APPENDIX: REVIEW AND INTERPRETATION OF THE EVIDENCE FOR "The Hard Truth about Artificial Intelligence in Healthcare: Clinical Effectiveness is Everything, not Flashy Tech"

Roy S. Zawadzki, MS

Treatment Paradigm and Clinical Outcomes

HeartFlow's highest quality evidence for the clinical efficacy of FFRCT comes from the Prospective Randomized Trial of the Optimal Evaluation of Cardiac Symptoms and Revascularization or "PRECISE" trial, which pits functional testing versus a "risk-stratified, precision" CCTA and FFRCT screening regimen. This trial is considered to be "pragmatic" in that the focus is on observing how CCTA + FFRCT performs in real-world clinical practice. Although not needed for regulatory purposes – the barriers for many MAMD "approvals" are quite low compared to, say, biologics – one can consider PRECISE to be a Phase 3 study (i.e. definitive), or at least can expect that the results will be used as such from the medical community.

Two important clinical trials precede PRECISE and provide evidence surrounding CCTA, the precursor to FFRCT: Prospective Multicenter Imaging Study for Evaluation of Chest Pain (PROMISE) and Scottish Computed Tomography of the Heart (SCOT-HEART). The earlier trial, PROMISE, found that over a two-year period, screening via CCTA did not demonstrate an improvement in clinical outcomes compared to functional testing. On the other hand, SCOT-HEART followed patients for five years and found CCTA, in addition to standard care to have a substantial benefit over just standard of care in the composite endpoint of non-fatal myocardial infarction (MI) or death due to MI (hazard ratio of 0.59, 95% confidence interval of 0.41 to 0.84). Notably, there was no difference in cardiovascular death or death from any cause between the two treatment arms, so these results are driven by the non-fatal MI rates. Furthermore, the rate of catheterization and revascularization was not significantly different across groups – recall that one of the main purposes of noninvasive screening is to increase the efficiency of catheterizations, which would likely be reflected indirectly with this endpoint.

Overall, I would not necessarily interpret the results from SCOT-HEART as a win for CCTA. Since revascularizations were not different between the treatment groups, the authors state the difference is associated with medical management. However, as the accompanying editorial points out: "the differences in medical management between the groups were modest…and do not seem to be sufficient to explain the much lower rate of myocardial infarction in the CTA group than in the standard care group."[i] So, we are left unable to ascertain what exactly about CCTA will help clinicians screen and treat potential CAD patients. PRECISE aimed to improve upon this by, firstly, implementing a stratification score before noninvasive screening alluded to by the SCOT-HEART authors, and, secondly, allowing for the option for moderate or higher risk patients to undergo FFRCT in addition to CCTA.

The initial one-year results of PRECISE were released in August 2023 in *JAMA Cardiology*. Compared to functional testing, the composite endpoint of clinical efficiency (invasive catheterization without obstructive CAD) and safety (death or non-fatal MI) showed a significant superiority in the precision strategy. It should be noted, however, that this was

driven entirely by the clinical efficiency outcome and there was a slightly higher, but not statistically significant, rate of non-fatal MI (1.2% vs. 0.5%). More time is needed to conclude whether the precision CCTA with optional FFRCT strategy results in superior clinical outcomes for patients. In response to the results, HeartFlow stated that the "CCTA+FFRCT-centered strategy is the superior diagnostic pathway for patients with stable chest pain and suspected coronary artery disease."[ii] I think such a claim is somewhat premature not only because we have yet to see the safety data but additionally because of limitations in the trial design that do not allow us to see the direct benefit of FFRCT.

Beginning with the primary endpoint, the outcome is analyzed as the time to the first event: all-cause death, non-fatal MI, or invasive catheterization without obstructive CAD. The inclusion of catheterization allows the trial to end much earlier (in this case, after a year) and have higher power for a given sample size due to more events. However, patients are only eligible for the catheterization if they were screened positive by the non-invasive testing, which means this endpoint is driven by false positives. Thus, the clinical outcomes are not yet clear for the false negatives and those who never received screening in terms of clinical outcomes. Additionally, a "positive" test does not necessarily mean that the patient went on to get an invasive catheterization because it is at the clinician's discretion, subjecting this endpoint to further selection bias.

