# 2013-2014: Evolving Challenges for Value-Prices LDTs

FOLEY HOAG LLP

September 2013

Bruce Quinn MD PhD
Senior Health Policy Specialist
Foley Hoag LLP
bquinn@foleyhoag.com
www.foleyhoag.com
617 832 1291

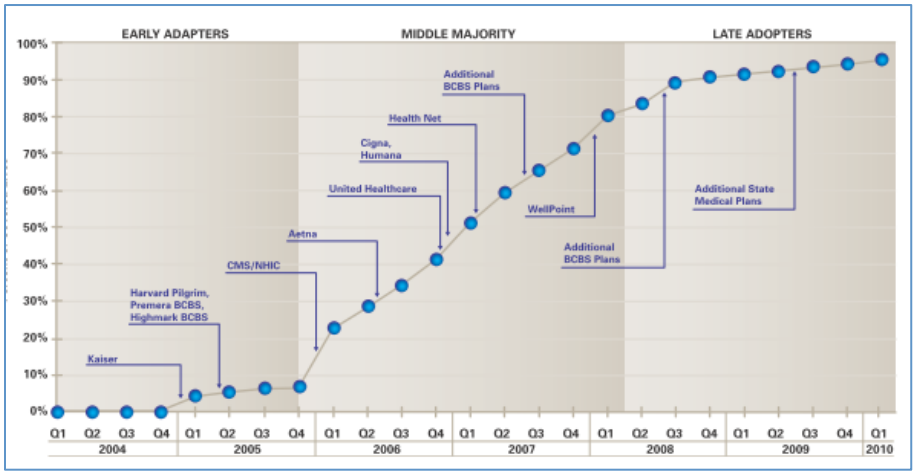
# EVOLVING CHALLENGES FOR VALUE-PRICED LDTs

In 2004, the Genomic Health Oncotype DX breast cancer molecular test was launched and within five years had substantially changed the standard of care for patients with early breast cancers. How has the marketplace changed in the past ten years, and what can companies expect when launching a new value-priced lab test today? Many conferences and publications have discussed "clinical utility" – how it should be defined, how it should be assessed, and what standards should be applied. However, there are many other changes in the payer industry and laboratory industry and in our healthcare system as a whole. This white paper examines major trends that directly impact the laboratory industry from the viewpoint of a startup company with a value priced diagnostic test.

# AN INFORMAL LANDSCAPE FOR 2013/2014:

# EVOLVING CHALLENGES FOR VALUE PRICED LDTs

The successful national launch of the Genomic Health Oncotype DX test was a landmark event in the evolution of the U.S. laboratory industry. The test was commercialized (with collaboration from NIH investigators), priced at over $3000, and offered the potential to substantially impact the standard of care pathway for small breast cancers in patients nationwide. The Oncotype DX test was not packaged as a commercial kit, but offered as a sole-source, very complex laboratory-developed test (LDT). As documented in a 2010 white paper by the consultancy Health Advances, the road from its introduction to coverage for 80% of U.S. insurance-covered lives took years - from the market launch in 2004 (accompanied by a New England Journal publication) until early 2008.[1] By the end of 2012, the company was profitable with income on the order of $10M per year against revenue of some $200M per year.[2]




[1] Figure 7 in: Health Advances, 2010, The reimbursement landscape for novel diagnostics. 40pp.
[2] http://investor.genomichealth.com/releasedetail.cfm?ReleaseID=719474

This often-cited chart of Oncotype DX's pathway to payer acceptance shows us the payer adoption rate for a test launched almost ten years ago, in 2004. What will the adoption path look like for new tests to be launched in 2013 or 2014?

This informal white paper describes some ongoing as well as emerging challenges for value-based diagnostics. Algorithmic Multi-analyte tests continued to be offered most commonly by commercial laboratories, either (a) as a sole product or as (b) flagship tests as part of a broader portfolio of molecular laboratory and pathology services.[3] Next generation sequencing tests for hereditary diseases and cancer are now offered by a range of stand-alone companies and by some academic medical centers.[4] A different and emerging application of next-generation or advanced sequencing tests is Trisomy 21 prenatal detection.[5]

There are some major signs of good news in the field of proprietary testing. For example, since early 2013, some major insurers have provided coverage for cell-free DNA testing for Trisomy 21. Notably, this coverage was announced shortly after a health association recommendation endorsed this technology.[6]

On the other hand, the ultimate fate of most new high-value tests that to be introduced in 2013/2014 will not be known for certain until several years have passed. Absent such certainty, this white paper lays out some of the currently active issues and considerations. In addition to this white paper, parallel sources of information include a 2012 Institute of Medicine report[7], a proposals toward more rational mechanisms of evidence generation through payment policy[8], and an increasing number

---

[3] Examples of the flagship test model include the Mammostrat breast cancer test at Clarient.

[4] For an academic center see, e.g., Heger M (2012) After a year of lessons learned, Michigan looks to expand comprehensive clinical cancer sequencing. Genomeweb.com (12/19/2012). For a leading commercial supplier, see http://www.foundationmedicine.com/

[5] Weaver C (2013) Tough calls on prenatal tests. Wall Street Journal, 4/3/2013. Sayres LC et al. (2012) Integrating stakeholder perspectives into the translation of cell-free fetal DNA testing for aneuploidy. Genome Med 4:49. Devers PL et al. (2013) J Genet Couns Noninvasive prenatal testing/noninvasive prenatal diagnosis: the position of the National Society of Genetic Counselors. 22:291-5. Dugoff L (2012) Application of genomic technology in prenatal diagnosis. NEJM 367:2249-51. Morain S et al. (2013) New era in noninvasive prenatal testing. NEJM 369:499-501.

[6] Genomeweb (12/19/2012) Wellpoint suggests it may cover Sequenom's 'MaterniT21 Plus,' Other non-invasive fetal aneuploidy tests. See also: A representative BCBS policy, http://scmedpolicies.wellpoint.com/medicalpolicies/policies/mp_pw_c150729.htm . See also: American College of Obstetrics and Gyneology (11/20/2012): Revised recommendation. [See e.g. Sequenom press release, 11/21/2012). See also: At least four companies work in this space, some with legal conflicts against one another: Hayden EC (6/27/2012) Fetal tests spur legal battle. A newborn industry based on non invasive genetic testing turns combative. http://www.nature.com/news/fetal-tests-spur-legal-battle-1.10894

[7] Institute of Medicine (2012) Genome-based diagnostics: Clarifying pathways to clinical use: Workshop summary. 105pp. See also: Faruki H & Lai-Goldman M (2010) Application of a pharmacogenetic test adoption model to six oncology biomarkers. Pers Med 7:441-50.

[8] Schulman KA & Tunis SR (2009) A policy approach to the development of molecular diagnostic tests. Nat Biotech 28:1157-9. Messner DA & Tunis SR (2012) Current and future state of FDA-CMS parallel reviews. Clin Pharm Ther 91:383-385. Lindor RA et al. (2013) Regulatory and reimbursement innovations for molecular diagnostics: parallel

of peer-reviewed policy studies of the payer adoption process.[9] The range of challenges that new tests face are understood by companies or academic labs that are actually immersed in the launch process, and a full perspective must bring together both the birds'-eye view and the ground-level view of the adoption and reimbursement process.

Business strategies always involve complex patterns of risks and mitigations. This document may seem to focus more on "risks" – for example, the increasing difficulty of thriving as an out-of-network lab. The "mitigations" to the issues discussed here are always company-specific, and emerge out of a careful analysis of the company's (or institution's) capabilities, timeline, business goals, and other considerations. To a large degree, the "risks" are more universal than the mitigations, and are more easily laid out in a summary white paper like this one.

