CHAPTER 1

Imaging technology assessment

Pari V. Pandharipande and G. Scott Gazelle
Massachusetts General Hospital & Harvard Medical School, Boston, MA, USA

KEY POINTS

- The ability to define the most appropriate roles for imaging and imaging-guided interventions in medical care will be a critical goal for the future success of radiology; as such, ways to overcome current barriers to successful technology assessments must be a priority.
- A six-level hierarchical model of technology assessments, focused in diagnostic testing, provides a useful framework for understanding how research concerning a technology's worth can progress from assessments of its technical efficacy to those of its societal value.
- Randomized controlled trials in imaging are commonly not feasible; alternative research methods (e.g., observational studies leveraging large data repositories) can provide practical solutions to addressing this problem.
- In order to benefit patients most in the broader context of healthcare, all technology assessments should be initiated with an eye toward future studies that will test the technology's worth at the societal level.

The continued success of radiology, as a discipline that can contribute meaningfully to the care of patients and improvements in public health, will depend upon our ability to define the most appropriate roles for imaging in modern medical care. In the past three decades, there have been astounding advances in imaging technologies; in parallel, utilization has increased rapidly [1–3]. Despite quick dissemination of such technologies into practice, the value that they add to patient care, in many instances, has yet to be clearly established [1–3]. Three barriers to related research efforts merit specific mention [2–4]. First, because imaging technologies typically address intermediate steps in a patient's care, it can be extremely difficult to conduct studies that reliably link imaging results to patient outcomes. Second, imaging technologies evolve rapidly, which frequently obviates the relevance of studies focusing on long-term patient outcomes. Third, there are few investigators who are adequately trained to conduct rigorous

Handbook for Clinical Trials of Imaging and Image-Guided Interventions, First Edition.
Edited by Nancy A. Obuchowski and G. Scott Gazelle.

assessments of imaging and image-guided interventions—until now, research focusing on technology development has had a greater priority. In the current era of healthcare reform, however, the value of healthcare is viewed increasingly through the lens of patient outcomes and societal costs, and efforts to overcome the barriers described earlier have emerged as a priority [2–5]. In this chapter, we first provide an overview of the different levels of evidence that define technology assessments in imaging [6–10]. We then introduce three case examples; together, they highlight common approaches and challenges to technology assessments in imaging and serve as a platform for discussing key research topics that will be covered in this book.

1.1 Six levels of evidence: a model of technology assessments in imaging

Models that are commonly used to categorize evidence in medicine do not translate well to diagnostic imaging, because they are typically centered on therapy rather than diagnosis. Since the 1970s, pioneers in technology assessment and imaging have contributed to the development and refinement of a six-level hierarchical model that is specific to diagnostic technologies [6–9]. The levels include (i) technical efficacy, (ii) diagnostic accuracy, (iii) impact on diagnostic thinking, (iv) impact on therapeutic planning, (v) impact on patient outcomes, and (vi) impact on societal outcomes [2, 6–10]. The model is hierarchical with respect to the scope of the evidence's impact: its spectrum "begins" with technical efficacy (*Level 1*) and "ends" with societal outcomes (*Level 6*). The model is also hierarchical with respect to the potential of a technology's impact: it implies, for example, that a given technology needs to demonstrate a certain level of diagnostic accuracy (*Level 2*) in order to produce favorable health and economic outcomes at the societal level (*Level 6*). However, a technology's success at one level does not necessarily predict its success at higher levels; for example, a highly accurate technology may not meaningfully—or positively—affect patient outcomes [6, 11]. Each level and the types of research appropriate to evaluate efficacy at that level are described below.

