To: Sunny Balwani[sbalwani@theranos.com]
Cc: Elizabeth Holmes[eholmes@theranos.com]
From: Daniel Young
Sent: Mon 4/14/2014 3:45:15 AM
Importance: Normal
Subject: RE: new draft of my response....
Received: Mon 4/14/2014 3:45:13 AM
email response to Tyler_2_DLY.docx

Looks very good. Please review my comments/additions.

-Daniel

From: Sunny Balwani
Sent: Sunday, April 13, 2014 5:35 PM
To: Daniel Young
Cc: Elizabeth Holmes
Subject: new draft of my response....

With some additions around CV paragraph...

Tyler.

Elizabeth forward me this email to respond to your comments.

Before I get into specifics, let me share with you that had this email come from anyone else in the company, I would have already fired them for this arrogant and insulting attitude. In your case, I am giving you the benefit of doubt that your intentions are in the right place, ~~because of Mr. Shultz,~~ and am taking the time to respond, even though your tone in this email all the way through last paragraph is not seeking to understand, but rather standing at higher perch of morality and judging others in the company. Perhaps this too was not your intent but this clearly comes across as your mind set.

Formatted: Highlight

Comment [EH1]: Stronger – emphasize what he did first especially in light of my comment on accuracy below

In my meetings with Daniel I found that the discrepancies between our CVs were due to Daniel calculating CV based on the median value of each precision run, while I was calculating CV of the entire data set for each level. When I asked him why we do this, he said that it was a way to average out the noise. I was under the impression that the coefficient of variation was meant to be, at least in part, a measure of how much noise exists in the data. By averaging out this noise before CV is calculated, the CV as a metric of assay performance becomes less meaningful. And because our calculations of CV are based on median rather than mean, this means that 2/3 of our data is entirely ignored both when calculating CV and acquiring a patient result.

Your basic understanding of statistics is still low and you do not grasp the meaning of the CV. We happen to be running 6 six replicates inside the device for our own comfort, going the extra mile in our early stages. For almost a decade, when we have run assays in pharma trials, we have run assays only in singlicates~~etons~~ or duplicates and achieve extremely high quality data. As a matter of fact, we have consistently seen that data generated by our platform – when free of human errors – has at least matched if not exceeded the quality from other **laboratories** (more on the point about **laboratories** later). Because we are running assays in 6 replicates, this does~~n't~~ *not* mean this is what is reported. What is reported is one our result.~~s with~~ We get a high degree of confidences ~~generated~~ by our algorithms ~~based~~ which take into account the number of ~~on~~ replicates we choose to run as well as other factors. This is a point you are struggling most ~~with~~ to grasp. Let me further simplify. It is like someone asking you to count M&Ms in a bag. Since we really want to be sure, we have 6 people count these instead of 1, even though the answer we give out is still 1 number and therefore in a large number of cases our answer is of higher confidence versus others who may only be counting the bag using 1 person. This doesn't mean we have to ~~share with others~~publish that we have 6 people counting and 6 answers instead of 1. We use internal algorithms to arrive at the right answer based on ~~what we see fit~~the use case. It may be a simple average of 6, a median, or some other more complex algorithm looking at statistical distributions and expected values. This is ~~a business decision as~~ the purpose of having a world class computational biosciences team to select what statistical model to ~~pick to~~use in order to generate ~~give out~~ the ~~best~~ most robust clinical data with the highest confidence from our *system (assay, hardware, software – including our algorithms, and other elements)*. What matters is that the answer we provide is something we ~~want to~~ have high~~er~~ confidence around. In this example, e~~E~~ach of these 6 people may in turn employ six more people so they can in turn provide the 1 right answer that feeds into our algorithm, which ultimately reports out ~~that picks~~the 1 right and robust answer. Similarly, what goes on inside the device is for internal calculations and verification purposes. The final reported result is all that matters, whether internally we run~~t~~ this on 1 replicate, 2, 6

Comment [DY2]: Consider adding: This was also shown recently when NASA compared one of our assays (one of the most challenging in the field) to leading competitors. They found that no other company could provide the performance that Theranos could.

Formatted: Font: Italic

or 12. It is important for us (ie, R&D, internal data generation, ~~future~~ and the algorithm-, amongst other areas~~generation~~) to know tip to tip variance, ~~so we can constantly refine and improve our product,~~ but not relevant in terms of how we quantify assay precision (CV). Daniel Young went over this in detail with you ~~before~~ over 3 sessions, as you were calculating this incorrectly before, and it still seems like your understanding around this is deficient. Please know these repeat sessions with Daniel himself were given to you solely because of your relationship with Mr. Shultz. We don't share our proprietary methodologies ~~internal workings~~ with other junior employees and certainly not with those that are not involved ~~with~~ in a given~~this~~ process. Moreover, Daniel and his team's time is also extremely previous given the amount of work they do almost 7 days a week. This was a privilege extended to you as a courtesy. Not a right.