---

review and coverage with evidence development. Sci Transl Med 5:176cm3. Hayes DF et al. (2013) Breaking a vicious cycle [tumor-based diagnostics]. Science Transl Med5:196cm6.

[9] Trosman JR et al. (2010) Coverage policy development for personalized medicine: private payer perspectives for the 21-gene assay. J Oncol Practice 6:238-42. See also: Deverka PA et al. (2012) Stakeholder assessment of the evidence for cancer genomic tests: insights from three case studies. Genet Med 14:656-62. From a company perspective: Pambianco D et al. (2012) Utility before profitability: The evolving evidence paradigm for molecular diagnostics. In Vivo (9/2012). Faulkner et al. (2012) Challenges in the Development and Reimbursement of Personalized Medicine—Payer and Manufacturer Perspectives and Implications for Health Economics and Outcomes Research: A Report of the ISPOR Personalized Medicine Special Interest Group. Value in Health 15:1162-71. For an international perspective: Personalized Medicine Coalition/Garfield S (2012) Advancing access to personalized medicine: A comparative assessment of European reimbursement systems. PMC, link at: http://www.personalizedmedicinecoalition.org/sites/default/files/files/PMC_Europe_Reimbursement_Paper_Final.pdf

**THE INSURANCE SYSTEM AND HIGH VALUE NOVEL DIAGNOSTICS**

1. Getting Your Claim In the Door
2. The Remote Payer
3. Pre-emptive Non Coverage
4. Risks of Physician Incentives in ACO's

**THE EVIDENCE ASSESSMENT SYSTEM AND HIGH VALUE NOVEL DIAGNOSTICS**

5. The Need for Evidence is Naturally High – It's No One's Fault
6. Technology Assessment Can be Biased against Diagnostic Tests
7. The Dossier – A Challenging Document
8. Value Pricing and Its Challenges

**EPILOG:**

9. It's Never Been Easy, But It's Worth It: Policymakers and Diagnostic Tests

**THE INSURANCE SYSTEM AND HIGH VALUE NOVEL DIAGNOSTICS**

1. Getting Your Claim In the Door
2. Pre-emptive Non Coverage
3. The Remote Payer
4. Risks of Physician Incentives in ACO's

## 1. GETTING YOUR CLAIM IN THE DOOR

Claims are the means by which providers and payers exchange information. The format and the codes of these (mostly electronic) documents are controlled by national regulations. For the laboratory provider, the claim is both a "statement of work" and an "invoice" to the third-party payer.

The barriers to filing a claim are getting higher and more complicated to navigate. For example, for many years a lab could submit a large proportion of claims to either (A) its local Medicare contractor (for all Medicare beneficiaries) or to (B) its local BCBS plan, for all BCBS beneficiaries. With these two registrations, the lab had access to roughly 150 million covered lives.

Regarding Medicare, today, over 20% of Medicare patients are in managed care plans that require individual claims submissions. And regarding the Blues plans, which cover about 100 million Americans,[10] the claims submission process for laboratories has rapidly become much more fragmented. For labs, through 2012, a "Blues Card" was issued by one of the 38 plans but accepted by providers (doctors, hospitals) enrolled in any local plan. This provided mobility and national coverage for the typical BCBS patient, and it allowed a lab to enroll with one regional Blues plan and submit claims for services to any Blues patient. In October 2012, BCBS suspended the "Blues Card" for laboratory services, and requires the performing lab to enroll in each of several dozen Blues plans nationwide, corresponding to where any particular BCBS patient lives and gets his/her BCBS insurance.[11] [12] The purpose of the new rules for laboratory claims is to let a particular Blues plan contract more aggressively with a limited number of in-network laboratories, and to benefit those in-network laboratories by excluding out-of-network laboratories. In short, it now takes *over 50* individual payer registrations, relationships, and contracts (traditional Medicare, 38 Blues plans, and 5-10 different Medicare Advantage plans) instead of two, to access the same 150M covered lives.

For both Blues plans and other commercial plans, payers are erecting much higher barriers to getting paid for the claim. Laboratories have long been classified by a health plan as either "in network" and "out of network." But increasingly, the plan may try to essentially cut off benefits for out-of-network tests, and harshly warn doctors that order tests from your lab that they are ordering "out-of-network"

---

[10] http://en.wikipedia.org/wiki/Blue_Cross_Blue_Shield_Association Accessed 1/2013.

[11] For one BCBS plan public explanation see: http://www.wa.regence.com/provider/blue-card-program/filing-claims/lab-dme-rx.html In contrast to "independent" (non hospital) labs, hospital labs may continue to use the Blues card system, e.g. enroll with only their state Blues plan but submit claims for any nationwide BCBS patient.

[12] http://www.thestreetsweeper.org/ckfinder/userfiles/files/RegionalLabBlueCardLawsuit.pdf

tests, which could endanger the doctor's continued participation in the health plan. Another tactic is to pay the out-of-network lab, but the payments made to such a lab may be a small fraction of the lab's charge. Finally, if your lab is an out-of-network plan, the payer may issue a check in the name of the patient, mailed to the patient's home, for the patient to cash. This check itself could reflect a 50-80% discount from the lab's original charge. The lab must now identify the patient's home address, and reach out by mail, by telephone, or through collection agencies to recoup money from the patient, taking the collection process far from the world of simple electronic healthcare transactions and drastically raising collection costs while cutting the realization rate on accounts receivable.

If being an out-of-network plan is an increasingly cold-shoulder environment, should a new lab work to become an "in-network lab" at as many health plans nationwide as possible? In attempting to do so, the lab may find the plan is not accepting any more in-network labs, and could even be excluding formerly participating labs from renewal to the next year's in-network contract. And if your lab does succeed in coming in-network, you are signing a contractual agreement to follow the plan's fee schedules and follow the plan's coverage decisions – which may include a coverage decision to *never* pay for your test.

This process may be equally daunting for smaller academic medical centers that might attempt to develop a national specialty-test outreach program. The largest labs bring economies of scale to the payer contracting problem, while smaller entities may seek assistance if they can outsource the billing function to a growing number of billing and payer-relations consultancies.

**2. "PRE-EMPTIVE NONCOVERAGE" – A new term of art?**

Most large private plans in the U.S. publish public coverage policies.[13] Sometimes these are "pre-emptive non coverage," a mechanism through which the insurers are able to state publicly that certain devices, procedures, or services are non-covered as "investigational."

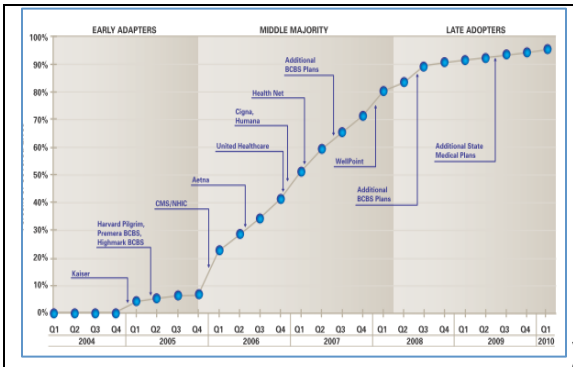
By 2012, such "pre-emptive non coverage" could appear:

- even before a product's FDA approval, or
- upon FDA approval, or
- market launch.

