

Sam Volchenboun Associate Professor of Pediatrics, Dean of Master's Education, Associate Chief, Research Informatics Officer Associate Director, Institute for Translational Medicine, University of Chicago



There is great potential for improvement if pharma companies can use agreed-upon standards for collecting data and for data interoperability

07.09.2021

Tags:

[USA](#), [Oncology](#), [Paediatric Oncology](#), [Cell & Gene Therapy](#), [CAR-T](#), [Data](#)

Dr Sam Volchenboun outlines the progress that has been made in the paediatric oncology field in recent years and his hopes for its future. Dr Volchenboun also explains the important role that data will play in the new era of targeted and personalised therapies and the impact of the work that his research group at the University of Chicago is doing to collect and standardise the data of children with cancer across the world.

Could you begin by introducing your research interests and how you came to focus on data?

I trained as a paediatric haematologist/oncologist and did my fellowship training at the Dana-Farber Cancer Institute in Boston and at Boston Children's Hospital. Having always had an interest in finding better ways to use data, whether data workflows or better use of data for trials, by the early 2000s I had become very frustrated. Therefore, I began some extra training, including a clinical informatics fellowship at MIT, and expanded my knowledge in AI, machine learning, and how to set

up systems for data collection.

Following that, I returned to my hometown of Chicago in 2007 and have been at the University of Chicago ever since. I maintain my clinical practice, taking care of childhood cancer patients, but have concurrently built a research group, [The Volchenboun Lab](#), dedicated to finding better ways of collecting and using data for research. With our flagship project the [Pediatric Cancer Data Commons \(PCDC\)](#), we drive research and cures through the collection, aggregation, and harmonization of disparate data and collaborative data sharing.

The Volchenboun Lab's initial focus was paediatric cancer. It can be very difficult to find enough patients for studies in this field as paediatric cancers are rare, with only a couple of hundred cases in the entire US. Therefore, we dedicated ourselves to collecting data and sharing it, including internationally, which has been a complicated but very rewarding process.

I also spend a lot of time in training and education, building master's programs, including a new program in precision health and precision medicine. I am very interested in building platforms to educate and train the next generation of clinicians.

While many of the blockbuster drugs that the industry is bringing forward are in oncology, how would you characterise the progress that has been made in paediatric oncology specifically?

There is a sense of frustration over the lack of drugs being developed specifically for paediatric cancer indications. This is being addressed with new legislation coming online and an increased focus within some pharma companies on developing therapies for children.

The bulk of our therapeutic options, as paediatric oncologists, are old-fashioned chemotherapies that cause a lot of side effects. Therefore, cellular and immuno-therapies with much lower side effect profiles are obviously tremendously exciting. Even though we have been able to reduce therapy for a lot of kids with cancer with better testing and surveillance, many still receive very intense chemotherapeutic treatment and many still die from their cancers.

I am optimistic for the progression of the field, with new targeted therapies coming in the next ten to 20 years that will hopefully mean doing away with the rather barbaric and toxic chemotherapy regimens.

How big a role will data play in this new era of more targeted therapies?

One of the biggest problems so far is that the deluge of genomic data coming in tends not to be collected with relevant clinical data, much of which sits languishing in the electronic health records. Therefore, the data that come with the genomic data are either that which the investigator collects on basic indicators like age and diagnosis, often presented on a spreadsheet, or clinical trial data, which can be very selective.

We are interested in, firstly, getting all the clinical trial data and linking it up to the genomic data to enrich it, something we have been successfully doing. Now we are talking about initiatives to go back to the electronic health record, get out the actual data on these kids, and try to centralize it so that it can be studied more effectively. That represents an enormous opportunity.

The issue of data is embedded into each country's healthcare system. In this sense, publicly funded welfare states like some in Europe seem to have a big advantage over the privatised and fragmented US system. To what extent can data in its current form realistically be integrated into the decision-making processes of agencies like the FDA?

The FDA has mandated a very particular standard for submission of data, developed by the international group [CDISC](#). Unfortunately, most groups collect their data in different formats before paying a lot of money to have it converted into this new format. I would love to see us collecting source data in the target standard, or at least into a standard that can be easily transformed.

