SimLEARN

Excellence in Veterans' Healthcare

EHT

A DEEP DIVE INTO SYNTHETIC DATA, MDCLONE, AND ITS BENEFITS FOR VA

CLIN 0011

Monthly Report of EHT Opportunities 6.3

MAR 2021

LETTER FROM THE EDITOR

Good Day VA!

Welcome, from SimLEARN's Emerging Healthcare Technology Integration (EHTI) portfolio!

We're delighted to be featuring Amanda Purnell, Ph.D., Clinical Data Specialist, VHA Innovation Ecosystem, in our March edition of the EHT Opportunities eZine! Specifically, we'll be diving into the fantastic work she is doing with synthetic data and the MD Clone project.

Amanda has been working as a Clinical Data Specialist for the Care and Transformational Initiatives (CTI) Portfolio of the VHA Innovation Ecosystem since October 2020. Prior to that year, she was the Senior Innovation Fellow, working on projects to democratize data to serve Veterans.

She has also held roles as an Innovation Specialist within Innovators Network since October 2016. She loves to empower others, ignite curiosity, and wonder about what's possible. Prior to her work in the VHA Innovation Ecosystem, she held leadership roles in implementing programs for prevention, integrative and complimentary care, and health behavior change at the VA. She has extensive experience and training in facilitation and advancing learning in adults. She has a PhD in Counseling Psychology from The Ohio State University.

We look forward to bringing you more future content like this from the very best and brightest VA has to offer!

Brian K. Stevenson
Brian.Stevenson@va.gov
AD of EHTI, SimLEARN (14HIL2)
Gulf War Veteran, USN

Table of Contents

EXECUTIVE SUMMARY

Through their partnership, the Veterans Health Administration Innovation Ecosystem (VHAIE) and MDClone are implementing an analytics platform with built-in synthetic data generation technology. This collaboration has the potential to significantly improve practitioners' ability to explore and learn about Veteran patient health and optimize care.

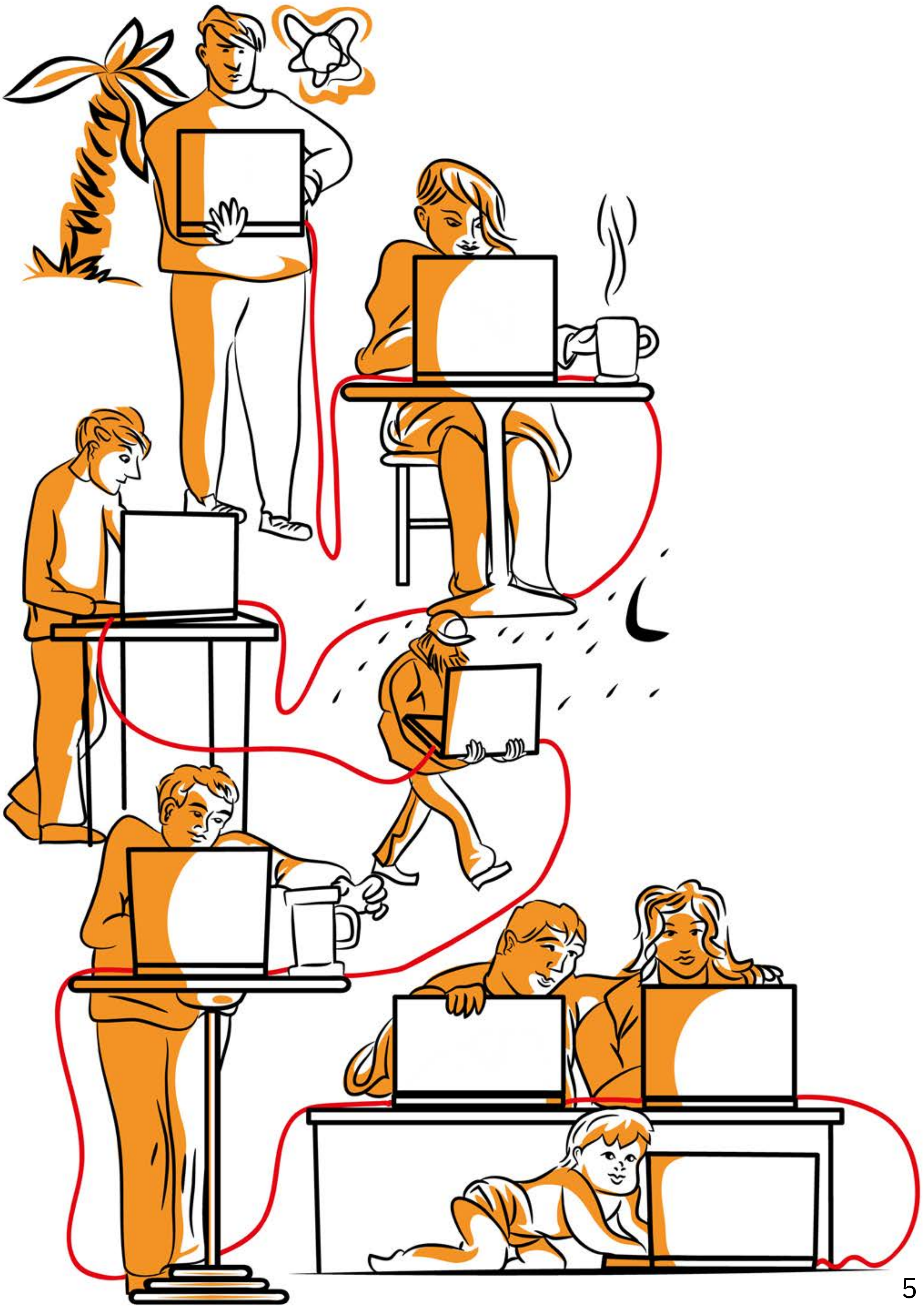
Our report begins with a high-level overview of the VHAIE-MDClone partnership, MDClone and its key benefits, and synthetic data. Read on for a Q&A with Dr. Amanda Purnell and the MDClone team who are pioneering this innovative solution. You are highly encouraged to reach out and join their effort! The more insight from actual users, the better!