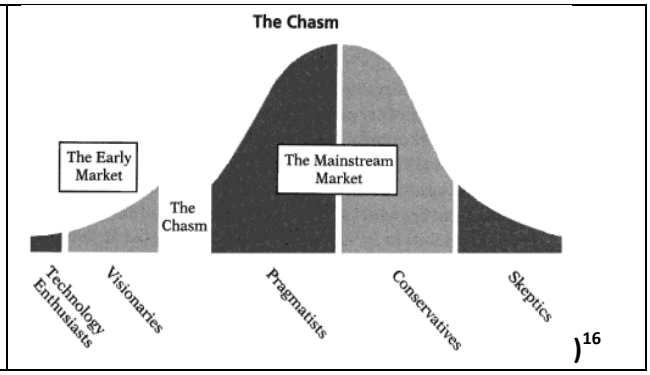
What implications does this have for the conventional market adoption cycle? The adoption of the Genomic Health Oncotype DX test seemed to be initially slow, then steadily rising through most of the

[13] These policies are due in part to a class action settlement around 2002. See for example: http://www.ama-assn.org/amednews/2002/12/16/prl11216.htm. I thank Dr. Bob McDonald, Aledo Consulting, for this history. The private payer published policies are part of the "technology assessment era": whereas 20 years ago, the claim would be summarily denied as "investigational," today a published policy reviews and then classifies the provider's service as investigational. I believe the 2002 requirement to publish coverage policies at private payers led to a need to write longer, more detailed, and more evidence based policies which require updating.

market, then trailing off due to late adopters. (Left figure, shown below). This is the classic adoption curve for new processes or technologies.[14] However, in his book "Crossing the Chasm," Geoffrey Moore proposed and inserted a substantial "gap" between early adopters and the rest of the marketplace – the "chasm." (Right figure, shown below.) From 2008 through the present, Medicare (in its local contractor programs, particularly the Palmetto MOLDX program) has often been an "early adopter." The question for labs in 2013 and in years forward is whether the bulk of the market will step up to coverage relatively quickly after local Medicare coverage (the conventional diffusion model), or else, whether Medicare LCD decisions will appear more as "early adopter" actions that are followed by a "chasm" that needs to be crossed using deliberate and different strategies (the Moore diffusion model).




[15]




[16]

One sign of a "chasm" rather than a continuity between a Medicare LCD decision and commercial adoption is the statement of Cigna medical director, that while the insurer may be aware of Medicare coverage decisions, they are not binding on Cigna, and the company makes its own evidence review and decision.[17]

### 3. THE REMOTE PAYER

Regardless of the quality of a company's trials, its potential clinical impact in clinical use, and the quality of its dossier or presentation, there are five or ten large insurers in the United States, another couple dozen substantial ones, and then a long tail of hundreds of smaller insurers.

---

[14] Everett Rodgers' work on diffusion of innovations dates back to the 1960s, with predecessor work tracing to the 1890s. http://en.wikipedia.org/wiki/Diffusion_of_innovations and Rodgers EM (2003) Diffusion of Innovations, 5th Ed., Free Press.

[15] Figure 7 in: Health Advances, 2010, The reimbursement landscape for novel diagnostics. 40pp. The early/middle/late model is found in Rodgers E (1962 and later editions) Diffusion of Innovation.

[16] http://news.cnet.com/8301-13505_3-20005119-16.html The "chasm" was popularized by Moore G (1991, 1999) Crossing the Chasm: Marketing and selling disruptive products to mainstream customers. Harper Business.

[17] Dark Report (8/19/2013). Cigna program addresses genetic test utilization. "Contrary to what some people believe, Cigna does not always follow what Medicare does with regard to covering genetic tests." (Dr David Finley, national Medical Officer, Cigna). One could also compare the coverage position of Palmetto MOLDX in CY2012 decisions to the status of the same tests' coverage in major commercial plans (available on the internet, see FN 13) in mid 2013. This exercise is left to the reader.

*At any or all of these health plans*, the decision-makers can be extremely hard to reach. Mail and email may go unanswered, letters seemingly unread, dossiers unread.

In short, no matter how well done the science, how careful the dossier, it may be impossible to get a hearing with virtually any of the U.S. payers, and there are a lot of them.

**4. NEW COST CONTROL MODELS (ACO'S) - BIAS AGAINST VALUE PRICED TESTS**

Medicare -- and commercial payers -- are promoting Accountable Care Organizations in which hospitals and doctors are ***collaboratively incented*** to reduce a patient's total per-year costs. [18]

To give a highly simplified example for discussion purposes, imagine a health plan contract where for every $1000 less that is spent, relative to a predicted level, the healthcare providers receive a rebate of $100 from the insurance plan. This 10% rebate could be very attractive, since hospitals might have only a 5% margin on $1000 of revenue charged ($50), but receive a higher 10% value ($100) on each $1000 not charged.[19]

If the hospital system is rational, faced with a 10% incentive on dollars not spent, it will continue to deliver services and charges whose profit margin is over 10%, while preferring the rebate to charging services whose profit margin is below 10%.[20]

***Example: A charge to avoid?*** For example, assume that a hospital gives $100,000 of a cancer drug, and Medicare pays it 106%, or $106,000. The 6% profit must cover additional costs such as cost of capital for drugs in stock. On the other hand, if actuaries anticipated the hospital, on average, to provide that $100,000 drug, but it does not, its costs will tally $100,000 lower than its benchmark. The hospital receives a bonus of $10,000. The bonus for not using the drug ($10,000) is higher than the profit from using the drug ($6,000).

***Example: A charge to continue making?*** On the other hand, take a $1000 PET scan. Say the gross margin is 50% (most of the cost is capital expense or fixed cost.) If the hospital runs the PET scan, it captures a gross profit margin of $500. It is better to run the scan than to receive the smaller bonus payment ($100) for not running it.

---

[18] http://www.alvarezandmarsal.com/hospital-acquisition-physicians-and-pricing-power#overlay-context= White paper, "Hospital Acquisition of Physicians and Pricing Power." (January, 2013). Similarly, Cassak D (2012) Physicians as Employees: will this spell the end of innovation? In Vivo (July).

[19] See, e.g., Reuters (3/2/2009): US Hospital profits fall to zero. "Plunging revenues from investments have forced median profit margins in US hospitals to zero, according to a Thomson Reuters report." In 9/2013, Bloomberg showed recent quarterly profits for the HCA hospital chain varying in the 3.5% to 5% range. http://www.bloomberg.com/quote/HCA:US

[20] The actual game-theory math would be much more complicated than presented here, involving multivariate calculations across fixed, variable, and semi-variable costs, among numerous other considerations. The math I am providing is a "cartoon" example only.

***What about a value-priced diagnostic test*** *?*– say, a $3000 genomic test. If the hospital orders the test, it will be run and billed to the payer by an outside lab. Therefore, no cash trades hands at the hospital. There is zero profit margin for the hospital. On the other hand, if the hospital foregoes the genomic test, $3000 is saved by the payer, and at the end of the year the hospital and doctor get a bonus payment of $300. Thus, *in a first-order and very simple analysis*, the hospital/physician dyad is incented by $300 to NOT order the value priced $3000 genomic test -- if possible.[21]

ACO's do offer opportunity for labs. For example, a laboratory with a new value-priced molecular test is seeking payer acceptance from (say) Illinois Blue Cross. Traditionally, the lab would hope to have a few credible thought leaders[22] endorse the value of the test. How much more powerful if several ACO's in the plan – say, three different six-hospital networks – in Illinois tell their BCBS plan that they need the test to manage care optimally and efficiently, particularly in the setting of carefully monitored costs which the hospitals are responsible for.[23] The voice of a well-respected ACO network will likely have more credibility than the voice of any single physician, no matter how well credentialed.