While I understand that calculating CV based on the medians is relevant for comparing our system to systems of our competitors, the fact that the CV of our cutoff level for Syphilis RPR drops from 43% to <20% by moving from CV of the entire dataset to CV of the medians tells me that a significant portion of our data is just noise. I believe that we should set two standards of CV that must be met in order for an assay to pass precision testing; a standard for the medians of each run, and a standard for each level's dataset as a whole.

Again, the variance across tips is not relevant as mentioned above- the variance for the reported value is what is quantified to assess assay performance.

Daniel also told me that for qualitative assays such as Syphilis RPR, the CV as metric of assay performance is less important than it would be for quantitative assays. I agree with him, at the end of the day the only thing that's important is delivering the correct result to our patients. However, given the high variation in our dataset, it is not surprising that when using a strict antibody index cutoff value of 1, our sensitivity was only 65% the first time we tested clinical samples and 80% the second time. The first issue I have with this is that there is no penalty for repeating an experiment. We repeat and delete rather than repeat and add. In our validation reports there is never any mention of how many attempts of precision or comparability testing it took to get the data that's presented. The second problem that I have is that our equivocal zone is adjusted and widened until we see the sensitivity and specificity that we want to report. Almost regardless of what the data looks like, we can adjust this zone until we get the 95% sensitivity that we want to see. Tellingly, out of the 247 patients that we tested, 66 of whom were Syphilis positive, more patients fell into our equivocal zone than we correctly diagnosed as being positive for Syphilis.

Equivocal zones are commonly used, and expected in such qualitative assays. The approach being used for settings such as ours is ~~was~~ based on common techniques. That being said, we do know that our equivocal range may be ~~is~~ wider ~~than where we would like~~ than ~~for~~ some other assays. However, in this case, the impact is that more patients will need confirmatory testing. But this is a **business** decision. We make these business decisions all day long. This is not ignoring data. Our recent internal dry-runs for PT for RPR is showing that our test is performing well as indicated in the validation report.

Comment [EH3]: Add text around recent AAP and "borderline" performance of reference method, especially in equivocal zone

No studies are simply repeated with the original data being ignored. There have been times when the initial data sets from initial studies may not be good enough because of many factors including the fact that many of our assays, algorithms, formulas, production methods, QC processes may have been in early stages. In such case when data suggests that to be the case, we improve our products, assays, software, algorithms, hardware, manufacturing processes, and more. We ask relevant teams – sometimes all teams – to identify the root cause of such issues and make changes and repeat~~s~~ our

experiment. THIS IS CALLED PRODUCT DEVELOPMENT THRU ITERATION. In this case when ~~may~~ we learn that our initial experiment was a result of a software bug or algorithms needing further refinement and debugging, misalignment of hardware tools or simply erroneous human processes, we discard that data in the R&D and product development stage. Nothing works the first time around in product development, not in startups, not in larger companies and certainly not when you are doing something extremely novel and unprecedented with limited resources. This is how every game-changing product is developed. I find it appalling that rather than seek and understand, you claim to be judging something you don't have a basic understanding of. I also think you are thoroughly confused about what is product development and what is validation in CLIA. Because of lack of resources, we ask people to do multiple things, wear multiple hats, and sometimes try to combine multiple steps into one. We don't need to explain these to every individual doing the experiments as there is a more senior, more experienced team that does data crunching and decides what experiments to rerun. Most junior level lab associates don't have access to this because they lack ~~any~~ the experience and knowledge. We hired them to run experiments, not do data analysis. This is also why we added you to the ELISA experiments team. Not because you brought a superior understanding of data analysis but because we needed ~~associates~~ people to run experiments. If you wanted to understand data analysis, the emphasis should have been seeking to understand. When you sent your last email, you wanted to understand more and we asked Daniel to spend his time with you. After that session, you shared you understood better. Seems like now you are trying to go even broader and grasp even more and your depth of understand is even shallower than before.

I then asked Daniel if he thought our Syphilis test was truly the most accurate and most precise Syphilis test on the market. He said that Theranos does not claim to have the most accurate or precise tests, and that if I could find any marketing materials that make such claims that I should forward them to him. A quick google search yields a handful of articles that explicitly make these claims. Daniel agreed that the authors make sweeping statements about our assay performances, but noted that Theranos never directly made any of these claims. If well-established institutions such as the Wall Street Journal have published misinformation about Theranos, it seems it would be in our best long-term interest to correct this information in order to uphold our image of bringing transparency to blood testing.

This is the point that irritates me the most. A g~~Q~~uick google search and what this led you to believe is mistaken - without paying attention, understanding and again, seeking to understand, but rather jump to conclusions and judgment.