IN COLLABORATION WITH

MDCLONE

VHA INNOVATION
Ecosystem

Care &
Transformational
Initiatives



VHAIE AND MDCLONE PARTNER TO DEMOCRATIZE DATA

The Veterans Health Administration Innovation Ecosystem (VHAIE) announced its partnership with MDClone, a digital health company, in December 2020, highlighting their collaboration to democratize data in VHAIE.

This effort is spearheaded by Dr. Amanda Purnell, Clinical Data Specialist at VHA. She is working in collaboration with VHAIE's Care and Transformational Initiatives (CTI), to provide unprecedented, secure access to clinical data to better understand and improve the health of over nine million Veterans.

The MDClone platform was designed to align with the movement of an entire project, not just a piece of it. Users of all levels will be able to use MDClone to ask important questions in real time and gain actionable insights with dramatically shortened timelines for quality improvement, innovation, and clinical research.

Most significantly, Veteran's privacy and health information will be protected via MDClone's synthetic data generation technology, which surpasses traditional de-identification methods.

With MDClone, VHAIE aims to broaden access to clinical data and empower its staff, accelerate transformation, and elevate teaming with external agencies, providers, and industry – all of which can positively impact the lives of Veterans nationwide.

WHAT IS MDCLONE AND WHY USE IT?

MDClone offers an innovative, self-service data analytics environment powering exploration, discovery, and collaboration throughout healthcare ecosystems, cross-institutionally, and globally.

Their platform allows users to overcome common barriers in healthcare in order to organize, access, and protect the privacy of patient data, while accelerating research, improving operations and quality, and driving innovation to deliver better outcomes.

Simulating real-world processes, MDClone enables:

- **Time Savings:** Set up a query and get results in minutes. For moderately complex studies, allot a few hours, not 6-8 months, as is typical for traditional research processes.

- **Ease of Use:** Free form querying means results are based on actual content, not how things are related to each other. Users can adjust their query at will. No expertise required.