**THE EVIDENCE ASSESSMENT SYSTEM AND HIGH VALUE NOVEL DIAGNOSTICS**

5. The Dossier Dilemma – A Challenging Document
6. The Need for Evidence is Naturally High – It's No One's Fault
7. Technology Assessment Can be Biased against Diagnostic Tests
8. Value Pricing and Its Challenges

## 5. THE DOSSIER DILEMMA - A CHALLENGING DOCUMENT

[21] Other factors may apply. For example, the hospital may order the $3000 test, and discovery the patient does not need a $30,000 drug, and receive a bonus for the $30,000 drug not dispensed. I would call this the "second order" rather than "first order" calculation.

[22] The term of art is KOL – key opinion leader. http://www.glossary.pharma-mkting.com/keyopinionleader.htm

[23] ACO's are incented to save on costs, but generally cannot spend uncompensated dollars to do so. In contrast to capitated models (HMOs), an ACO is reimbursed fee-for-service during the year while bonuses or penalties are tallied at year's end. Say the base case is $1000 spent and 0 rebate, and the hospital has a 5% or $50 margin. If the hospital saves the plan $200, assume the hospital still has a 5% margin on $800 spent ($40) plus a 10% rebate on $200 saved - $20, for a net hospital margin of $60. Now assume the hospital spends $100 more to save the plan $200 (net savings being $100), and the $100 is chargeable when it occurs. The hospital has spent net $900, with a $45 margin, plus the hospital gets a bonus of $10 for the health plans $100 net savings, or $55. Now in the final example, the hospital itself spends $100, which is uncompensated, to save the plan $200. The plan's payment to the hospital is $800, with a $40 margin. The hospital also gets another $20 for the $200 the insurance plan saved, adding to $60. But now the hospital must subtract the $100 in *uncompensated* cost-saving services, and the hospital has a final net margin of $60 - $100 or -$40. Thus, providing uncompensated services to save the health plan money is not feasible for the hospital in the fee for service ACO model.

Administrative biomedical reviews often involve dossiers. For an FDA advisory board, the FDA prepares a dossier for the panelists to read, and the company prepares its own. This is an example of "dueling dossiers," countless pairs of which can be found on the FDA website as the public evidence of its many drug and device advisory boards.

For companies trying to communicate with insurers, it is inefficient and likely ineffective to just give an insurer a packet of publications on the new genomic test. When will the reviewer read the papers? Will he read them at all? How much content knowledge does he have? The payer reviewer will generally have less content knowledge than an FDA reviewer or FDA panelist, who are chosen for their particular fields of expertise. So a dossier that summarizes and precedes the bundle of peer reviewed publications is important.

**But how long should it be?**

> **Too long**, and it will be unread (the payer medical director has a much lower burden of attention and less fear of public or supervisory review than an FDA reviewer).
>
> **Too short**, and whatever has been edited out for brevity will be deemed missing, or inadequately explained, and cause a non-coverage decision.

Since the dossier is written by the company, it is viewed with a great deal of skepticism. The payer thinks: a few favorable papers are cited, then a few tables are cherry-picked, but how many unfavorable papers or editorials have been left out? (And the payer thinks: How do I know? Where would I find them?) Within the company-written dossier, anything that seems to be wrong, or hyped, will be treated ruthlessly by the reader.

I have concluded…Dossiers: Can't live with them, can't live without them.

## 6. SOME NEW TESTS ARE ACTUALLY ONLY WEAKLY VALIDATED OR OFFER LITTLE ADVANTAGE

Discussions with payers and presentations made by payer decision-makers at conferences suggests that policymakers and payers learn to be skeptical because some novel diagnostic tests that are brought to their attention have relatively little advantage over existing practices.[24] This observation is not particular to diagnostic tests; the same problem occurs in the slow evolution of drugs, medical devices, and surgical procedures. Blockbuster advances in any area of medical technology are rare, and most advances are incremental.

---

[24] Hayes DF et al. (2013) Breaking a vicious cycle [tumor-based diagnostics]. Science Transl Med5:196cm6. "…Few tumor biomarker tests have been adopted as standard clinical practice….insufficient investment in research and development [and insufficient] scrutiny of biomarker publications by journals…a multitude of substandard studies…very few studies that address clinical utility at a high level of evidence." The paper as a whole calls for both reimbursement and regulatory reform.

The strongest category of evidence, for the most straightforward problem, is that obtained for combination diagnostic tests which are studied in FDA clinical registrational trials for new cancer drugs. These enter the market, necessarily, with a lot of data published and with a lot of technical validation. Even the usually skeptical reviewers at the BCBS TEC organization seem to be satisfied with such data (for example, see their review regarding the accuracy and value of the BRAF test in choosing the drug Zelboraf for melanoma.)[25]

One important new category of tests is multi-analyte tests in which the lab report's results (typically a single variable) are based on an algorithm.[26] Validation for IVDMIA-type tests is also appropriately held to very high standards. If the test, for example, is to be predictive of 10-year recurrence rate of "Cancer X," there is virtually no way to second-guess the validity of the test (unless a new decade of data collection were undertaken.) Say a particular clinic gives the test to 60 women, 20 of whom score a 10% ten-year recurrence risk. The clinic director notes that within 3 years, 4 of the 20 women in this 10% risk pool recur, rather than the predicted 2 women of 20. By itself, this predictive performance seems under par, but is really impossible to interpret. Perhaps it is a quirk, or perhaps, it is a warning that the test was miscalibrated by a factor of two and should not be used. How would you know? (You would need a lot of women, and follow them for 10 years). Because of this, it is important that the foundational work for the ten-year prognostic test be adequate in scale, and be closely reviewed for validity before it is widely used.

Numerous problems could cause a variation in the predicted versus the observed accuracy of an IVDMIA/MAAA test. The original study, while correct, may have been by chance better correlated with outcome than the test would generally perform. One must also ask: Was the population in the original trials representative enough of the individual doctor's next patient? Do the original clinical pathways (from 5, 10, or sometimes 15 years ago in the case of a 10 year study) remain applicable today, in the face of constantly evolving imaging, surgical, radiotherapeutic, and pharmacologic interventions. The test may be correct and worthwhile, but the questions need to be asked to be assured that it is. The Institute of Medicine recently issued a lengthy report detailing problems when an improperly validated IVDMIA-type test was put into use in a clinical trial at a major academic medical center.[27]

---

[25] November 2011: Special Report, Companion Diagnostics: Example of BRAF gene mutation testing to select patients with melanoma for treatment with BRAF kinase inhibitors.
http://www.bcbs.com/blueresources/tec/vols/26/special-report-companion.html
See also interview with the BCBS report's lead author, in: Reinke T (2012) Targeted Medications: New focus on companion tests. Managed Care February 2012,
http://www.managedcaremag.com/archives/1202/1202.companiondiagnostics.html

[26] The FDA termed these "IVDMIAs" in vitro devices-multianalyte index assays." The AMA CPT code system has a new category for these tests in 2013, calling them MAAAs: Multi analyte assays with algorithms. While their may be several reasons for the alternate terminologies, AMA and laboratory stakeholders would have wanted to avoid any term beginning with "IVD," which has generally been reserved for only FDA-approved kit tests in comparison to LDTs – lab developed tests. The term "home brew," which was in fairly wide use a decade ago and described lab-developed versus kit tests, is now politically incorrect.