On the topic of treatment, the proportion of revascularization was higher in the experimental arm (9.2% vs. 5.2%), which is driven by PCIs. The authors do not address this in the paper, which leaves me to speculate why this may be the fact. A possible explanation is that the results from the CCTA + FFRCT and risk-stratification are causing doctors to undertake such procedures more quickly. I do not believe this is necessarily positive considering the results of the landmark International Study of Comparative Health Effectiveness with Medical and Invasive Approaches (ISCHEMIA) trial. In ISCHEMIA, conservative optimal medical treatment (OMT) did not result in different clinical outcomes than revascularization (the rate of revascularization, i.e. cross-over, in the control group was 20%). In turn, I believe there is still some debate in the medical community regarding how a CCTA strategy can help manage patients, conservatively.[iii]

The main issue I have with PRECISE as it relates to FFRCT is that we are limited in what causal conclusions we can make based on the trial. Essentially, in the experimental arm, we are unable to disentangle the individual effects of the risk-stratification, CCTA, and FFRCT. One could argue that this is a pragmatic trial, so it was not designed for this purpose. Even so, these results may incorrectly be used by some to explain why the FFRCT is worth the additional price tag on top of CCTA. Furthermore, we may run into a scenario like SCOT-HEART where we are unable to convincingly explain how FFRCT is improving the clinical workflow. In the supplemental appendix of the trial, it is shown that only 30% of the experimental arm received CCTA with FFRCT via the physician's discretion. This produces selection bias that is not overcome by the trial design so we cannot simply just compare the results on the FFRCT subgroup to ascertain the independent effect of the product.

A smaller, investigator-initiated trial (i.e. not by HeartFlow) called the FORECAST trial compared CCTA + FFRCT with a relatively unrestrictive "standard-of-care" testing pathway that could include CCTA but not FFRCT. They found that while catheterization utilization was significantly less (19% vs. 25%), there was no difference in clinical outcomes. Mean total cardiac

costs were higher for the FFRCT group but were not statistically significant (£114, or +8%, with a 95% confidence interval of −£112 [−8%] to +£337 [+23%]). FORECAST gives us a clearer view of the marginal effect of FFRCT, which appears to be largely neutral albeit on a slightly different population.

Another issue is that in the control group, the physician can choose the functional testing methodology for the patient. This is problematic because, as we will cover in the next sub-section, not all functional testing methodologies have equal diagnostic performance. I am concerned that while all centers will have cutting-edge CCTA equipment, courtesy of HeartFlow, they may not have employed the cutting-edge functional testing equivalents. Roughly 10% of patients in the control group received stress MRI, 30% stress echocardiography, another 30% MPI/PET, and 10% exercise ECG without imaging. Is this mix of functional testing truly the highest standard of care? Once again, this is not necessarily problematic for a pragmatic trial and the paper's limitations do indeed state that "the pragmatic trial design precludes evaluation of different [functional testing] choices." Regardless, this evidence cannot be used to prove the superiority of FFRCT.

Predictive Performance FFRCT vs. CCTA vs. Functional Testing

For the diagnostic accuracy of ischemia, we will primarily focus on the results surrounding a paper published in the *Journal of the American College of Cardiology* in 2019 on the PACIFIC study that took 208 patients with suspected stable CAD and ran CCTA, FFRCT, PET scan, and SPECT. Then, they compared the results with the ground truth diagnosis via an invasive FFR measurement. The endpoints included sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), accuracy, and AUC. In the paper, the results were presented in multiple ways. Firstly, the performance was computed on a per-patient and per-vein basis: in the per-patient case, if the patient had one or more vessel classified as notably blocked then they were labeled as a positive case. In addition, the results were reported per-protocol (PP) and intention-to-diagnose (ITD). A PP analysis essentially disregards patients who were unable to be screened by all methods with FFRCT taking in almost all of these cases. The ITD analysis, on the other hand, will mark patients unable to be screened as a positive case. In summary, there are four sets of results: per-vessel PP, per-patient PP, per-vessel ITD, and per-patient ITD.