One of the problems we have is that most clinical trials themselves are still written in a word processor when we should be building our trials in a more structured format so that the data contained within them can be formatted automatically. There is a problem all the way back to the source that needs to be solved. Until then, we are just going to keep playing catch-up to transform our data into the preferred format. I foresee a move over the next five years back towards creating trials in a structured format, facilitating better data collection.

The fact that there are only a small number of children with cancer perhaps represents an opportunity to make the data collection process easier. What, however, are the limitations of the data in its current form and how does this play into your decision to expand the international reach of your project?

Obviously, we are thrilled that paediatric cancer is very rare. However, even having started out collecting data in the US through the Children's Oncology Group, which runs clinical trials at over 150 sites, we nevertheless very quickly realized that to be effective, we needed to go international.

Getting data from other countries means more data, but another significant motivation for this international push was helping standardize how we collect and store the data. The way in which clinical trial data are collected and represented has traditionally differed between the US, Europe, and the rest of the world. Therefore, we have attempted to bring stakeholders together, define the exact terms at each stage of the trial process for a specific disease, and map those terms into a dictionary so that going forward, everybody is speaking the same language.

We have now done this for over ten different paediatric tumour types through collaborations in the US, Europe, Australia, South America, and Japan. Once we come to a consensus, these dictionaries are published, so that anyone who wants to use them, can. Even the US National Cancer Institute has adopted many of our dictionaries, stored them, and made them widely available. This will contribute to agreed-upon standards throughout the world.

Once the standards are established, the difficult and time-consuming process of negotiating contracts and data sharing agreements begins. Following the establishment of these agreements, statisticians then need to transform all the data into that standard.

Overall, it is a long and expensive process, but we have managed to collect and transform the data for tens of thousands of patients so far and hope to increase that number in the next couple of years, with the aim of making the data available to researchers who want to use it to connect to genomic data and do other types of research.

What challenges are there in expanding internationally into countries and regions that may lack the infrastructure of the US or Western Europe?

Part of our international efforts are focused on countries and parts of the world that have had difficulties in collecting and using data. Most children in the world who die of cancer are never diagnosed with cancer, meaning that there is clearly a grave problem with how we collect and use data globally. Using standards to collect data and providing better ways to collect and store data could be key for these regions. We hope to play a bigger role in education, training, and resource provision in the next couple of years, taking our work and extending it to areas that might lack the financial resources to do such work themselves.

This will require more funding; we have been very fortunate so far to have received good funding to develop our dictionaries and collect data, but that has tended to be directed at high-income countries that already have a lot of infrastructure in place. One good thing is that even small technological innovations — such as providing a form on a cell phone for a physician to enter some electronic health data or assigning a patient a unique identifier — can have enormous impacts.

A certain level of *enforceability* will no doubt be key to ensuring that the data you are collecting becomes standardized and usable. In your view, what role will regulators have to play?

We are so far taking a grassroots, bottom-up approach, assuming that if we build it, people will start to use it. However, this is not going to work unless regulators like the FDA and EMA start to mandate these standards. Without such mandates, pharma will do whatever is cheapest and easiest; if having a person copy and paste data from one system to another is cheaper than building a better way to collect data, they will continue to do so.

There needs to be a combination of both a —“carrot”— and a —“stick”— approach. In this analogy, the carrot is data that look better and are easier to move and use. The stick is the threat to industry that if they do not abide by these standards, they cannot be involved in certain trials. It is still unclear to me exactly how this will work, but there certainly needs to be some regulation and enforceability of standards.

Has the Pediatric Cancer Data Commons Consortium already started these discussions with regulators and private industry?

We have not been involved in lobbying and are very much a grassroots organisation at this point. Additionally, the FDA has come up with its own rules around standards as mentioned. I am not sure what it would take to push pharma to be more open to a generalized platform for data collection; it has to do with both regulation and finances. Regardless, given that it is not uncommon for clinical trials to be completely adapted manually to whatever the local system is, copying and pasting orders and medications, for example, there is a clear and obvious need for change. I think it will change as more financial incentives are put in place to promote greater agility and speed in opening studies and the better collection of data.

However, I have noticed that parent associations very much want to share their data and be involved in these processes. Having this kind of support from both patients and their families is going to be very important in changing both how we share and collect data.