- **Security and Privacy:** Data is not seen until the last step. Users will have the opportunity to pull synthetic data. Only those users with special access can pull original data.

Founded in 2016, MDClone works with health systems, payers, and life science companies in the U.S., Canada, and Israel.

中性脂肪
HDLコレステロール
LDLコレステロール
GOT(AST)
GPT(ALT)
γ-GTP(γ-GT)
ALP
空腹時血糖
肝機能
5.5%
(一)
7.0mg/dl以下
男:38.0～48.9%
女:34.0～43.9%

SYNTHETIC DATA EXPLAINED

Synthetic data differs from real data in that it is not linked to real people, events, or circumstances. Instead, it is generated via computer programs that mirror real-world information by observing attributes of real data. Synthetic data sets can be reliable and accurate when based on trends and traits of real data.

Simply put, synthetic data maximizes data utility without the risk to patient privacy.

Why use synthetic data?

- Explore data independent of Internal Review Board (IRB) constraints
- Access data instantly
- Explore data dynamically
- Maintain patient privacy
- Share data worldwide

What are the benefits of synthetic data for clinical research?

Synthetic data sets contain:

- None of the actual individuals from the original data set
- Same statistical characteristics as the original population across attributes and within sub-groups
- Same format as the original data
- Same suitability for analysis, but without privacy concerns

Why use synthetic data for clinical research?

- Synthetic data maintain variables and inter-variable correlations
- No need to remove data elements (like de-identified data)
- Patients cannot be reidentified (no one-to-one correspondence)
- Hypotheses can be validated before seeking to use original data

Source: "Driving Real-World Clinical Research with Synthetic Data," MDClone webinar