I saw these articles. These articles claim Theranos is better. I personally agree with that but that is my opinion just like many of these articles are opinions of bloggers and authors. An overwhelming majority of patients who haves experienced our ~~method~~ systems over the last decade, not just at Walgreens, agree with this opinion based on their experiences. When journalists who come and experience what we do say that this is the best way to equip and operationalize a ~~do~~ lab, there is no disagreement around that.

In specific, you mention the WSJ article. Here is what the author says: "Theranos's processes are faster, cheaper and more accurate than the conventional methods and require only microscopic blood volumes, not vial after vial of the stuff." Does the article say Theranos is better than Immulite running in 1 lab on 1 device? This says more accurate than conventional methods. This brings me to other major point that either you don't understand or simply choose to ignore ~~and~~ in makinge your misinformed claims and assertions against the company. Let me shed some light for you.

At this point, Theranos is not selling any devices. We are a high complexity CLIA laboratory. As such in general we are compared to other high complexity CLIA laboratories (read the WSJ language carefully – it says conventional methods which in our case is other laboratories and more specifically, the larger national laboratories that use multiple devices for a given assay in different locations ~~so – towe~~ compare apples to apples). When on our web site we site that as a CLIA lab, our CVs are such and such (we only make 1 claim on Vitamin D – more on this later), we are comparing these to other labs. Do you know what is the typical inaccuracy~~CV~~ for Vitamin-D in other labs? It is usually much higher than 25%. Do you know what is Vitamin-D CV across different devices and different Labs in even 1 company like Quest and Labcorp? It is over 40%. Many of hospital and payer partners routinely tell us they have never see Vitamin-D CV from larger Labs lesser than 50% in their lives. The same applies for an overwhelmingly large number of other assays; the CVs for these is usually much higher when you measure it across multiple devices, different reagent lots and different days. For Vitamin D, the CLIA governing body doesn't even define an acceptable CV (only state of NY does) because this is one of the most volatile and difficult assays. We go through the excruciatingly difficult task of calibrating our lots and batches of devices, assays, reagents, cartridges, plastics, movements inside devices, and other even more difficult things that you have no visibility into, so that we can have a platform that gives us this capability to have tight CVs across ~~devices~~the locations at which people give samples and also across large numbers of devices. We will make these claims against other devices when we start selling our devices.

Comment [EH4]: Add pre-analytical error and variance point – this is what WSJ and everyone else is talking about by the word "accuracy". Second aspect of what they mean by that is variability over time. Read the articles – they expound on exactly this. This is our whole point and mission on "actionable information" and what we ALWAYS talk about when talking about accuracy as you can see from the context around any performance claim, including on our website and in articles.

Add future of personalized medicine point on trending and associated predictive power. Other labs and EMR systems cant and caution against relying on longitudinal data.

The most important thing about measuring CV is not when you are in a controlled, pristine environment where all elements and all processes are controlled, but rather when everything is as close to real world as possible and measuring CV on samples from the moment samples are collected to all the way results are reported (from cradle to grave). That's the only true CV that matters in clinical decision making. And as you may have heard us say several times in the context of actionable information, that's the CV we are able to generate and control. The device and instrument makers don't measure CV that way because they can't; their devices are only in a 'real world' environment when they are in laboratory settings and there they have no control over sample collection, sample transport. As you may have already read, over 90% of errors happen in pre-analytical processing and in our case, these processes are much better controlled and as we grow as a company, these will only get better and for our competition, they only get worse. On the other hand, the commercial high complexity labs never will measure, let alone report these CVs because they have no controls over majority of the process themselves. Equally importantly, it is only when you have this level of confidence in your overall process that you are able to trend such data for personalized medicine – another thing we walk about when we talk about CV. You took an erroneous, myopic – and still incorrect view – what CV is, let alone what the purpose of CV is in general and what we are going after. The smartest physicians, lab directors, hospital CEO, pharma executives etc understand the value of this and that we already are starting from a far superior place. And you are telling and teaching us what we should allow to be published about ~~u~~is in online media. Next time, think!

Again, if you really were seeking to understand and not preach, all of these points would be obvious to you and you would be focusing on and would have understood why everyone who understands the overall laboratory process knows what we do is so much better already and as we build out more software, processes, controls etc, will only get better with time.

~~Moreover~~Regardless of the above more important points, the CV you are comparing is from multiple Theranos devices to 1 competitor device. If you are measuring CVs across multiple devices and multiple reagent lots from the same vendor the results are horrendous. WSJ article in particular makes no

reference to our RPR method to being the best and neither do we. This is called an NP-complete problem. Even when we have data to show we are better than 2 larger labs, we can't prove we are better than *every* lab. It is logistically impossible. This is why we, Theranos, never make this claim on our web site. However, an average person of average intelligence can easily experience and tell you that this is the case.