[27] Institute of Medicine (2012) Evolution of translational omics: Lessons learned and the path forward. 274pp. http://www.iom.edu/Reports/2012/Evolution-of-Translational-Omics.aspx Problems with the Duke test were also covered in the trade press and national media. Singer N (2011) Duke suspends researcher and halts cancer

One other difficulty with prognostic IVDMIA type tests is the fact that *when a series of patients are tested with two tests that use the same general model and make the same claim*, the two tests can give *highly discordant results* to one another, yet each one may be *equally valid and accurate*. (See Appendix). This topic was recently discussed in the trade press, but as the Appendix shows, it is *a mechanical consequence of simple probability rules* (on top of any actual differences between the two tests). Therefore, this particular problem will not go away and may be expected to gather more attention in the next few years as more comparative studies between two IVDMIA tests are performed.[28]

**7. Technology Assessment Can be Biased against Diagnostic Tests**

This is the most technical portion of the present essay, and the reader may want to hear the "take home lesson" and skip to section 8. This is the take home lesson:

> While technology assessments are important, and may reveal important flaws in a test that become clear once they are pointed out, the current technology assessment process has many degrees of freedom where it can go awry and produce unnecessarily negative results on diagnostic tests.[29]

While Section 6 was directed to real difficulties in the tests themselves, the technology assessment culture can create problems unfairly even where none exist. To name one example, technology assessments apply a hierarchy of evidence in which double-blinded randomized trials are at the highest rank of clinical validity. But there is no such thing as a double-blinded diagnostic test.

> Imagine a clinical trial where patients in one arm get a PET scan, and in the other arm not. But the treated physician does not know the patient's status, since this is a double-blinded trial. For the treating physician to be "*double blinded*" in his decisions for the patient, he has to be given a real PET report for patients in one arm of the trial, and a fake PET report for patients in the other arm, and actually use each for major patient management decisions, and yet not know which is which.

---

studies. New York Times (7/20/2010). Couzin-Frankel J (2011) More trouble for Duke as FDA audits center. Science (1/28/2011.) Kolata G (2011) How bright promise in cancer testing fell apart. New York Times (7/7/2011). The trade journal *Cancer Letter* covered the story through numerous articles spanning 2 years. http://www.cancerletter.com Documentation from an FDA inspection, including a 27-page site visit report, are available at:
http://www.fda.gov/MedicalDevices/ProductsandMedicalProcedures/InVitroDiagnostics/ucm289100.htm

[28] Ray T (2012) Genomic Health Study Highlights Discrepancy between Oncotype DX and MammaPrint Results. Genomeweb.com, 12/19/2012.

[29] See also e.g. Becla L et al. (2012) Health technology assessment in the era of personalized healthcare. Int J Tech Assmt Health Care 27:118-26. "…Suitable models should be developed. Conclusions: Integrative, systems biology-based approaches toward personalized medicine call for novel assessment methods. The translation of their highly innovative technologies into the practice of health care requires the development of new HTA concepts."

This example is obviously nonsensical and a little mind-boggling. But the example shows what is required for a "double blinded" physician in a two-arm trial randomized on using or not using a diagnostic test and with the ideal state of double-blinding. The example shows in a nutshell that what may be easy and commonplace for a pharmaceutical trial (double blinding) can turn nonsensical for a diagnostic test.

Since diagnostic tests are *not* tested in double-blinded diagnostic test trial, the highest level of evidence is unattainable for them. The next-best level of evidence is usually described as pivoting the two-arm trial on the diagnostic test. But this is often nonsensical, for other reasons. Imagine a two-arm trial pivoted on the Herceptest. In Arm A, 100 women get a standard of care therapy and live one year. In Arm B, 100 women get Herceptest, and the 20 who are Her2eu positive get Herceptin, of which half respond and live one full year longer. The benefits in these ten patients (who live one year longer) are washed out by the other 90 patients are also in the test-using Arm B but who have exactly the survival they would have had in Arm A.[30] If the trial were run, the "B" arm would have a 13.2 month average survival, or a 1.2 month advantage, *very unlikely* to be statistically significant, and even if it were, *clinically unimpressive at best*. This is the actual trial result: the outcomes comparing the two arms, pivoted on the test, would be the exactly same even if the trial were run on 1000 patients. On the other hand, to use the Herceptin is entirely sound: give each patient in 100 a Herceptest, then give the 20 cases with positive tests the drug Herceptin, and know that half the 20 getting Herceptin live an extra year and on average all 20 live six months longer.

Another bias toward randomized trials involves the concern that retrospective trials can only establish correlation and not causality. Correlation is "bad." To have an example fresh in the reader's mind, wearing shorts is found to be highly correlated with eating ice cream. This correlation seems reliable in retrospective studies in Philadelphia, Denver, and Chicago. But it is not causal: If you take someone in Toronto in January, and put him in shorts, he will not start eating ice cream. Thus, we conclude, the association was a correlation but not causal.

But with diagnostic tests, often what one wants, and what one uses, is correlation (!). Troponin is very useful in the E.R. because it is correlated with heart attack. Troponin does not cause heart attack (if you inject someone with a little troponin, it doesn't cause a heart attack). But the correlation between troponin and heart attack is extremely useful in clinical practice in an emergency room. One of the reasons for doing RCT's in the first place – because you must seek causality rather than correlation – may be absent for the diagnostic test.[31]

---

[30] For example, in Arm A, survival is 12 months. In Arm B, 20 patients get Herceptin and 10 live one year longer. The average survival in Arm B as a whole is 13.2 months, a 1.2 month advantage, and unlikely to be detectable, if detected, unimpressive.

[31] For example, if I take a mercury thermometer as a gold standard, I validate an electronic thermometer by finding in 100 patients, the readings are accurate within 1%. This is an *observational* and *correlational* study. Transferred to an RCT, I would use the mercury thermometer in 100 patients, and the electronic one in 100 patients, and look for survival or other outcome differences. The RCT model is nonsensical relative to the question being asked, but observational and correlational studies are considered a low level of evidence in technology assessments, resulting in data that should not generally be trusted for clinical decision-making.

Another bias that appears in some technology assessments of diagnostic tests is the requirement for "survival" effects. But this is not always germane. For example, there is a real world trial (and I simplify here) where in one arm, 100 patients get thoracic surgery for a presumed operable tumor, and 30 are discovered to have inoperable tumor while the chest is open.[32] These 30 patients underwent "futile surgery." In the PET-CT imaging arm, 100 patients are scanned, and in 20 of them, an inoperable tumor's extent is determined without surgery, and they are spared unnecessary ("futile" or open-and-shut) surgery. However, given there are 100 patients in each arm, 80 of which are treated identically, and 20 of which (in the PET arm) have inoperable tumor, there is unlikely to be a survival difference in the two arms. The benefit must be sought elsewhere (20 surgeries avoided) and not in the metric "survival." There is a benefit, and it is tangible, but it is not survival.

The authors of technology assessments may also stumble unawares against problems that are far better understood by regulatory experts. In a drug trial, there is typical a fixed benefit which is a predetermined endpoint – e.g., cholesterol drops by 20%, or blood pressure drops by 20%, etc. This is the "benefit." The trial will collect innumerable other data points for adverse events – someone has nausea, someone has a headache, someone a rash, and so on, with a noisy distribution of all of these numerous possible adverse events between the two arms. Regulatory experts understand the repeated-measures problem when considering the adverse events. I have seen technology assessments that acknowledge a certain effect was found – say, a standard test was 85% accurate and a new test 90% accurate – but the reviewer will then go on to belabor any differences in randomization. Of course, if you collect 15 or 20 variables – race, education, age, gender, tumor size, and so on – eventually one recorded cohort variable or another will differ at the "p .05" level or to some unspecified degree that alarms the tech assessment reviewer.[33] He may write this up as a serious and concerning flaw rather than a quirk of the multiple measures phenomenon.

