Radio(geno)mics, deep learning, and habitats". *Cancer*, Volume 124 (2018), pages 4633-4649. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6482447

[65] Anja Nilsen, *et al.*, "Reference MicroRNAs for RT-qPCR Assays in Cervical Cancer Patients and Their Application to Studies of HPV16 and Hypoxia Biomarkers". *Translational Oncology*, Volume 12, Number 3 (2019), pages 576-584. https://pubmed.ncbi.nlm.nih.gov/30660934

[66] Yangming Ou and Andreas Schuh, "DRAMMS Software Manual". https://www.cbica.upenn.edu/sbia/software/dramms/_downloads/DRAMMS_Software_Manual.pdf

[67] Yangming Ou, *et al.*, "Validation of DRAMMS among 12 Popular Methods in CrossSubject Cardiac MRI Registration". In *Workshop of Biomedical Image Registration*, Proceedings 2012, pages 209-219. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5462118/pdf/nihms861246.pdf

[68] Nikolaos Papanikolaou, *et al.*, "How to Develop a Meaningful Radiomic Signature for Clinical Use in Oncologic Patients". *Cancer Imaging*, Volume 20 (2020). https://cancerimagingjournal.biomedcentral.com/articles/10.1186/s40644-020-00311-4

[69] Iacopo Petrini, *et al.*, "ED-B Fibronectin Expression is a Marker of Epithelial-Mesenchymal Transition in Translational Oncology". *Oncotarget*, Volume 8, Number 3 (2017), pages 4914-4921. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5354880

[70] Gibin G. Powathil, *et al.*, "Systems Oncology: Towards patient-specific treatment regimes informed by multiscale mathematical modelling". *Seminars in Cancer Biology*, 2015, pages 13-20. https://pubmed.ncbi.nlm.nih.gov/24607841

[71] Sergey F. Pravdin, *et al.*, "Mathematical Model of the Anatomy and Fibre Orientation Field of the Left Ventricle of the Heart". *BioMedical Engineering OnLine*, Volume 12 (2013), pages 13-20. https://biomedical-engineering-online.biomedcentral.com/articles/10.1186/1475-925X-12-54

[72] Esther Puyol-Antón, *et al.*, "A Multimodal Spatiotemporal Cardiac Motion Atlas from MR and Ultrasound Data". *Medical Image Analysis*, Volume 40 (2017). https://www.sciencedirect.com/science/article/pii/S1361841517300890

[73] Kun Qian, *et al.*, "Energized Soft Tissue Dissection in Surgery Simulation". *Computer Animation and Virtual Worlds*, 2016. https://onlinelibrary.wiley.com/doi/abs/10.1002/cav.1691

[74] Zahra Raisi-Estabragh and Steffen E. Petersen, "Cardiovascular Research Highlights from the UK Biobank: Opportunities and challenges". *Cardiovascular Research*, Volume 116, Number 1 (2020), pages e12-e15. https://academic.oup.com/cardiovascres/article/116/1/e12/5645361

[75] Stefania Rizzo, *et al.*, "Radiomics: The facts and the challenges of image analysis". *European Radiology Experimental*, Volume 2 (2018). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6234198/

[76] Felipe Romero, "Philosophy of Science and the Replicability Crisis". *Philosophy Compass*, Volume 14, Number 11 (2019). https://onlinelibrary.wiley.com/doi/full/10.1111/phc3.12633

[77] Daisuke Sato, *et al.*, "Formation of Spatially Discordant Alternans Due to Fluctuations and Diffusion of Calcium". *PLoS One*, 2013. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0085365

[78] Abbas Shirinifard, *et al.*, "3D Multi-Cell Simulation of Tumor Growth and Angiogenesis". *PLoS One*, 2009. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0007190

[79] Elizabeth R. Smith *et al.*, "New Biological Research and Understanding of Papanicolaou's Test" *Diagnostic Cytopathology*, Volume 46, Number 6 (2018), pages 507-515. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5949091/