We also wanted to share with you our own first-hand experience with "top-of-the-line" competitor's devices - an experience that supports the overall sentiment in the laboratory industry. Namely, to support our growing sample volumes, we have more than one of the same devices from our competitors (before we can completely replace these devices with Theranos devices). We are in the process of finalizing comparability/verification of these "identical" devices. We have found that the performance/accuracy/CV across these devices are not adequate for some tests (specifically for lipids) and we are forced to seek corrective actions from the vendor. Unfortunately, most labs are not able to identify these problems like we just did, and patient care suffers as a consequence. Our fully integrated laboratory solutions will vastly improve on this broken and outdated model.

Comment [DY5]: I added this. Please review. This refers to recent data using Advia's 1, 2, and 3.

We believe that our approach, our methodology, our technology, our platform is superior to other high complexity CLIA labs. ~~It~~sIt's not necessarily singularly best in every given moment on every assay on every sample on every day in the hands of any person when compared to a single reference method *device*. We can never prove that, like I said, that's a ~~logictically~~logistically impossible problem to prove. However, in the long term as we generate more and more data, we will make direct claims on our web site. In the meantime, we don't. Anything – every single letter – that we put in our marketing content and on our web-site is vetted by some of the most competent law firms who are subject matters on such claims and contrary to your comments to Daniel – on the implications of those claims. We are a VERY conservative company. We ask these advisors not just to make sure our claims are correct, but also that they don't lead an average person to draw wrong conclusions. We don't take anything lightly. We run a very tight ship on these matters.

Formatted: Underline

I then thought back to our previous discussion when I asked about our claim of having <10% CV for our assays. We checked the Theranos website together and found that we only make this claim for Vitamin D. I checked the 2-Tip validation data (we were running 2-tip protocol at the time) and found that the CVs for our three levels were 18%, 16%, and 19% when calculated based on the median of each precision run and 23%, 23%, and 25% when calculated based on the entire dataset. Here are scatter plots of the results from VitD precision testing, they don't seem to meet the standard we claim on our website for Vitamin D.

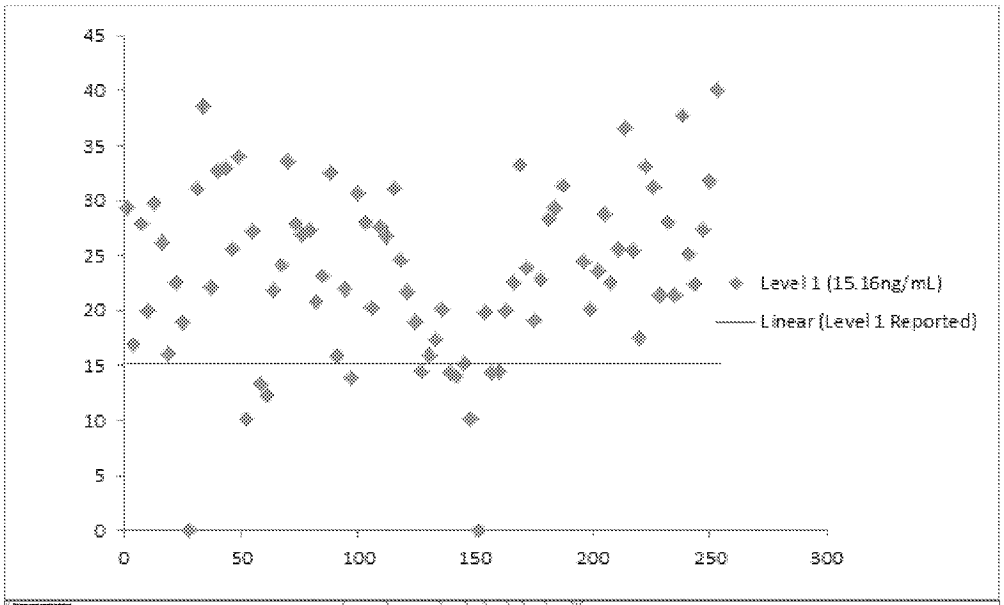
Comment [EH6]: Add point about 10% CV on 6 tip data and when bias is corrected and how CV is calculated – in Fmain range

Comment [EH7]: What data is this? Daniel – add comments

The data below appears to be form the 2-tip precision data, including each tip value. As noted several times, the CV metric across that data in not how we quantify our assay system CV. We have referenced our 6-tip Vit D precision below in reference to the table summaries. Also note that y¥ou are wrong here – m¥Median was not used here – averages are. Tthis is why you are generating wrong numbers.

Comment [DY8]: These appear to be from the 2-tip precision data, including the data from each individual tip, which as noted above is not the result that we report.

