Adrian H.: [00:04] Hey, this is Adrian Hernandez, and welcome to the NIH Collaboratory Grand Rounds podcast. We're here to give you some extra time with our speaker and ask them the tough and interesting questions you want to hear most. If you haven't already, we hope you'll watch the full Grand Rounds Webinar recording to learn more. All of our Grand Rounds content can be found at Rethinkingclinicaltrials.org. Thanks for joining.

Adrian H.: [00:27] Hi there, this is Adrian Hernandez from the NIH Collaboratory, and today we have Amy Abernathy who did a recent Collaboratory Grand Rounds reflecting on research at scale, exploring what is possible with high quality real world data, examples from Flatiron Health. So Amy, thanks for joining us.

Amy Abernathy: [00:44] Thank you, Adrian. It's great to be here with you today.

Adrian H.: [00:47] So first, give us a little background. What's the problem that you all have been aiming to solve?

Amy Abernathy: [00:52] So at Flatiron, we have been thinking about how can you simultaneously accelerate research and improve day to day cancer care delivery by having a system of better software, data and the ability to use those data to make decisions on a day to day basis. The real problem that we've been trying to solve is how do you get to high quality, interoperable, readily analyzable cancer data that we can use to analyze how care is being provided and get that information back to cancer doctors every day, as well as use that data to then try and understand what works for whom, and how to make our understanding of cancer treatments better.

Adrian H.: [01:40] Well, you know, it certainly sounds ambitious to both do something better for our care delivery, while also delivering research grade results. How have you guys actually been doing this? What's been the approach?

Amy Abernathy: [01:56] So, from the standpoint of building the datasets and building the underlying infrastructure, our point of view has been that you have to elevate the quality of all of the data to be of an adequate standard to answer all the kinds of questions sitting in front of you at one time. So, whether those are research questions, quality improvement questions, the data itself has to be of a high enough standard it can be used simultaneously for all these different purposes. That's the general premise. The backdrop of how do we do it is through a really two sided business model. The first thing we do at Flatiron, is we build software for cancer care providers. So, we build an electronic health record that's used by about 2,800 oncologists in the United States, so it's the largest community based electronic health record in the US. We also build other software tools such as quality monitoring systems, and other solutions that oncologists at academic organizations and other centers can use our software, and that means that we then tie into their background electronic health record such as Epic or Cerner.

Amy Abernathy: [03:16] Through those software systems, what this then means is that we're in the direct day-to-day workflow of the oncologist, and that we are either the EHR of record, or we are tied into the electronic health record at, for example, a large academic medical center. Regardless of how we get there, this then means that we have full access to the electronic health record for all the cancer patients receiving care in those organizations. That's the first part of what we do at Flatiron. Through the access to the electronic health record, we then pull in the data into one central repository and prepare it in a set of cleaning activities that then get us to datasets that are readily available for analysis. So what I've just told you about building software for oncologists, that is about 50% of Flatiron's work. The other 50% of Flatiron is focused on the data there in the electronic health record, and how do we clean it up for day-to-day analysis? In order to do this, you have to step back and think about, what is an electronic health record?

Amy Abernathy: [04:28] An EHR is comprised of two kinds of data streams. First, there's structured data. This is information that's already available in a digital format, and in general, you can think that you could put it in a spreadsheet if you need to. These datasets require standardization and harmonization because we're pulling across many source systems, but in general, we can now get it into one common format that's readily analyzable. The key challenge is that about 50% of the critical data points that you need for oncology research as well as quality management, exist in unstructured documents. These are digital PDFs that represent the medical case notes, the radiology reports and pathology reports, it might even be the condolence card, which is the best signal that this patient has passed away. There what we do, is we have an entire cadre of human data abstracters who pull these critical data points out and put them in the right place in the overarching dataset. What Flatiron builds is software solutions that allow those human data abstracters, really data curators, to do their job more effectively including efficiency and high quality, consistent work.