Another dilemma of evidence is the "statistical valley of death."[34] A test is to be used in a certain clinical circumstance where a randomized control trial would be the optimal evidence. But there is already so much retrospective, high quality evidence, that a randomized trial would be borderline unethical, lack equipoise, be difficult for IRB approval, or, if approved, simply not enroll effectively because of apprehension of the control arm for patients and doctors. Thus, the level of evidence is "frozen" at the level of retrospective evidence since it is too ethically difficult to launch a two-arm RCT. While the "equipoise" probably is well known in the clinical trial policy literature – usually in the context of trials that were thought to violate equipoise, but gave surprising results[35] – the problem may arise

---

[32] The trial data shown in my example is a thought experiment, but is similar to a real trial: Fischer B et al. (2009) Preoperative staging of lung cancer with combined PET-CT. NEJM 31:32-39. (Correction, 2011, 364:980-1).

[33] See: CTAF (2010) Gene expression profiling for the diagnosis of heart transplant rejection. http://www.ctaf.org/assessments/gene-expression-profiling-diagnosis-heart-transplant-rejection

[34] The term "valley of death" is used in pharma trials when there is enough money for a Phase II trial but not for a Phase III trial, and in similar scenarios.

[35] A 2012 example found that intra aortic balloon pumps, considered an accepted technology, were actually found to be ineffective (equal to placebo) in an RCT. NEJM (2012), Thiele et al., 367:1287-96. This confirms that equipoise should not be based on observational trials. See similarly: Rettig RA et al. (2007) False Hope: Bone marrow transplantation for breast cancer. Oxford University Press. But these examples notwithstanding, there are equipoise-rooted dilemmas: for recent articles, see Joffe S & Miller FG (2012) Equipoise: Asking the right

even more often now and in the future for genomic diagnostics, because of the ease with which retrospective data from paraffin blocks can be gathered. I call this the "pre hoc" equipoise problem. There is no such burden of retrospective data with a new molecule in a drug trial.

Related to the "statistical value of death" problem is the Evidence Based Medicine reviewer who finds two trials each with a 30% survival benefit between the two arms, favoring use of the test or drug, but then the reviewer bemoans the fact there are not five or six trials for him to review. The problem is that after a few favorable trials in the same direction, further randomized trials become unethical, or at the least, impossible to enroll effectively. (This is the post hoc equipoise problem.) With human clinical trials, there is a tangible "shell" over the ethically possible volume of the clinical trial data that technology assessment evidence reviewers rarely mention. On the other hand, regulatory reviewers are highly aware of this difficulty, for example, if faced with a trial design which would involve randomizing half of the subjects away from an FDA-validated and on-label arm and into an experimental arm.

## 8. VALUE PRICING AND ITS CHALLENGES

The Genomic Health Oncotype DX model dates to 2002/2003 (since the pivotal publication was in late 2004). As has been well-publicized, and noted in the introduction, the road to widespread coverage by US payers was very slow, from the 2004 market introduction through 2008 or 2009. And it was very, very expensive in terms of marketing and administration, let alone costs for R&D.

Working in a field that is dominated by commodity pricing, the value pricing model is seen as a panacea by lab industry executives (see, e.g., a recent trade journal report which is typical[36] ). (This is variably called a "value pricing" model, a "pharmaceutical pricing" model, or a "market based pricing" model.) However, this model requires not only "value pricing" argumentation on paper but a market monopoly in practice. The value pricing argument fails because the test is not a monopoly. To illustrate the concept easily with an example taken from the drug world, let's say that I can produce a value economic argument that Prozac is worth $6000 a year in QALY's, or $500 per month. *But the day it becomes generic*, fluoxetine is worth $5 per month, which is 99% less that the value I just demonstrated in my economic model. Value based pricing requires both clear and tight economic scenarios, and a relative monopoly in its marketplace.[37]

---

questions for clinical trial design. Nat Rev Clin Oncol 9:230-5. Miller FG & Joffe S (2011) Equipoise and the dilemma of clinical trials. NEJM 364:476-80. Gifford F (2000) Freedman's 'clinical equipoise' and the slide scale, all dimensions considered 'equipoise.' J Med Philos 25:399-426.

[36] "The key we've found is adopting a pharma-like value perspective in support of value-driven payments for diagnostics." Gray Sheet, Personalized Medicine and Palmetto (12/3/2012, report of a conference).

[37] For an excellent and thorough discussion, in a somewhat different direction than I have taken, see Hornberger J (2013) Assigning value to medical algorithms: implications for personalized medicine. Personalized Medicine 10:577-588. See particularly, "Principles for valuation in medical products," p. 582ff.

The Oncotype DX breast cancer test arguably addressed a singular decision point in the therapy pathway (whether or not to give chemotherapy after a breast cancer lumpectomy).[38] It is rare that new tests address a clinical area where there is a dearth of information and in which the new test being developed promises to have a sole source position. More typically, the clinical scenario comprises multiple and uncertain patient presentations, heterogeneous patients, and multiple therapeutic choices, and the new information provided cannot be regarded as a "monopoly" type of information that can be framed in only "one" scenario for care (Either X without our test or else Y with our test.) My experience in watching the reactions of payer policymakers to value pricing economic models suggests that the value pricing scenario is often viewed by the listener as too artificial and reduced and stylized into a narrow argument that the test can win. That is, there may not be anything wrong with the math inside the pharmacoeconomic model, but the payer will disagree with the assumptions and premises as being too narrow. I do not want to discuss any particular example, but the general problem would be "Our test is cost saving by $5000 in Scenario A." The payer thinks, well, yes, but he thinks Scenario B, C, D, and E are all more likely.

Or, As is well known in evidence-based medicine policy circles, value based arguments are highly subject to differing interpretations. For example, some publications establish the Genomic Health test as truly cost-saving, while on the other hand a United Kingdom NICE model released in 2012 gives the same test a fairly high cost per QALY, concluding it is not cost saving.[39] [40] (This particular disagreement evolved over time with several stages, but for this white paper, the point is, that different bodies can disagree about the same evidence.)

Policymakers are aware of the level and frequency of public disagreement in the debate about health economic outcomes models. Peer-reviewed models ought to be credible; the work of NICE ought to be credible, too. Over time, the recurring patterns of disagreements among experts have cast into question the modeling on which value based diagnostics pricing relies.[41] [42] [43]

---

[38] There are now multiple tests in this space; for several years the Oncotype DX test had a near-monopoly position, e.g., over 90% market share.

[39] http://www.nice.org.uk/nicemedia/live/13283/58040/58040.pdf Genomic Health released a dissenting press release (3/3/12). See also NICE Q&A response, http://www.nice.org.uk/nicemedia/live/13283/57996/57996.pdf

[40] See also: Silverman E (2013) NICE vs Not So NICE: A nasty squabble over reimbursement. http://www.pharmalive.com/nice-vs-not-so-nice-a-squabble-over-reimbursement-in-the-uk

[41] See also the Institute of Medicine workshop, "Assessing the Economics of Genomic Medicine," July 17-18, 2012. A conference report is forthcoming in 2013. http://www.iom.edu/Activities/Research/GenomicBasedResearch/2012-JUL-17.aspx

[42] For an example of creative thinking in this area, see: Califf RM et al. (2008) Considerations of net present value in policy making regarding diagnostic and therapeutic technologies. Am Heart J 156:879-85. Also: Trikalinos TA et al. (2009) Decision-analytic modeling to evaluate benefits and harms of medical tests. Med Decis Making 29:E22.