Formatted: Font color: Red



~~You are wrong here. Median was not used here – averages are, this is why you are generating wrong numbers.~~

For a while I've been giving our assays the benefit of the doubt until we see how the new 6-Tip method performs. Here is a comparison of the 7 assays we run on Theranos devices to their predicate methods. While we are now performing better than we were with the 2-Tip method, you can see that of the 7 assays we run on the Theranos system, there is only one level from one assay that shows less variation than our competitor's technology.

The below summaries are incorrect largely because your have used the median values, while for all quantitative tests, we use means. Moreover, the algorithm also takes into account the statistical distributions of the data in arriving at the final reportable. Details for each of these tests are below:

For TSH, our 6-tip data resulted in 21.9%, 19%, and 19.7% CVs at the three levels (from low to high levels) – a significant improvement at the low level compared to the 2-tip format.

Immulite 3rd generation TSH		Theranos TSH		
level (ulU/ml)	total CV	6-Tip		
0.016	12.5%	Level (ulU/ml)	**CV whole dat**	**CV medians**
0.32	5.3%	0.02	42.9%	34.1%
1.3	4.6%	2	24.6%	17.9%
3.3	4.8%	20	27.7%	20.8%
7.3	5.1%			
19	4.5%			
39	6.4%			

Comment [EH9]: Add comment on this

Comment [EH10]: Add single device and pre-analytic error/variance point again

Comment [EH11]: Add summary comment on all below is incorrectly taken from whatever he took it from.

Comment [DY12]: I added this – please review and keep in or take out.

Again your number are off. Our precision numbers are instead 8.4%, 3.5% and 4.6%. This is pretty close to the predicate and better in some cases.

Immulite fT4	
Level	CV total
0.51	10.2 %
0.85	7.1%
1.13	6.4%
1.49	6.0%
2.91	3.6%
4.82	3.6%

Theranos fT4		
	6-Tip	
Inter mean	whole dat cv	CV medians
1.63	28.8%	14.5%
5.42	11.0%	4.0%
6.68	5.2%	3.9%

Calculations in the report show CVs values of 19.2%, 9.2% and 7.6%.

Immulite TT4	
level	CV total
1.8	11.7%
2.6	10.8%
5.2	8.5%
7	6.1%
8.2	5.6%
13	6.0%
16	5.6%

Theranos TT4		
	6-Tip	
Level	CV whole Dat	CV medians
1.91	16.0%	13.9%
3.37	16.0%	14.0%
15.8	18.3%	14.6%

Immulite tPSA
"<4.6% for 3 levels of controls"

Theranos tPSA		
	6-Tip	
Level	CV whole Dat	CV medians
1.4 (ng/ml)	33.8%	13.0%
3.37 (ng/ml)	17.1%	10.8%
10.2 (ng/ml)	24.1%	11.8%

Values in the report are 12.4%, 9.4%, and 7.3%.

Diasorin VitD	Theranos VitD

Level	CV
7.2	5.5%
14.7	4.2%
21.7	4.0%
35	2.9%
73	3.2%
62.7	3.1%
93.6	3.2%
115	4.2%
128	4.8%

6-Tip		
Level	CV whole Dat	CV medians
11.7 (ng/ml)	18.6%	12.5%
28.7 (ng/ml)	19.1%	9.5%
73.6 (ng/ml)	12.1%	9.8%

Oraquick HCV	
Sensitivity	99%
Specificity	100%

Theranos HCV	
Sensitivity	99%
Specificity	94%

TST vValues from the report: 7.5%, 6.2%, and 9.3%. Definitely on par with the reported Immulite values, tthough that is not the point here as as I explained above since Immuloite values are calculated only on 1 devices.

Immulite TST	
Level	Total CV
27.1 ng/dL	24.3%
86.1 ng/dL	13.0%
152 ng/dL	10.3%
280 ng/dL	9.1%
414 ng/dL	8.2%
991 ng/dL	7.2%

Theranos TST		
6-Tip		
Level	CV whole Dat	CV medians
90 ng/dL	19.4%	11.6%
300 ng/dL	12.5%	5.1%
1,000 ng/dL	17.4%	13.0%

Furthermore, Theranos has an inherent advantage in these comparisons due to the way we run our precision testing. While our competitors conduct their precision testing over 20 days, we do ours in 5. Accordingly, we can see that our precision experiments are not indicative of longer-term assay

performance once we begin running patient samples; our Daily Quality Control failure rate is far greater than would be predicted by our QC reference range calculations, and our internal comparison of Theranos results in proficiency testing yielded less than satisfying results. I am not sure if this analysis has been done, but we should examine our Daily QC results as if it were a prolonged precision experiment to more accurately evaluate long-term assay performance.