[80] Brita Singers Sørensen and Michael R. Horsman, "Tumor Hypoxia: Impact on Radiation Therapy and Molecular Pathways". *Frontiers in Oncology*, 2020. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0007190

[81] Jie Su, *et al.*, "Automatic Detection of Cervical Cancer Cells by a Two-Level Cascade Classification System". *Analytical Cellular Pathology*, Volume 2016. https://www.hindawi.com/journals/acp/2016/9535027

[82] Jing Rui Tang, *et al.*, "Evaluating Nuclear Membrane Irregularity for the Classification of Cervical Squamous Epithelial Cells" *PLOS One*, 2016. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5065206/pdf/pone.0164389.pdf

[83] Wen Tang and Tao Ruan Wan, "Constraint-Based Soft Tissue Simulation for Virtual Surgical Training" *IEEE Transactions on Biomedical Engineering*, Volume 61, Issue 11 (2014). https://ieeexplore.ieee.org/document/6820761

[84] Giacomo Tarroni, *et al.*, "Large-scale Quality Control of Cardiac Imaging in Population Studies: Application to UK Biobank". *Nature Scientific Reports*, 2020. https://www.nature.com/articles/s41598-020-58212-2.pdf

[85] Michal R. Tomaszewski and Robert J. Gillies, "The Biological Meaning of Radiomic Features". *Radiology*, 2021. https://pubs.rsna.org/doi/pdf/10.1148/radiol.2021202553

[86] Paul Van Liedekerke, *et al.*, "Simulating Tissue Mechanics with Agent-Based Models: Concepts, perspectives and some novel results". *Computational Particle Mechanics*, Volume 2 (2015), pages 401-444. https://link.springer.com/article/10.1007/s40571-015-0082-3

[87] Maolin Xu, *et al.*, "An Analysis of Ki-67 Expression in Stage 1 Invasive Ductal Breast Carcinoma Using Apparent Diffusion Coefficient Histograms" *Quantitative Imaging in Medicine and Surgery*, Volume 11, Number 4 (2021), pages 1518-1531. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7930667

[88] Zhihui Wang, *et al.*, "Accelerating Cancer Systems Biology Research through Semantic Web Technology", Volume 11, Number 4 (2021), pages 1518-1531. *Wiley Interdisciplinary Review of Systems Biology Medicine*, Volume 5, Number 2 (2013), pages 135-151. https://pubmed.ncbi.nlm.nih.gov/23188758

[89] Hadi Wiputra, *et al.*, "Cardiac Motion Estimation from Medical Images: A regularisation framework applied on pairwise image registration displacement fields". *Nature Scientific Reports*, Volume 10 (2020). https://www.nature.com/articles/s41598-020-75525-4.pdf

[90] Hao Zhang, *et al.*, "The Human Explanted Heart Program: A translational bridge for cardiovascular medicine". *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease*, Volume 1867, Number 1 (2021). https://www.sciencedirect.com/science/article/abs/pii/S0925443920303434

[91] Jinao Zhang, *et al.*, "Deformable Models for Surgical Simulation: A Survey". *IEEE Reviews in Biomedical Engineering*, Volume 11 (2017). https://ieeexplore.ieee.org/document/8107531

[92] Xingyu Zhang, *et al.*, "Atlas-Based Quantification of Cardiac Remodeling Due to Myocardial Infarction". *PLOSOne*, 2014. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0110243

[93] Alex Zwanenburg, *et al.*, "Image Biomarker Standardization Initiative (Reference Manual)". *arXiv*, 2019. https://arxiv.org/pdf/1612.07003.pdf

[94] Alex Zwanenburg, *et al.*, "The Image Biomarker Standardization Initiative: Standardized Quantitative Radiomics for High-Throughput Image-based Phenotyping". *Radiology*, Volume 295, Number 2 (2020). https://pubs.rsna.org/doi/10.1148/radiol.2020191145