[43] Elkin EB et al. (2011) Economic evaluation of targeted cancer interventions: critical review and recommendations. Genet Med 13:853-60.

Finally, value based pricing can be most readily accepted by insurers (assuming the data and modeling is correct) when a test is literally cost-saving. This is rare.[44] Usually, the test simply is proposed to have an "acceptable" cost effectiveness such as $40,000 per QALY, which means it is actually cost-increasing and that is a reason to postpone coverage.[45]

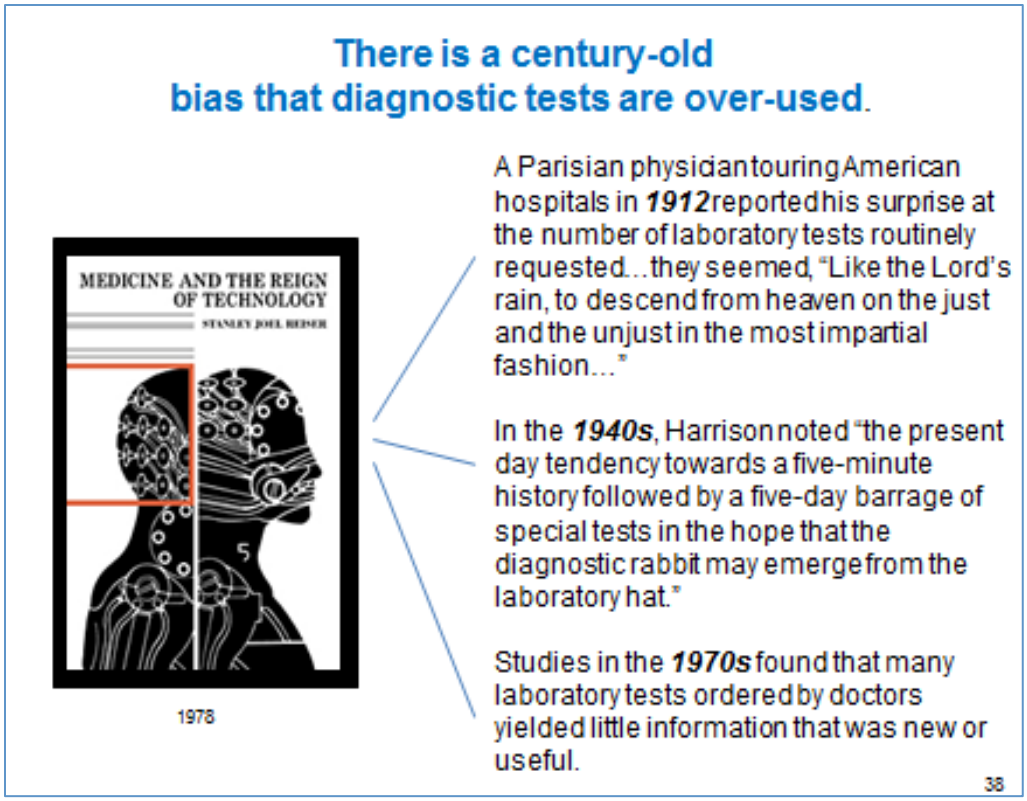
New personalized medicine tests easily generate red ink for the developer based on development time, market price, and return on investment considerations. [46] Without value pricing, which repays risk and investment, only commodity products (products provided at the cost of production) are available, and innovation stagnates.[47]

**EPILOG:**

**9. IT'S NEVER BEEN EASY, BUT IT'S WORTH IT: POLICYMAKERS AND DIAGNOSTIC TESTS**

Although it goes without saying that diagnostic tests are often crucial for making therapeutic decisions, one can also trace a longstanding bias against diagnostic tests, going back a full century. For example, I have used this slide:

[44] Fang C et al. (2011) Cost utility analysis of diagnostic laboratory tests: Systematic review. Value in Health. Epub.
[45] For example, a pill that is taken once and provides a one year life increase for $10,000 is highly cost effective per QALY. However, the immediate one time cost of such a pill for the US would be about 3 trillion dollars. (300M x 10K).
[46] http://www.nature.com/nrd/journal/v8/n4/full/nrd2825.html The microeconomics of personalized medicine. Davis JC et al Nat Rev Drug Disc 8:279-86.
[47] See McKenzie RB, Lee DR (2008) In Defense of Monopoly. How market power fosters creative production. Univ. Michigan. For a free online summary, see http://object.cato.org/sites/cato.org/files/serials/files/regulation/2009/11/v32n4-3.pdf




Similarly, just a couple years ago the then-chief medical officer of CMS spoke at an NIH conference, and described the coming wave of genetic tests in cancer as "a tsunami of costs." Yet all diagnostic tests, including imaging, in cancer care may be only 5% of cancer costs,[48] so it is unlikely that the necessary genomic tests for cancer will really constitute a "tsunami of costs."

In my own experience, I attended several AMA CPT editorial meetings and related public meetings in the last several years. Some of the panelists who were payer medical directors were *visibly* livid at the large number of genetic tests that were "*bombarding*" their claims processing systems.[49] Consultants who interview "focus groups" of payer medical directors with model dossiers for new diagnostic test concepts are often barraged by criticisms of the potential dossiers, as seen through the eyes of the medical directors.

When a test does have a favorable "public rating" it is usually because the public image is simplistic and not entirely accurate, e.g. the Herceptest, and then, it is vulnerable to being "taken down" (e.g., papers a few years ago that Herceptest had many flaws, was not reliable, and moreover, RUO her2neu tests

[48] Sullivan R et al. (2011) Lancet Oncol 12:933-80. Delivering affordable cancer care in high-income countries. See also: Based on data I saw at an IMS conference, a rough rule of thumb is that half of cancer care costs are hospital care and 30% drugs. Other costs include physicians, radiotherapy, diagnostics, and hospice.
[49] I am not arguing with their perceptions: they felt livid and they felt bombarded.

were commonly used in the laboratory market and were even worse).[50] This observation placed in this section because when this "expose'" occurs it is well-accepted, I argue, because it comfortably unseats the Her2neu test from its pedestal and delivers it to the "see, it couldn't be that good" status that new diagnostic tests start at.

Relative to the poor "PR" of diagnostic tests, there have been biases that the quality of clinical lab operations, themselves, are not very reliable. An early manifestation of this concern was the one-time groundswell of publicity against the lab industry in the 1960s and 1970s that led to the promulgation and sequential strengthening of CLIA legislation at the federal level.[51] The pathway to the CLIA legislation was a real "war against labs" conducted by the commissioner of health in New York, among others, with study commissions in full swing, with New York Times editorials and so forth. More recently, in policy circles at least, circa 2008, 2009, there was news of the epidemic failure of basic breast cancer ER/PR testing in Newfoundland[52] and later of her2neu testing in Quebec.[53] More recently, the New York Times covered the failure of a complex oncology LDT at Duke University, an event which also triggered investigations by the IOM and FDA. (Fn. 25, supra). In mid-2013, a New York Times Op-Ed called for the FDA to more strictly regulate diagnostic tests.[54]

To close this section, a recent example of bias against novel diagnostics by policymakers is the August, 2012, announcement by CMS staff that they would pay *nothing* for the value of an IVDMIA or multi-analyte test over the value of its components (generally commodity measures, such as RNA levels of 15 genes).[55] This neatly sidestepped arguments over the valuation of the algorithm … by assigning a value of nothing.[56] (This particular proposed policy is currently on hold at CMS and under review in 2013).