Comment [EH13]: Add comment – this is straight out false. Same comment on QC "failure" – need to explain this, and same comment again on AAP

Comment [EH14]: Add comment – he should not be the one suggesting this and also incorrect comment

You are again wrong here but that is primarily because you seem to starting with the assumption that everything you read on Google and everything other labs publish is word of truthe. It is not.

Let me address each of your 3 insulting accusations here separately.

First, the CLIA regulations define a day as 8 hours, not 24 hours. We run our precision 20 hours amongst multiple shifts. This gives us ~~some advantages~~ an advantage only in number of calendar days it takes us to complete the experiment (calendar day time ~~managagement~~management for the business) but~~and~~ also has some disadvantages, though these disadvantages to you who is starting from a place of doubt don't seem to be obvious or relevant.

Second, our daily QC control "failure" rate is higher because of how we have constructed our QC tests. Other devices in the upstairs lab play tricks with their QC runs and don't bubble up all errors in raw format to the users like we do. Our QC "failures" are not because of reagent stability like you claim. There is absolutely no data that shows that and it is astonishing you are implying that this is the case without any data. Our QC "failures" are because of newness of some of our ~~our overall~~ processes which we are improving every day. Again, if you had started from a place of understanding and intention to help, ~~this is why your email would have been like~~you would not have made the statements you did. Also know that the QC "failures" on Edisons ~~is~~ are because we display all errors to the CLIA technicians so in these early days, we know more, learn more and catch all possible error conditions no matter what the root case may be. We are working on automatic QC software for CLIA that will mitigate the QC ~~error~~ flag messages and only give ~~errors~~ flags that are relevant to CLIA or to patient sample processing. Other devices and vendors have been doing this for decades. We need to write more software to capture and mitigate these which we are doing this quarter. This is product development, this is how startups are built. I find it particularly ~~disgusting~~ disappointing that these facts are also lost on you when you explicitly said you wanted to be in a start up environment that requires building from the bottom up.

Comment [EH15]: Add point here that QC failure is not failure as per my comment above

Let me now address the third point about internal pre-trial PT results. These first internal pre-trial PT tests were for information gathering and process improvement purposes. The purpose of these pre-trial runs ~~was~~ were to test new processes we have been introducing to the CLIA lab, highlight where to focus and improve ~~our~~ those processes, SOPs, procedures and ~~where~~ accordingly~~to~~ focus our software resources. Tthe results of these initial internal pre-trial PT runs was 'less than satisfactory' because we identified ~~few~~ a bugs in our algorithms and software ~~because of~~as this process was intended to do. This was the intent behind this internal pre-trial run. There is absolutely nothing in these internal pre-trial PT results that says that our internal reagent, assay or cartridges stability is at question. To the contrary, ongoing stability studies for reagents, antibodies, controls and cartridges show extended stability and our hardware QC procedures ensure device performance is stable and meet our strict requirements. Even then, during this debugging process, we questioned everything. All teams worked together to reproduce the entire process manually, reran tests, calibrators, and poured through our code to see where bugs might be. We found ~~the~~ a bugs and we are working on fixing the algorithms. ~~More~~ But most importantly, this was an internal pre-trial test run that you are making these very serious statements

Comment [EH16]: Add comment – to the contrary, we tested (list all test) showing no reagent, antibody, cartridge, hardware or other problem

about. This was the purpose of this internal pre-trial test run – to find bugs. For you to use this internal test data that was designed to find bugs in our internal ~~the PT~~ processes and claim ~~that the~~ fault ~~is~~ with assays and question ~~reagent~~ reagent stability is ~~d~~disingenuous~~eeply disappointing~~.

Let me add a final point to this. I saw an email from Daniel this weekend which you sent in February I believe where you questioned the legality our PT method and where you cited a reference from CMS regulations. This is of the utmost seriousness to our business – not only are you questioning our integrity, but our license to operate as a business. That comment and that accusation based on absolute ignorance about the integrity of our company and its core team members is so insulting to me that if you were not Mr. Shultz's grandson, I would have personally thrown you out of this building. This is a privilege you are over abusing – at least at this company. This email from you is the end of this. Only email on this topic I want to see from you is an apology to Daniel and his team and possibly to Elizabeth. Please cc me on this email so I see that this happened.

Formatted: Tab stops: 5.68", Left

Your assumption of being right based on your very limited knowledge and understanding of Math, Statistics, Laboratory industry, medical device precision methods, data - and now laboratory regulations - is a very discouraging news for your own growth in business. Your sense of responsibility may be commendable but your lack of desire to seek accurate understanding, accurate information and contribute but rather ~~your tendency of~~ telling others what right is disturbing.