On its website, HeartFlow touts the superior AUC of 0.94 (95% CI: 0.92-0.94), which comes from the per-vessel PP results. This was statistically significantly superior with the next highest AUC coming from the PET scan at 0.87 (95% CI: 0.83-0.90). In the per-patient results, the advantage over PET vanishes: the AUC of FFRCT is 0.92 while the AUC of PET is 0.91. Therefore, it is key to discern which set of results is more important in which scenarios. In my opinion, the highest priority is finding the presence of any significant CAD blockages rather than the number of them, which favors the per-patient results as opposed to the per-vessel results. Firstly, if at least one significant blockage is found via imaging, then the patient would go on to catheterization where any missed blockages would be detected. Secondly, the majority of patients only appear to have 1-vessel CAD so the per-vessel and per-patient metrics would be equivalent for these patients.[iv] Thirdly, consider that the most likely treatment option, likely

medical management, would not necessarily change by the number of blocked vessels, but rather the severity of them and severe blockages are easier to spot, in general.

One may debate my previous arguments on whether the per-vessel or per-patient results are more important. Regardless, the ITD results do not show superior diagnostic performance for FFRCT. ITD presents the most practical and realistic case where a patient comes in wishing to be screened. Ultimately, 25% of patients were not able to be evaluated by FFRCT while only four (1.9%) of patients failed the PET procedure. In the ITD, per-vessel analysis FFRCT was equivalent to both CCTA and PET with the upper bound of the 95% confidence of PET being the higher than that of FFRCT. The AUC for FFRCT was 0.83 (95% CI: 0.79-0.86), the AUC for CCTA was 0.80 (95% CI: 0.77-0.84), and the AUC for PET was 0.86 (95% CI: 0.83-0.89). For the ITD per-patient analyses, the AUC of FFRCT, 0.79 (95% CI: 0.73-0.85) was inferior to that of PET 0.90 (95% CI: 0.85-0.93).

The non-evaluability of images poses a major practical issue for HeartFlow and the 25% rejection rate is arguably a lower bound. 84% of non-evaluable images were due to motion-related higher heart rates with the rejected patients having a statistically significantly higher heart rate than the included patients (61.8 vs. 56.5). Furthermore, additional beta-blocker prescriptions and higher rates of retrospective acquisition were also predictive of non-evaluable images. In the real world, patients are often more severe and would thus be more likely to exhibit the characteristics associated with non-evaluability. When performing CCTA screening, the staff are likely not as skilled at handling difficult cases, and it may not be reasonable to expect high-quality CCTA imaging devices as in the trial. The study was located in Amsterdam, which may not suffer problems that many underfunded American medical centers in lower socioeconomic areas may.

Real-world evidence on the practicality of use is needed. Fortunately, a large (n = 90,553) retrospective study in the UK named FISH&CHIPS was just completed recently but the full results have not been released yet. Thus, we do not know yet whether rejection rates continue to be high or if there are any burdens on the health system. Many concerns over implementation may be allayed by this data.