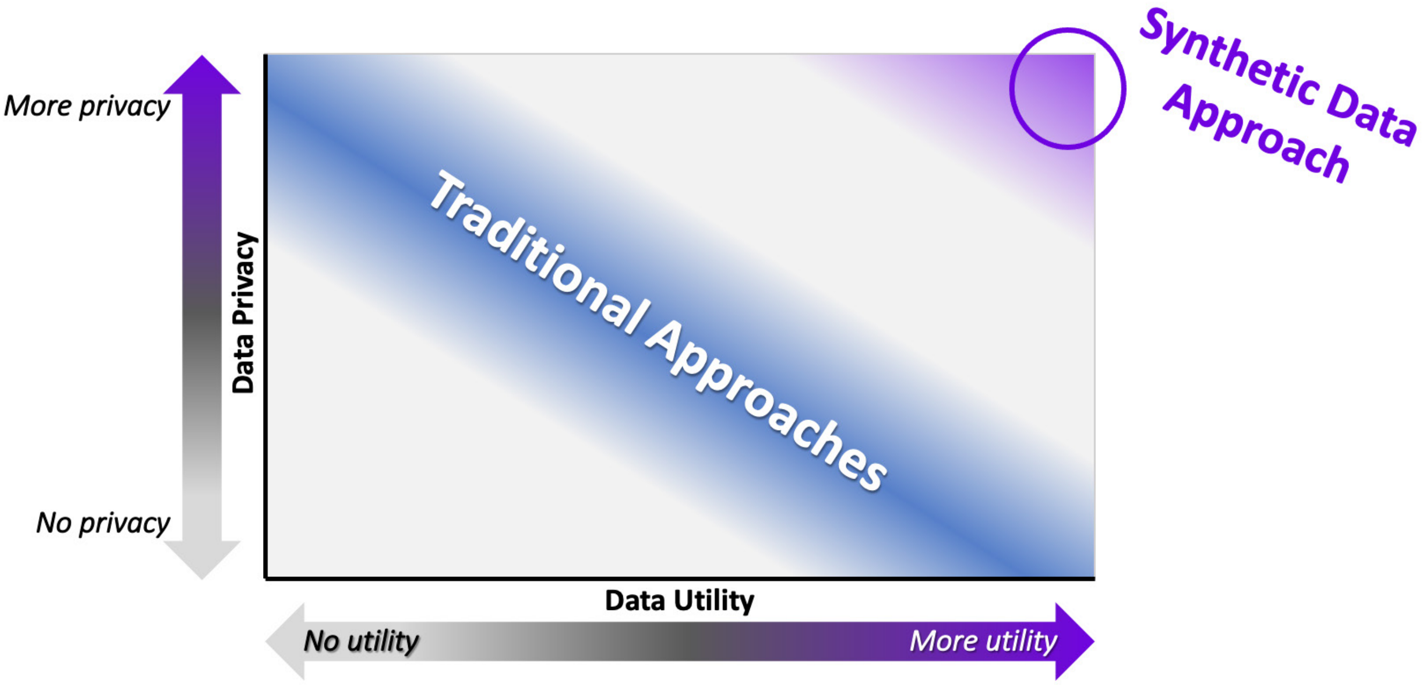


Image: MDClone

“Simply put, synthetic data maximizes data utility without the risk to patient privacy.”

- MDClone

Synthetic Data ≠ De-identified Data

Synthetic data is not the same as de-identified data. Synthetic data is populated with novel synthetic patients who are not real people and instead reflect the statistical properties of the overall population. The data display has the same overall shape, meaning drawing conclusions and relationships between variables is maintained. On the other hand, de-identified data has obscured characteristics and the data removed cannot be regenerated.

Let's look at the visual example to the right. When comparing the original data to the synthetic data, you will see similarities. However, the woman in the synthetic data population wearing a dark purple shirt and orange hair is not real. She is a synthetic person generated with overall characteristics from the original data. In contrast, de-identified data has identical descriptive features but also unrecoverable missing data.

Synthetic Data: Close Estimate to Real Data

Synthetic data enables broad access to diverse users and rapid, safe, and repeatable analysis of data in hospitals or other health organizations where patient privacy is a primary value.

Use of synthetic structured data provides a close estimate to real data results and is thus a powerful tool in shaping research hypotheses and accessing estimated analyses without risking patient privacy, according to published studies like "Analyzing Medical Research Results Based on Synthetic Data and Their Relation to Real Data Results: Systematic Comparison From Five Observational Studies" in JMIR Medical Informatics.