Comment [EH17]: Tighten/possibly integrate with intro and earlier comments

I am sorry if this email sounds attacking in any way, I do not intend it to be, I just feel a responsibility to you to tell you what I see so we can work towards solutions. I am invested in this company's long-term vision, and am worried that some of our current practices will prevent us from reaching our bigger goals. I'm sorry I wasn't able to catch you for a conversation, I know how busy you are, but if you would like to discuss anything I've mentioned in person, I would be more than happy to do so.

From: Elizabeth Holmes
Sent: Friday, April 11, 2014 4:35 PM
To: Sunny Balwani
Subject: FW: Follow up to previous discussion

From: Tyler Shultz
Sent: Friday, April 11, 2014 3:38 PM
To: Elizabeth Holmes
Subject: RE: Follow up to previous discussion

Hi Elizabeth,

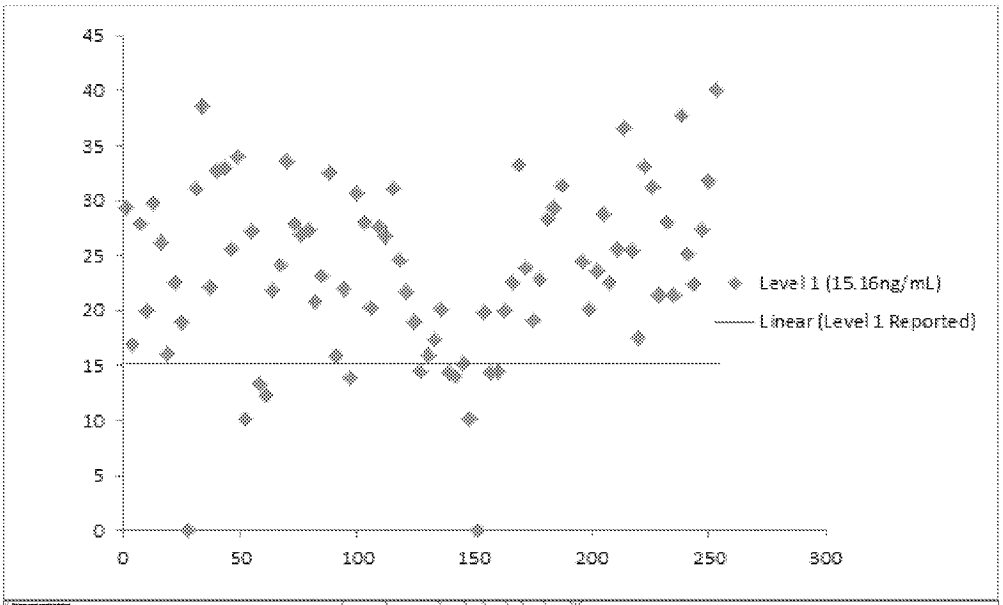
In my meetings with Daniel I found that the discrepancies between our CVs were due to Daniel calculating CV based on the median value of each precision run, while I was calculating CV of the entire data set for each level. When I asked him why we do this, he said that it was a way to average out the noise. I was under the impression that the coefficient of variation was meant to be, at least in part, a measure of how much noise exists in the data. By averaging out this noise before CV is calculated, the CV as a metric of assay performance becomes less meaningful. And because our calculations of CV are based on median rather than mean, this means that 2/3 of our data is entirely ignored both when calculating CV and acquiring a patient result.

While I understand that calculating CV based on the medians is relevant for comparing our system to systems of our competitors, the fact that the CV of our cutoff level for Syphilis RPR drops from 43% to <20% by moving from CV of the entire dataset to CV of the medians tells me that a significant portion of our data is just noise. I believe that we should set two standards of CV that must be met in order for an assay to pass precision testing; a standard for the medians of each run, and a standard for each level's dataset as a whole.

Daniel also told me that for qualitative assays such as Syphilis RPR, the CV as metric of assay performance is less important than it would be for quantitative assays. I agree with him, at the end of the day the only thing that's important is delivering the correct result to our patients. However, given the high variation in our dataset, it is not surprising that when using a strict antibody index cutoff value of 1, our sensitivity was only 65% the first time we tested clinical samples and 80% the second time. The first issue I have with this is that there is no penalty for repeating an experiment. We repeat and delete rather than repeat and add. In our validation reports there is never any mention of how many attempts of precision or comparability testing it took to get the data that's presented. The second problem that I have is that our equivocal zone is adjusted and widened until we see the sensitivity and specificity that we want to report. Almost regardless of what the data looks like, we can adjust this zone until we get the 95% sensitivity that we want to see. Tellingly, out of the 247 patients that we tested, 66 of whom were Syphilis positive, more patients fell into our equivocal zone than we correctly diagnosed as being positive for Syphilis.