1.1.1 Level 1: technical efficacy

Technical efficacy is judged by measurements that pertain to image generation (in diagnostic imaging) and procedural feasibility (in interventional radiology) [6, 7, 11]. These measurements are preclinical; they serve as the basis for continual, iterative improvements until the technology is ready for testing and use in clinical practice [6, 7, 11]. Commonly applied metrics may include spatial and contrast resolution (for diagnostic techniques) and the precision of catheter, needle, or probe localization (for interventional techniques). A review of technical efficacy

parameters is commonly required by regulatory authorities—such as the US Food and Drug Administration—prior to a technology's dissemination into practice.

1.1.2 Level 2: diagnostic accuracy

Diagnostic accuracy is judged by multiple metrics of test performance. Here, we will discuss the three most common: (i) sensitivity and specificity, (ii) positive and negative predictive value (PPV and NPV, respectively), and (iii) receiver-operating characteristic (ROC) curves [11–14]. Each reflects a different window into the ability of a technology to detect disease. Sensitivity and specificity reflect a technology's ability to detect true positive (among disease-positive) cases and true negative (among disease-negative) cases within a study population. PPV and NPV also reflect the technology's ability to detect true positive and true negative cases but instead among *test positive* and *test negative* cases, respectively. Therefore, PPV and NPV—unlike sensitivity and specificity—depend heavily on disease prevalence within the study population. As a result, study-derived estimates of PPV and NPV are specific to populations with similar disease prevalence as the study population.

ROC curves, which plot sensitivity (y-axis) against 1-specificity (x-axis), provide insight into the trade-offs that occur when using different thresholds for calling a test "positive" [11, 13, 14]. When studies include data from multiple readers and/or multiple sites of care, a summary ROC curve may be generated, providing insights into the capabilities and limitations of an imaging technology that are more generalizable. (See Chapter 9 for a detailed discussion of ROC curves.)

Research evaluating the diagnostic accuracy of imaging technologies is frequently challenged by the identification of a suitable reference standard. Here, we highlight three common predicaments [11]; further details can be found in Chapter 7. First, in certain circumstances, an imaging test result may be less likely to be verified (e.g., undergo follow-up biopsy) than others. This can occur, for example, when a positive imaging test is typically verified in routine care, but negative or indeterminate results are not. In such situations, the resulting "verification bias" could distort corresponding estimates of test performance. Second, imaging technologies sometimes can detect more underlying pathology than a commonly accepted reference standard. Consider, for example, a patient who undergoes a liver biopsy for the detection of hepatic steatosis. The biopsy (reference standard) may be falsely negative, due to sampling of an unaffected area of the liver; however, imaging of the whole liver may more accurately reveal patchy areas of fatty infiltration throughout the liver. Using such a reference standard (a so-called tarnished gold standard) can result in inappropriately low estimates of sensitivity and specificity. Third, in many scenarios, the only reasonable reference standard is patient outcomes based on long-term follow-up, for example, disease progression or regression. Given the large number of additional tests and treatments that an individual may experience during follow-up, it can be very difficult

to attribute long-term outcomes to the results of a given imaging test. When defining a reference standard in a study of diagnostic accuracy, investigators typically must address one or more of the mentioned obstacles or, at minimum, understand how such obstacles may affect their results.

1.1.3 Levels 3 and 4: impact on diagnostic thinking and therapeutic planning

The impact of imaging technologies on diagnostic thinking (Level 3) is generally assessed by determining changes in physicians' diagnoses and/or their diagnostic confidence following the performance of an imaging study [2, 6, 7, 10, 15–17]. Moving progressively "closer" to patient outcomes, Level 4 studies measure the impact of imaging technologies on therapeutic planning (Level 4) by judging a technology's effect on patient management [2, 6, 7, 10, 15, 16]. The principal challenge in obtaining such measurements is knowing what would have happened, under real clinical conditions, had the imaging test not been performed. To determine this, hypothetical questions are generally posed to physicians, such as, "*what would you have done if this imaging test were unavailable*?" [15, 16]. These hypothetical questions are known to be cognitively burdensome and subject to various biases; nevertheless, they can also provide information that is critically important for assessing the value of new imaging technologies. The National Oncologic PET Registry (NOPR) provides one of the best examples of how this information can be requested, reported, and used to gain insight into the projected value of a new imaging technology [18, 19].