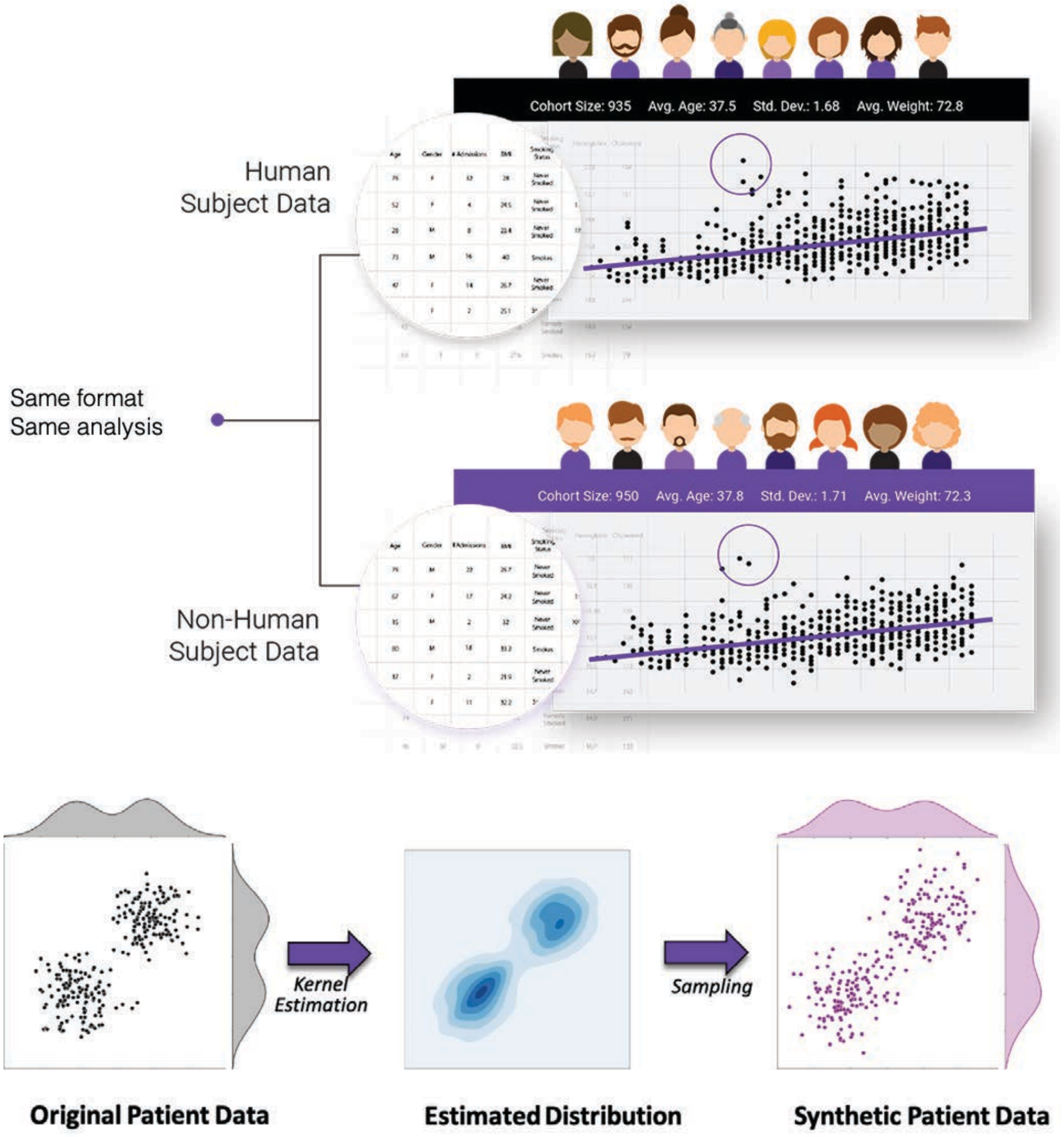

Cohort Size: 935 Avg. Age: 37.5 Std. Dev.: 1.68 Avg. Weight: 72.8
Human Subject Data
Same format Same analysis
Cohort Size: 950 Avg. Age: 37.8 Std. Dev.: 1.71 Avg. Weight: 72.3
Non-Human Subject Data
Kernel Estimation
Sampling
Original Patient Data
Estimated Distribution
Synthetic Patient Data

Image: Randi E Foraker and Sean C Yu in JAMIA Open

WHY IS VA USING SYNTHETIC DATA?