In real-world terms, what is the ultimate goal in the use of this data?

The idea is that, with larger, more standardised data sets, questions can be asked that could not with smaller data sets. Many paediatric cancer subsets are so rare that thousands of patients are needed just to get to the 100 or 200 patients that are relevant for a study. Paediatric cancer treatments and diagnostics have moved forward with better stratification of patients through molecular and other types of testing, with patients divided into increasingly granular groups for treatments and outcomes. The more data we have, the more people can look at the different groups, their outcomes, and how they were treated to try to come up with better stratification schemes.

For instance, the neuroblastoma group has taken the data in the PCDC and come up with increasingly better ways to stratify patients into risk treatment groups, so that patients who do not need as much therapy are given less, and those who do are given more. That has only been made possible by the large number of patients — currently over 22,000 — that we have put together. Therefore, we are hoping to power these studies that could not be powered sufficiently by a small cohort of patients, once we get patients from all over the world on board.

Additionally, using only patients from one country in studies means that the data will often be very localized to a restricted set of racial and ethnic groups. By collecting data from all over the world, we will be able to better develop models and algorithms that can take into account the global diversity around disease.

The word “algorithm” is potentially anxiety-inducing for consumers and patients. How do you ensure that individual human experiences are not reduced to mere numbers in your work?

We must be very careful about how we use these terms and the data. For example, because of the nature of deep learning algorithms, if a physician used such a method to determine which chemotherapy regimen to use, they would likely not be able to explain to the patient how that recommendation was determined. This would be very confusing for the consumer/patient. Therefore, a balance must be struck between utilising machine learning algorithms in the decision-making process and being able to explain to the patient how decisions are being made. In paediatrics, most of the algorithms being utilised are more like decision trees — choosing options based on whether a patient has certain laboratory findings or genetic markers for example — which is a more visible and explainable format using standardized regression algorithms.

However, as we collect more and more data, explaining our decision making to patients will become more challenging. As an example, we did a study looking at images for neuroblastoma at diagnosis and then at outcome to see if the patient’s outcome could be predicted from the initial diagnostic image. This involved a lot of image manipulation and machine learning. Finally, there were not enough patients to do a full study, but we started to see the possibilities of being able to look at large amounts of data and make predictions and build, in this case, a real machine learning algorithm to do predictions. Once we start to get into those types of analyses, it is going to become increasingly difficult to explain to patients what we are doing.

The education is not just around the maths, but also the ethics and culture of how we discuss these issues and educate the community. All of this needs to be done concurrently.

How much of an issue is the fact that many of the companies developing those systems are pure software firms without operations in the healthcare space?

I run a [master's program in biomedical informatics](#), which takes people with both medical and computer science backgrounds, brings them together, and tries to train them on building systems for healthcare. If computer scientists work on a problem alone, it is likely to miss the mark in terms of usability, while if physicians work alone, the solution may not be scalable and sustainable from an IT standpoint.

Bringing people together and educating them on literacy around databases, algorithms, security, and privacy is very important. Organisations like the American Medical Informatics Association (AMIA) are training physicians and scientists to understand computer science and computer scientists to understand medicine. There is already a new breed of professionals with expertise in both computer science and healthcare, which we hope to see develop even further.

Despite the initial excitement from patient groups around the launch of CAR-T therapies, there remain a lot of unanswered questions. As a clinician, what is your impression of CAR-T?

CAR-T is not my area of expertise, but we do have a CAR-T program at our hospital through which kids have been treated, and I have many colleagues and friends around the country involved in running these trials. There is always initial excitement around new therapies, which is then later tempered with concerns about long term outcomes and side effects. From my point of view, this is another illustration of why we need to collect better data.

For example, we are very interested in collecting data for late effects; not just those in the first year after treatment, but in the following ten or even twenty or more years. These data will help us understand the late effects of these new therapies and how they can be detected. However, there is a challenge in, for example, convincing a 20-year-old to continue giving up their data five years after they have been cured. Creating systems that can collect these data will be crucial. It has never been more important to have standardized ways to collect data directly from patients and their families and then put it all in one place for researchers to use.

One of the arguments being made by industry sponsors about the poor data currently available for CAR-T therapies is that they are only being used as a last resort. What is your take?