1.1.4 Level 5: patient outcomes

Patient outcomes that are associated with imaging can be evaluated with a wide spectrum of measures [2–5]. Metrics may address short- or immediate-term outcomes (e.g., rates of allergic reactions, complications, or postprocedural mortality) or long-term outcomes (e.g., rates of overall survival, disease-specific survival, progression-free survival, or life expectancy and quality of life). As mentioned, patient-level outcomes are difficult to measure in a way that can be confidently attributed to imaging. To reliably measure patient outcomes associated with imaging, randomized controlled trials (RCTs) are typically needed. In such studies, patients are randomized, for example, to undergo an imaging technology under investigation versus the standard of care [20–23]. Because differences in outcomes between the study groups are commonly small, large sample sizes are often required to achieve adequate power. This, combined with the need to conduct long-term follow-up, can make imaging-related RCTs prohibitively expensive. Nevertheless, when a substantial proportion of the population may be affected, the information yielded may justify such costs; the National Lung Screening Trial [20] exemplifies one such circumstance among several [20–23].

Additional study designs permit outcomes evaluations and, in many cases, may represent a more practical and feasible approach. Two merit specific

mention. First, **population-level observational studies** that leverage "big data"—for example, data repositories or registries that are large enough to require specialized programming and analytics for their analysis—may provide a new window into imaging-related outcomes [4]. While such research is still subject to typical biases of observational studies, it may offer new opportunities to evaluate relationships between imaging technologies and patient outcomes, particularly in settings where such analyses were not previously feasible due to limitations of study power. Second, "pragmatic" trial designs are gaining traction, though they are in their infancy as an outcomes research method [4, 24]. **Pragmatic trials** are those that attempt to deliver a low-burden intervention to a wide variety of practice settings, with less restrictive inclusion criteria as compared to typical RCTs [4, 24]. Data collection is typically passive, occurring, for example, through automated collection of specific outcomes that are already recorded in participating centers' electronic medical record systems [4]. The primary advantage of pragmatic trials over traditional RCTs is their greater generalizability to "real-world" settings; the primary disadvantage is that because of their less stringent study designs, they are more vulnerable to bias [4, 24].

1.1.5 Level 6: societal efficacy

Societal efficacy is judged by whether an imaging technology represents a good use of societal resources. By definition, societal efficacy must account for both health outcomes (clinical effectiveness) and economic outcomes (costs) associated with the use of an imaging modality [25–27]. Ideally, costs and effectiveness should be evaluated using a societal perspective and a lifetime horizon [25–27]. Using a societal perspective to estimate costs is important because resource allocation decisions are often made at the societal level. Using a lifetime horizon for estimating costs and effectiveness is also important because in many circumstances, the risks, benefits, and costs of imaging technologies are strongly influenced by the downstream care pathways that occur as a result of the imaging test result(s). Put another way, the health and economic outcomes associated with subsequent care, rather than the imaging technology itself, strongly influence the value of imaging. Most analyses of societal efficacy employ cost-effectiveness analysis, adhering closely to the guidelines put forth by the Panel on Cost-Effectiveness in Health and Medicine, convened in the 1990s by the US Public Health Service [25–27].

In the United States, a consistent challenge of cost-effectiveness analysis, from the standpoint of policymakers, pertains to the interpretation of results [28–30]. In a standard cost-effectiveness analysis, the primary results are expressed in terms of one or more incremental cost-effectiveness ratios (ICERs) that are associated with corresponding diagnostic testing and/or treatment options. An ICER represents the additional cost required to achieve an additional unit of outcome (commonly life-years or quality-adjusted life-years). In theory, whether or not a technology is an efficient use of resources is judged by

comparing such ICERs to a societal willingness-to-pay (WTP) threshold, which represents the price that a society is willing to pay to increase life expectancy (or quality-adjusted life expectancy) in a target population [25–27]. In the United States, there is no "hard" threshold beyond which a technology is considered too expensive, a circumstance that often challenges the interpretation of ICERs resulting from cost-effectiveness analyses [28–30]. Recently, a $100,000–$150,000 WTP threshold has been suggested as a guideline for US analyses, but this value will undoubtedly be subject to debate and revision in future years, to account for inflation and an evolving economic and healthcare climate [30]. See Chapter 11 for further details.