When one stands in downtown Washington DC today – say, a few blocks from the White House or Capitol – it is easy to imagine what the same streets looked like in the 1780s and 1790s, when Washington and Jefferson rode by on horseback or in carriages. Today, telecommunications signals fill the air, the buildings are full of internet and computer wiring, and jets fly overhead at 40,000 feet. The technology revolution of the past 200 years was all accomplished in a world of patents. Yet today, the value of patents is still regularly questioned. Similarly, if one compares the practice of medicine in 1905 by William Osler in Baltimore, and the practice of his successors today on the same grounds of

---

[50] For discussion see Phillips KA (2008) Closing the evidence gap in the use of emerging testing technologies in clinical practice. JAMA 300:2542-4. See also references in: Grimm EE et al. (2010) Achieving 95% cross-methodological concordance in Her2 testing: causes and implications of discordant cases. Am J Clin Pathol 134:284.

[51] Numerous articles archived by the author.

[52] http://theoncologist.alphamedpress.org/content/13/11/1134.full

[53] http://www.darkdaily.com/public-learns-about-errors-breast-cancer-testing-canadian-province-quebec#axzz2GmcHzFUK

[54] New York Times (7/7/2013). The gap in medical testing [FDA, LDT]. http://www.nytimes.com/2013/07/08/opinion/the-gap-in-medical-testing.html

[55] CMS wrote as a proposal in August 2012: "Medicare does not recognize a calculated or algorithmically-derived rate or results as a clinical laboratory test." In: ACLA comment to CMS on CLFS policy, 9/28/2012. At: http://acla.com/sites/default/files/ACLA%20Comments%20on%20CMS%20Proposed%20Payment%20Determinations%20for%20CLFS%202013.pdf

[56] This CMS policymaking is discussed by Hornberger (2013), FN 37.

Johns Hopkins, the value and impact of modern diagnostic tests – from PET imaging to genomics - is awesome. Yet every advance has been hard won. **Technological change is only accomplish when adventure and investment are repaid; that is, when market prices are above marginal cost.** Providing the appropriate and hopefully socially optimal repayment for the breakthroughs in molecular testing are a challenge that deserves to be faced head on.

# APPENDIX

**Two IVDMIA type tests can give highly discordant results to one another, yet be equally accurate**

While this is easily demonstrated on a whiteboard, that whenever two IVDMIA tests are compared "head to head" the results are likely to be quite discordant, the results will likely upset an evidence-based medicine audience.

Take a population of 100 women, who have had a clean lumpectomy for early stage breast cancer. 15 of 100 will have a recurrence within one year (a fairly realistic cohort). The 15 of 100 who will recur are shown in in the top three rows: (e.g. 3 rows x 5 columns highlights the group of 15 patients who recur):

Archival Tissue
100 Patients
15 / 100 (Top 3 rows)
Will Actually Recur

RECUR

| | | | | |
|---|---|---|---|---|
| **1** | **2** | **3** | **4** | **5** |
| **6** | **7** | **8** | **9** | **10** |
| **11** | **12** | **13** | **14** | **15** |
| 16 | 17 | 18 | 19 | 20 |
| 21 | 22 | 23 | 24 | 25 |
| 26 | 27 | 28 | 29 | 30 |
| 31 | 32 | 33 | 34 | 35 |
| 36 | 37 | 38 | 39 | 40 |
| 41 | 42 | 43 | 44 | 45 |
| 46 | 47 | 48 | 49 | 50 |
| 51 | 52 | 53 | 54 | 55 |
| 56 | 57 | 58 | 59 | 60 |
| 61 | 62 | 63 | 64 | 65 |
| 66 | 67 | 68 | 69 | 70 |
| 71 | 72 | 73 | 74 | 75 |
| 76 | 77 | 78 | 79 | 80 |
| 81 | 82 | 83 | 84 | 85 |
| 86 | 87 | 88 | 89 | 90 |
| 91 | 92 | 93 | 94 | 95 |
| 96 | 97 | 98 | 99 | 100 |

**MAAA TEST "A"**

High Risk = 15 Cases

66% will recur

RECUR

| | | | | |
|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 |
| 6 | 7 | 8 | 9 | 10 |
| 11 | 12 | 13 | 14 | 15 |
| 16 | 17 | 18 | 19 | 20 |
| 21 | 22 | 23 | 24 | 25 |
| 26 | 27 | 28 | 29 | 30 |
| 31 | 32 | 33 | 34 | 35 |
| 36 | 37 | 38 | 39 | 40 |
| 41 | 42 | 43 | 44 | 45 |
| 46 | 47 | 48 | 49 | 50 |
| 51 | 52 | 53 | 54 | 55 |
| 56 | 57 | 58 | 59 | 60 |
| 61 | 62 | 63 | 64 | 65 |
| 66 | 67 | 68 | 69 | 70 |
| 71 | 72 | 73 | 74 | 75 |
| 76 | 77 | 78 | 79 | 80 |
| 81 | 82 | 83 | 84 | 85 |
| 86 | 87 | 88 | 89 | 90 |
| 91 | 92 | 93 | 94 | 95 |
| 96 | 97 | 98 | 99 | 100 |

**MAAA TEST "B"**

High Risk = 15 Cases

66% will recur

RECUR

| | | | | |
|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 |
| 6 | 7 | 8 | 9 | 10 |
| 11 | 12 | 13 | 14 | 15 |
| 16 | 17 | 18 | 19 | 20 |
| 21 | 22 | 23 | 24 | 25 |
| 26 | 27 | 28 | 29 | 30 |
| 31 | 32 | 33 | 34 | 35 |
| 36 | 37 | 38 | 39 | 40 |
| 41 | 42 | 43 | 44 | 45 |
| 46 | 47 | 48 | 49 | 50 |
| 51 | 52 | 53 | 54 | 55 |
| 56 | 57 | 58 | 59 | 60 |
| 61 | 62 | 63 | 64 | 65 |
| 66 | 67 | 68 | 69 | 70 |
| 71 | 72 | 73 | 74 | 75 |
| 76 | 77 | 78 | 79 | 80 |
| 81 | 82 | 83 | 84 | 85 |
| 86 | 87 | 88 | 89 | 90 |
| 91 | 92 | 93 | 94 | 95 |
| 96 | 97 | 98 | 99 | 100 |

The first MAAA test, **TEST A,** identifies 15 women at high risk: Patients 1-10 (who will recur) and patients 16-20) who will not. The 15 high risk patients have a 66% recurrence risk, making this a very good test. (The remaining 85 patients are low risk, with a 5/85 or 6% recurrence).

The second MAAA test, **TEST B**, also identifies 15 women at high risk, who have a 66% risk of occurrence. However, only 5 of the 15 high risk women in the two test results are the same. (Only women 1-5 are in the high risk group of 15 women in both tests.) 10 of the 15 high risk women in the two test results are different. This occurs even though both tests have exactly the same performance (15 women with 66% risk and 85 women with 6% risk).

Many non expert reviewers will be very puzzled that the tests will have the same performance, yet identify 10 of 15 high risk women differently. Note that the result would be the same if the tests were predictive rather than prognostics, e.g. if the 15 women in the top three rows were responders to chemotherapy out of 100 women given the chemotherapy.

The tests <u>can</u> be compared usefully for overall performance, but <u>cannot</u> be compared for "discrepancy" by giving each of 100 women the two tests and comparing individual results one to another. (The example here shows the maximum discrepancy or minimum concordance possible).