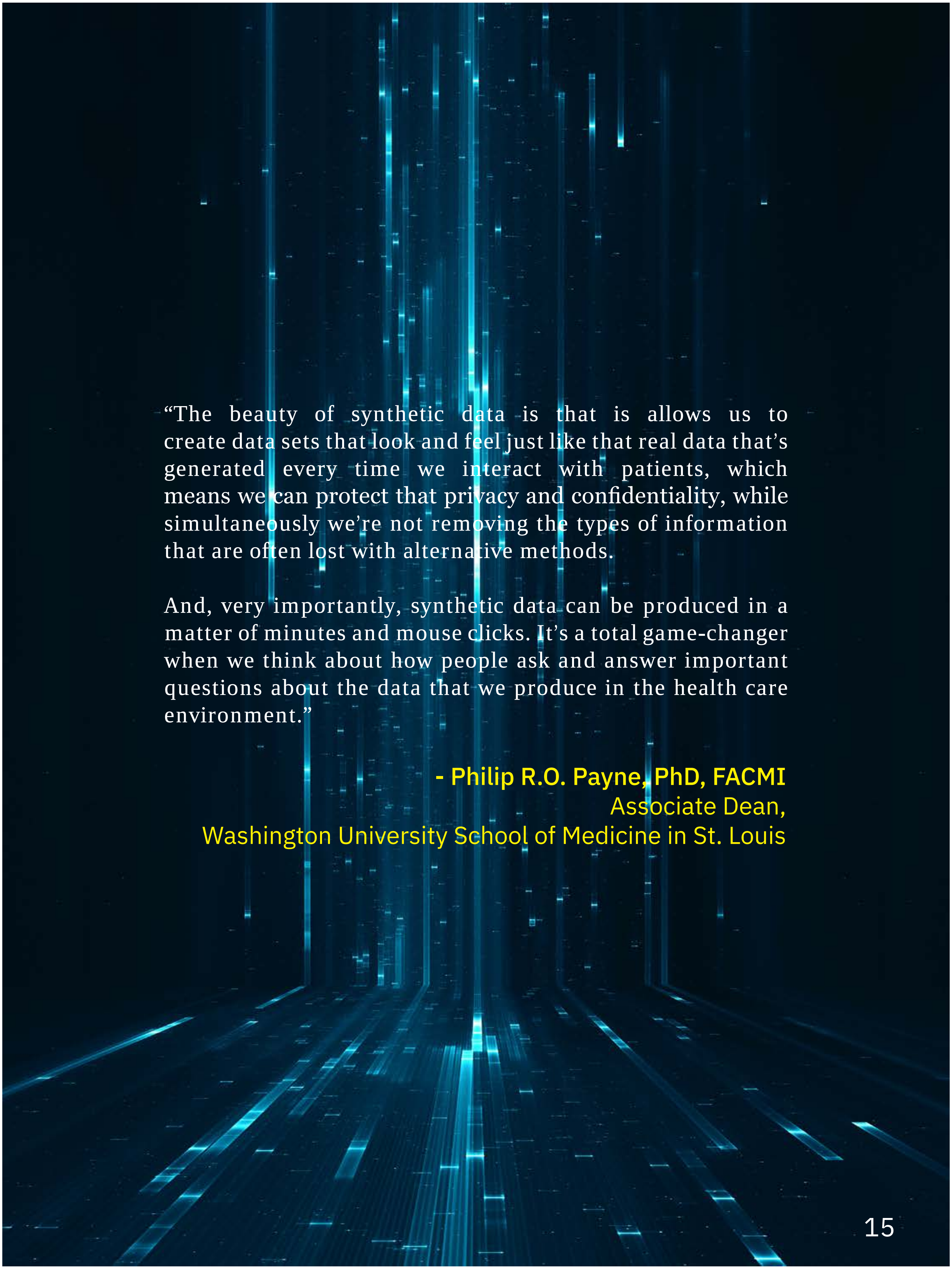
Synthetic health data has all the characteristics of health records – such as information about blood pressure, diabetes, weight, and illnesses – with the main difference being it is artificially created, as the name suggests, rather than generated by actual events.

Synthetic data is different from de-identified data, which simply hides attached identities but still uses real Veteran information. It has been shown that de-identification is time consuming, difficult, and may still carry the risk of re-identification with modern computer analysis techniques. With synthetic data, there are no identities to hide.

VA protects Veteran health data and privacy, meaning there are multiple steps involved for anyone who wishes to use Veteran data for their work. This can often hamper innovation and efficient development. For quality improvement leaders, researchers, and medical providers who need to quickly access Veteran health information, especially those on the frontlines of the COVID-19 pandemic, synthetic data can be a reliable, private, and quick alternative to real data.

Synthetic data allows data scientists to simulate events or circumstances that have not yet happened in the real world but might in the future. Where real data does not exist, synthetic data can create and test how different interventions may work if certain real-word events happen, like a future pandemic.

Source: Excerpts from VAntage blog, 7 December 2020



"The beauty of synthetic data is that is allows us to create data sets that look and feel just like that real data that's generated every time we interact with patients, which means we can protect that privacy and confidentiality, while simultaneously we're not removing the types of information that are often lost with alternative methods.

And, very importantly, synthetic data can be produced in a matter of minutes and mouse clicks. It's a total game-changer when we think about how people ask and answer important questions about the data that we produce in the health care environment."

- Philip R.O. Payne, PhD, FACMI
Associate Dean,
Washington University School of Medicine in St. Louis

Q&A CHAT WITH DR. AMANDA PURNELL AND THE MDCLONE TEAM

This content was collected from an interview on March 10, 2021 and has been organized and paraphrased for format and readability.