1.2 Case examples: introducing technology assessment methods into study design

Technology assessments that correspond to each of the described levels of evidence present investigators with common challenges, several of which have been discussed earlier. As in any discipline, however, the best way to know and anticipate such challenges is to gain experience in doing the research itself. In lieu of this—as a proxy—it is worthwhile to think through the design of typical assessments of imaging technologies. Drawing from the principles presented earlier, we introduce three hypothetical areas of research inquiry, selected to span both diagnostic and interventional imaging technologies. The examples illustrate how technology assessment research methods may be applied in a variety of clinical settings and serve as a platform for discussions throughout the remainder of this book.

Case example 1.1 A New Liver Blood-Pool Agent for Detecting Colorectal Metastases at CT: Evaluation of Diagnostic Accuracy

Unlike in many other cancer settings, isolated liver metastases in patients with colorectal cancer may be treated for cure [31]. As a result, accurate detection of liver metastases from colorectal cancer is an important component of treatment planning in these patients. Suppose that a hypothetical blood-pool agent ("Agent A") is developed specifically for the identification of colorectal metastases in the liver at CT and that its safety in human subjects has been established. Whether or not Agent A is superior in detecting liver metastases, relative to the current standard of care (e.g., CT with Agent B), is unknown. Readers may consider what types of study designs could *(i) determine whether Agent A can more accurately identify liver metastases than Agent B, (ii) identify appropriate thresholds for positivity when using Agent A (vs. Agent B) as a CT contrast agent, and (iii) determine whether the use of Agent A can result in improved patient outcomes or more efficient use of societal resources than the current standard of care.*

Case example 1.2 A New Biologic for Metastatic Breast Cancer: Evaluation of Treatment Response with ^{18}F-FDG-PET Imaging

Consider a new type of drug—Biologic A—that is found to be highly effective in the treatment of a subset (20%) of patients with metastatic breast cancer. Unfortunately, no test is available that allows for *a priori* determination of whether a given patient will respond. Biologic A carries a small risk (1% per year) of death from severe immunosuppression, and its cost exceeds $100,000 per month. As a result, there is substantial interest in identifying patients, as early as possible in the course of their therapy, who are likely to be responders. At ^{18}F-FDG-PET imaging, uptake of ^{18}F-FDG correlates with tumor viability, and decreased uptake following treatment generally correlates with tumor responsiveness, but the thresholds for uptake in separating respondents from nonrespondents are imperfect. Moreover, some individuals declare themselves as respondents 1–2 months later than others. Readers may consider what types of study designs could *(i) determine the extent to which FDG-PET is capable of distinguishing responders from nonresponders and (ii) identify optimal thresholds—guided by patient outcomes—for distinguishing these populations*.

Case example 1.3 Minimally Invasive Therapies for Early Lung Cancer: A Comparison of Imaging-Guided Radiofrequency Ablation versus Stereotactic Ablative Radiotherapy

In certain situations, early lung cancers may be identified in patients who cannot undergo surgery, for example, due to advanced age or multiple comorbidities. In these circumstances, minimally invasive approaches may offer an opportunity for treatment and cure. An increasing number of options are available [32–34]. In one option, imaging-guided radiofrequency ablation (RFA), a radiologist inserts a probe into the tumor under imaging guidance and then delivers radiofrequency energy to "kill" tumor cells [32, 33]. In another option, stereotactic ablative radiotherapy (SABR), a radiation oncologist delivers a highly localized, therapeutic radiation dose to treat the tumor [34]. Each option has different risks and benefits and a different profile of complications. Readers may consider what types of study designs could *(i) determine and compare the clinical efficacy of each treatment, (ii) measure and compare quality of life associated with each treatment, and (iii) determine the relative cost-effectiveness of the different treatments*.