That is always the issue with new therapies. For cancers that have an 80 percent cure rate with traditional chemotherapy, it is impossible to propose new, relatively unknown treatments as an alternative to parents as a first option. It is almost always the case that these new therapies are attempted on children where other therapies have not worked. However, this does not necessarily mean that the data are bad, just that more data are needed to make a good interpretation, including data not only from the CAR-T study, but all of a patient's other therapies, how their kidneys were functioning pre-therapy, whether they have received radiation, etc. There are so many variables and the more data we have, the more we are able to control for those variables.

Regulators commonly ask for 15 yearsâ?? worth of data. Is that reasonable, especially in terms of paediatric patients?

For paediatric oncology, 15 years of data may not be enough. For example, 15 years after a two-year-old patient receives treatment they will be 17 and will probably not have had any children, meaning a lack of data on fertility. Moreover, chest radiation during Hodgkin lymphoma therapy creates a risk of breast cancer for women in their 30s, which could be more than 15 years from when a patient was treated. We need to create systems where data can continue to be collected, especially for kids that are treated when they are younger.

The fundamentals of your practice have changed a lot in the past 20 years. What do you see as the most important game-changers coming online today?

Data are still collected in a very old-fashioned way. When electronic health records were emerging, they basically took the old-fashioned way of writing notes and created an electronic version of the same thing. We need to emerge from that into something much more sustainable and granular. As of today, there is no way to automatically ask if a patient has responded to therapy via their medical record; a doctor must read the notes.

The biggest advances are going to be in smarter ways of collecting data. In five or 10 years, the notes will probably be automatically generated from listening devices that might even write the note in advance for the physician or nurse to look at. How we put data into electronic health records is going to change immensely.

The other big change is going to be our ability and desire to collect data from patients when they are not at the hospital via wearables, sensors, surveys, and cell phones. This is going to be critical to understanding what has happened to the patient, how are they sleeping, whether they are being active, whether they are stressed, or if they have a fever, etc. Collecting all these data and getting it into the health record is going to be very helpful.

We are also going to see a movement to get health records out of hospitals and under patientsâ?? control. Patients will want to control their health data, including who is able to see it. Parents get very frustrated when they want a second opinion and have to move all the data from one place to another. A big change is coming in the way people think about how they handle and use their healthcare data.

A system with more nuanced endpoints based on patientsâ?? perceptions also comes with negative aspects, especially in a system like that of the US, where doctors are scored and charge fees according to their reputations. Moreover, as patients take ownership of their data, might they also start demanding equity for its use?

Agencies like the FDA have an important role to play here as they develop endpoints for drugs and whether they work or not. Having endpoints that are less about survival and more about quality and length of life are going to be very important. I am hopeful that we will continue to see the trend towards more nuanced data, not just yes/no or cured/not cured.

The question of whether, as patients take more control of their medical data, they have a right to be compensated for the use of that data is an interesting but difficult question and one that pharma companies are surely thinking about. These data are clearly very valuable when it comes to

research.

The other point around equity is how the research is done and how the results are used. At the University of Chicago we are very focused on our Southside population, a population that has not always been well represented in terms of research and outcomes. We are very focused on diversity and inclusion when it comes to research studies, developing new therapies, and looking at new ways to diagnose and treat patients.

What would you like to share from a physician's perspective with company sponsors that have been working on data?

In one of my presentations, I show a picture of a research office with three laptops, each for a different pharmaceutical company and each using a different system to enter data. There is great potential for improvement if pharma companies can use agreed-upon standards for collecting data and for data interoperability. Even if the companies do not intend to share their data, I would like to see a move towards collecting data in a way that ultimately, once it has been used or been published, means that the data can be shared and then studied by a larger group.

The need for openness extends beyond pharma to electronic health record companies as well. They also bear a responsibility to allow us to collect data in a standardized non-platform-specific way. Traditionally, proprietary records made it difficult for patients to move around — something that was perhaps good for individual companies or healthcare providers on a business level but is not good for patients or for research.

[To learn more about the work of the Volchenbom Lab and the Pediatric Cancer Data Commons \(PCDC\), click here](#)

[See more interviews](#)