I then asked Daniel if he thought our Syphilis test was truly the most accurate and most precise Syphilis test on the market. He said that Theranos does not claim to have the most accurate or precise tests, and that if I could find any marketing materials that make such claims that I should forward them to him. A quick google search yields a handful of articles that explicitly make these claims. Daniel agreed that the authors make sweeping statements about our assay performances, but noted that Theranos never directly made any of these claims. If well-established institutions such as the Wall Street Journal have published misinformation about Theranos, it seems it would be in our best long-term interest to correct this information in order to uphold our image of bringing transparency to blood testing.

I then thought back to our previous discussion when I asked about our claim of having <10% CV for our assays. We checked the Theranos website together and found that we only make this claim for Vitamin D. I checked the 2-Tip validation data (we were running 2-tip protocol at the time) and found that the CVs for our three levels were 18%, 16%, and 19% when calculated based on the median of each precision run and 23%, 23%, and 25% when calculated based on the entire dataset. Here are scatter plots of the results from VitD precision testing, they don't seem to meet the standard we claim on our website for Vitamin D.



For a while I've been giving our assays the benefit of the doubt until we see how the new 6-Tip method performs. Here is a comparison of the 7 assays we run on Theranos devices to their predicate methods. While we are now performing better than we were with the 2-Tip method, you can see that of the 7 assays we run on the Theranos system, there is only one level from one assay that shows less variation than our competitor's technology.

Immulite 3rd generation TSH		Theranos TSH		
level (ulU/ml)	total CV		6-Tip	
0.016	12.5%	Level (ulU/ml)	**CV whole dat**	**CV medians**
0.32	5.3%	0.02	42.9%	34.1%
1.3	4.6%	2	24.6%	17.9%
3.3	4.8%	20	27.7%	20.8%
7.3	5.1%			
19	4.5%			
39	6.4%			

Immulite fT4		Theranos fT4		
level	CV total		6-Tip	
0.51	10.2%	Inter mean	**whole dat cv**	**CV medians**
0.85	7.1%	1.63	28.8%	14.5%
1.13	6.4%	5.42	11.0%	4.0%
1.49	6.0%	6.68	5.2%	3.9%

2.91	3.6%
4.82	3.6%

Immulite TT4	
level	CV total
1.8	11.7%
2.6	10.8%
5.2	8.5%
7	6.1%
8.2	5.6%
13	6.0%
16	5.6%

Theranos TT4		
	6-Tip	
Level	**CV whole Dat**	**CV medians**
1.91	16.0%	13.9%
3.37	16.0%	14.0%
15.8	18.3%	14.6%

Immulite tPSA
"<4.6% for 3 levels of controls"

Theranos tPSA		
	6-Tip	
Level	**CV whole Dat**	**CV medians**
1.4 (ng/ml)	33.8%	13.0%
3.37 (ng/ml)	17.1%	10.8%
10.2 (ng/ml)	24.1%	11.8%

Diasorin VitD	
Level	CV
7.2	5.5%
14.7	4.2%
21.7	4.0%
35	2.9%
73	3.2%
62.7	3.1%
93.6	3.2%
115	4.2%
128	4.8%

Theranos VitD		
	6-Tip	
Level	**CV whole Dat**	**CV medians**
11.7 (ng/ml)	18.6%	12.5%
28.7 (ng/ml)	19.1%	9.5%
73.6 (ng/ml)	12.1%	9.8%

Oraquick HCV	
Sensitivity	99%
Specificity	100%

Theranos HCV	
Sensitivity	99%
Specificity	94%

Immulite TST	
Level	Total CV

Theranos TST
6-Tip

27.1 ng/dL	24.3%	Level	CV whole Dat	CV medians
86.1 ng/dL	13.0%	90 ng/dL	19.4%	11.6%
152 ng/dL	10.3%	300 ng/dL	12.5%	5.1%
280 ng/dL	9.1%	1,000 ng/dL	17.4%	13.0%
414 ng/dL	8.2%			
991 ng/dL	7.2%			

Furthermore, Theranos has an inherent advantage in these comparisons due to the way we run our precision testing. While our competitors conduct their precision testing over 20 days, we do ours in 5. Accordingly, we can see that our precision experiments are not indicative of longer-term assay performance once we begin running patient samples; our Daily Quality Control failure rate is far greater than would be predicted by our QC reference range calculations, and our internal comparison of Theranos results in proficiency testing yielded less than satisfying results. I am not sure if this analysis has been done, but we should examine our Daily QC results as if it were a prolonged precision experiment to more accurately evaluate long-term assay performance.

I am sorry if this email sounds attacking in any way, I do not intend it to be, I just feel a responsibility to you to tell you what I see so we can work towards solutions. I am invested in this company's long-term vision, and am worried that some of our current practices will prevent us from reaching our bigger goals. I'm sorry I wasn't able to catch you for a conversation, I know how busy you are, but if you would like to discuss anything I've mentioned in person, I would be more than happy to do so.

Thanks,

Tyler