1.3 Conclusion

In this chapter, we have provided a broad introduction to the assessment of imaging technologies, reviewing a six-level hierarchical model for categorizing evidence in diagnostic imaging [6–9]. In addition, we have provided hypothetical areas of research inquiry, to set the stage for further discussions of research design and methods in subsequent chapters (e.g., ROC curves, survival estimates, quality-of-life evaluation, etc.). When considering the hierarchy and the clinical examples presented, our primary take-home point is the following: while robust studies of an imaging technology's technical efficacy and diagnostic accuracy are

important, they are insufficient in demonstrating the technology's value to patients. In order to benefit patients most in the broader context of healthcare, technology assessments—*at all levels*—should be initiated with an eye toward future studies that will test the technology's worth at the societal level.

References

1 Smith-Bindman, R., D.L. Miglioretti, and E.B. Larson, Rising Use Of Diagnostic Medical Imaging In A Large Integrated Health System. *Health Aff*, 2008. **27**(6): p. 1491–1502.
2 Gazelle, G.S., et al., A framework for assessing the value of diagnostic imaging in the era of comparative effectiveness research. *Radiology*, 2011. **261**(3): p. 692–698.
3 Pandharipande, P.V. and G.S. Gazelle, Comparative effectiveness research: what it means for radiology. *Radiology*, 2009. **253**(3): p. 600–605.
4 Lee, C.I. and J.G. Jarvik, Patient-centered outcomes research in radiology: trends in funding and methodology. *Acad Radiol*, 2014. **21**(9): p. 1156–1161.
5 Carlos, R.C., et al., Patient-centered outcomes in imaging: quantifying value. *J Am Coll Radiol*, 2012. **9**(10): p. 725–728.
6 Fryback, D.G. and J.R. Thornbury, The efficacy of diagnostic imaging. *Med Decis Making*, 1991. **11**(2): p. 88–94.
7 Thornbury, J.R., Eugene W. Caldwell Lecture. Clinical efficacy of diagnostic imaging: love it or leave it. *AJR Am J Roentgenol*, 1994. **162**(1): p. 1–8.
8 Fineberg, H.V., Evaluation of computed tomography: achievement and challenge. *AJR Am J Roentgenol*, 1978. **131**(1): p. 1–4.
9 Fineberg, H.V., R. Bauman, and M. Sosman, Computerized cranial tomography. Effect on diagnostic and therapeutic plans. *JAMA*, 1977. **238**(3): p. 224–227.
10 Pearson, S.D., et al., Assessing the comparative effectiveness of a diagnostic technology: CT colonography. *Health AFF (Millwood)*, 2008. **2008**(27): p. 6.
11 Gazelle, G.S., Cost-Effectiveness of Imaging and Surgery in Patients with Colorectal Cancer Liver Metastases. 1999, Harvard University, U.S. Copyright # TX 4-973-835.
12 McNeil, B.J. and S.J.Adelstein, Determining the value of diagnostic and screening tests. *J Nucl Med*, 1976. **17**(6): p. 439–448.
13 Goodenough, D.J., K. Rossmann, and L.B. Lusted, Radiographic applications of receiver operating characteristic (ROC) curves. *Radiology*, 1974. **110**(1): p. 89–95.
14 McNeil, B.J. and J.A. Hanley, Statistical approaches to the analysis of receiver operating characteristic (ROC) curves. *Med Decis Making*, 1984. **4**(2): p. 137–150.
15 Abujudeh, H.H., R. Kaewlai, and P.M. McMahon, Abdominopelvic CT increases diagnostic certainty and guides management decisions: a prospective investigation of 584 patients in a large academic medical center. *AJR Am J Roentgenol*, 2011. **196**(2): p. 238–242.
16 Esses, D., Birnbaum, A., Bijur, P., Shah, S., Gleyzer, A., Gallagher, E.J., Ability of CT to alter decision making in elderly patients with acute abdominal pain. *Am J Emerg Med*, 2004. **22**(4): p. 270–272.
17 Nagurney, J.T., Brown, D.F., Chang, Y., Sane, S, Wang, A.C., Weiner, J.B., Use of diagnostic testing in the emergency department for patients presenting with non-traumatic abdominal pain. *J Emerg Med*, 2003. **25**(4): p. 363–371.
18 Hillner, B.E., et al., The National Oncologic PET Registry (NOPR): design and analysis plan. *J Nucl Med*, 2007. **48**(11): p. 1901–1908.
19 Aberle, D.R., et al., Reduced lung-cancer mortality with low-dose computed tomographic screening. *N Engl J Med*, 2011. **365**(5): p. 395–409.