Amy Abernathy: [05:44] Our cadre of data abstracters is about 11,000 individuals who are oncology nurses and tumor registrars. They're trained to do this work in a very precise way, and then what we do is constantly evaluate the quality of the data that's coming out of our data curation process, to make sure that it is of a consistent quality to be able to answer the questions at hand. So really the way that Flatiron does this is this combination of building software for oncologists, and then building solutions that allow human curators to do work at scale very efficiently and in a very high quality way.

Adrian H.: [06:44] Just hearing your Grand Rounds and your answer just now about integrating into care delivery has been really important, but then on the other end is insuring quality data. It's certainly notable that you all are aiming to have quality data that can fit a regulatory purpose, or be regulatory worthy. How do you see that for the future for real world evidence and actually clinical trials?

Amy Abernathy: [07:00] Such an interesting question. When we think about the availability of high quality data to answer really important questions, so these are questions that are intended to change the way we treat patients and therefore improve public health, we need to make sure that the datasets themselves are of adequate quality and background to be able to be sure that the answers are accurate and credible. The way that we think about this is first to evaluate the quality of the data itself. So we, on a consistent and regular fashion, describe issues such as data completeness, whether or not variables are reliable, and whether or not the variables are valid in estimating or demonstrating the particular value of interest. So for example, if the data point is intended to indicate whether or not the cancer has progressed, how good is it that we pull the data out of the electronic medical record and do that demonstration? We spend a lot of time describing the validity of the data point itself.

Amy Abernathy: [08:09] When we think about generating regulatory grade datasets, we think about datasets that have a series of features where not only is the dataset highly curated, but we've described the quality of the data that's present and then we've also documented very carefully what can and cannot be done with the data in answering a series of critical questions. So for example, we think about these specific use cases, whether that is to support a label expansion or to understand patients who are not treated in clinical trials because they met exclusion criteria. We then ask, "How good is this dataset gonna be in answering those questions?" And we grade the dataset within the context of those kinds of questions.

Adrian H.: [08:56] Wow, that sounds really important so people have a clear understanding of what purposes they can be used for, and it sounds like a range of purposes. So, there's been a lot that you all have done. What's next? What's next on the horizon?

Amy Abernathy: [09:15] That's really interesting, and now that the datasets are becoming available, and we generate these datasets almost like registry cohorts, so there's lung cancer and breast cancer, et cetera, what's been remarkable is to see how quickly we can start learning new and important things, and so I think that the next part on the horizon is now moving from data cleaning to both analyzing datasets for new discovery right now as well as thinking about, what are the new methods we need on the analytic front in order to be able to glean more and more insights? What we're seeing is these very high quality longitudinal data sets allow us to be ready to answer a whole bunch of questions, but we don't necessarily always have the methods to be able to do that, and so we're thinking about, what's the methodological development that needs to go along with the dataset development?

Amy Abernathy: [10:02] The other thing is, what we now have are huge, large scale labeled datasets. So, these are highly curated, labeled datasets, and provide us now an underlying substrate to start to think about, how do we build machine learning and other algorithms to start to take away some of the work that's currently being done in a very painstaking and manual way by people? So, we're thinking about, where can we leverage machine learning to start to take away some of those tasks? One practical example is oral cancer data, or oral drug cancer data, needed to be hand-abstracted for a period of time, and now we used the labeled datasets to start to train machine learning algorithms, so now a lot of those data points in our data sets are generated automatically. So, these are the things that we're working on next.

Adrian H.: [10:58] Wow, that's terrific. So, we'll look forward to hearing more about that as that evolves. So Amy, I want to say thanks from everyone for joining us on today's podcast, and thanks for listening to this podcast. Our next podcast will be with Aaron Mckethan on policy and priorities: rethinking university research with state data. So, hopefully everyone can join that as well. Thanks again, and look forward to future podcasts.

Adrian H.: [11:31] Thanks for joining today's NIH Collaboratory Grand Rounds podcast. Let us know what you think by rating this interview on our website, and we hope to see you again on our next Grand Rounds, Fridays at 1:00 PM Eastern time.