Radiation exposure caused by CCTA is an additional practical concern of physicians. Hanson et al write: “These acquisitions [by CCTA] generally require higher doses of radiation and should be performed and centers with expertise in performing high quality CCTA. In light of these limitations, we believe ischemia testing will continue to play a role in patients who are not ideal candidates for CCTA.”[v] Further, Duhamel et al. state “a CCTA first strategy in patients with suspected CAD would expose a huge number of patients to ionizing radiation, while alternative functional imaging modalities, such as echocardiography and CMR, are available.”[vi]

Of course, we may not have the choice but to expose a patient to radiation when imaging. Choosing the modality is thus subject to a cost-benefit evaluation. Let’s suppose for sake of argument that PET has equal performance to FFRCT. In general, PET has a lower average radiation dose compared to FFRCT (3.10 vs. 5.31 mSv).[vii] In a case where high-quality images are required for FFRCT to perform optimally, several images of higher radiation may be needed – 5.31 mSv may be a lower bound. In the PRECISE trial, the control group had a significantly lower radiation average exposure than the CCTA + FFRCT group (4.7 vs. 5.2 mSv) with a median of 1.5 vs. 4.8 mSv. One must contextualize these results considering that the control group did not only just include PET. Furthermore, I do not know whether the difference in radiation doses is

practically significant and whether such radiation exposure is worth the tradeoff in diagnostic accuracy. Either way, it is important to consider that "non-invasive" data acquisition is not always straightforward and itself may harm the patient. In a situation where a large percentage of patients are rejected by FFRCT, it may be unethical to subject them to radiation.

Raw diagnostic performance is not necessarily the only defining factor in whether we should choose a screening modality – the cost of screening matters too: an insurer may want to reimburse the cheaper but slightly less accurate diagnostic over the more accurate but expensive one for many reasons including access and cost-effectiveness. We will briefly touch upon this with a few basic calculations translating predictive performance to practical numbers using the methodology of Liu et al.[viii] We can first compute the positive predictive value (PPV) or the likelihood that a patient predicted to have an ischemia actually has one. This will tell us the number needed to screen (NNS) in order to find one patient. We can, additionally, include this with the cost per test in order to quantify the "cost-effectiveness" of the device by multiplying the NNS by the cost of the test. A diagnostic that has a high NNS would mean that mean that, on a population level, we must order more tests in order to find a positive case and, hence, spend more money. The table below presents these calculations where the NNS is based on the *JACC* paper (per-patient and ITD) while the costs for the tests are from a web search from MDsave.com with the exception of CCTA + FFRCT, which is from the Medicare pricing announced by the company.

	NNS (95% CI)	Cost per Test	Total Cost (95% CI)
CCTA + FFRCT	1.67 (1.56, 1.78)	$1736	$2223 ($2084, $2382)
PET	1.20 (1.13, 1.31)	$2000	$2409 ($2272, $2631)
CCTA	1.53 (1.42, 1.66)	$806	$621 ($577, $673)
SPECT	1.13 (1.06, 1.28)	$4162	$1595 ($1493, $1800)

The results show that CCTA is superior to the other testing methods with respect to identifying positive cases. Of particular note is that the CCTA + FFR is on par with the PET scan but beaten by the SPECT measurements. Ultimately, it is difficult to find actual costs and, further how the insurance is negotiating these prices, so I invite whomever to replace the prices with what they believe to be reasonable. There is much more analysis to be done by those who know much more about health outcome evaluation modelers. I would interpret these numbers as a starting point in ascertaining the additional value of FFRCT on top of CCTA. As such, increases in performance, particularly around rejections, can be directly linked to cost-effectiveness.

DISCLAIMER

This piece is an analysis of the company HeartFlow based on my interpretation of publicly available peer-reviewed information. This analysis is intended solely for educational purposes and should not be considered as medical advice. I have no affiliation with HeartFlow, and there are no conflicts of interest to declare.

REFERENCES

[i] https://www.nejm.org/doi/full/10.1056/NEJMe1809203
[ii] https://www.heartflow.com/newsroom/late-breaking-aha-acc-guideline-directed-ccta-ffrct-precise-trial/
[iii] https://academic.oup.com/eurheartj/article/42/37/3853/6352228
[iv] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4893304/
[v] https://pubmed.ncbi.nlm.nih.gov/37743129/
[vi] https://www.sciencedirect.com/science/article/pii/S1875213623001663?via=ihub
[vii] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5710451/
[viii] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6857505/