20 Jarvik, J.G., et al., Rapid magnetic resonance imaging vs radiographs for patients with low back pain: a randomized controlled trial. *JAMA*, 2003. **289**(21): p. 2810–2818.
21 Hillner, B.E., et al., Impact of positron emission tomography/computed tomography and positron emission tomography (PET) alone on expected management of patients with cancer: initial results from the National Oncologic PET Registry. *J Clin Oncol*, 2008. **26**(13): p. 2155–2161.
22 Smith-Bindman, R., et al., Ultrasonography versus computed tomography for suspected nephrolithiasis. *N Engl J Med*, 2014. **371**(12): p. 1100–1110.
23 Hoffman, U., et al., Coronary CT angiography versus standard evaluation in acute chest pain. *N Engl J Med*, 2012. **367**(4): p. 299–308.
24 Lurie, J.D. and T.S. Morgan, Pros and cons of pragmatic clinical trials. *J Comp Eff Res*, 2013. **2**(1): p. 53–58.
25 Hunink, M.G., Outcomes research and cost-effectiveness analysis in radiology. *Eur Radiol*, 1996. **6**(5): p. 615–620.
26 Weinstein, M.C., et al., Recommendations of the Panel on Cost-effectiveness in Health and Medicine. *JAMA*, 1996. **276**(15): p. 1253–1258.
27 Gazelle, G.S., et al., Cost-effectiveness analysis in the assessment of diagnostic imaging technologies. *Radiology*, 2005. **235**(2): p. 361–370.
28 Winkelmayer, W.C., et al., Health economic evaluations: the special case of end-stage renal disease treatment. *Med Decis Making*, 2002. **22**(5): p. 417–430.
29 Neumann, P.J. and M.C. Weinstein, Legislating against use of cost-effectiveness information. *N Engl J Med*, 2010. **363**(16): p. 1495–1497.
30 Neumann, P.J., J.T. Cohen, and M.C. Weinstein, Updating cost-effectiveness—the curious resilience of the $50,000-per-QALY threshold. *N Engl J Med*, 2014. **371**(9): p. 796–797.
31 Chan, K.M., et al., Outcomes of resection for colorectal cancer hepatic metastases stratified by evolving eras of treatment. *World J Surg Oncol*, 2011. **9**: p. 174.
32 de Baere, T., G. Farouil, and F. Deschamps, Lung cancer ablation: what is the evidence? *Semin Interv Radiol*, 2013. **30**(2): p. 151–156.
33 Ridge C.A., S.B.Solomon, and R.H.Thornton, Thermal ablation of stage I non-small cell lung carcinoma. *Semin Interv Radiol*, 2014. **31**(2): p. 118–124.
34 Padda, S.K., et al., Early-stage non-small cell Lung cancer: surgery, stereotactic radiosurgery, and individualized adjuvant therapy. *Semin Oncol*, 2014. **41**(1): p. 40–56.