Contributors include:

- Dr. Amanda Purnell, Clinical Data Specialist, VHAIE
- Arifa Nathoo, Customer Success Partner, MDClone
- Dr. Vinod Aggarwal, Clinical Expert, MDClone
- Sean Watford, Data Scientist, Booz Allen Hamilton

JOIN THE MDCLONE EFFORT AT VA!

If you want to learn more about the VHAIE-MDClone project, please contact Dr. Amanda Purnell. Her team aims to have an application that users can log into and test this spring. Read the VAntage blog article on MDClone to learn more!

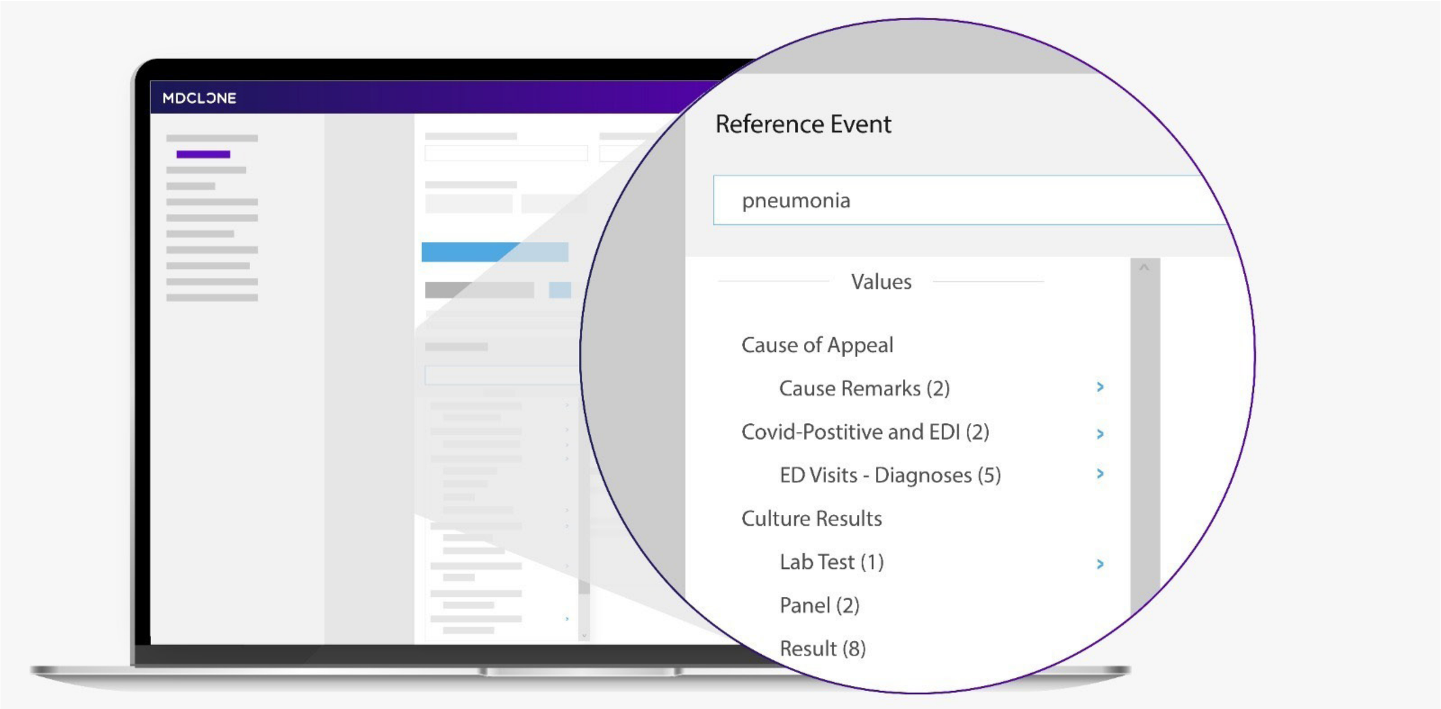


Image: MDClone

Q1

VHA's initial collaboration with MDClone will center around suicide prevention, chronic disease management, precision medicine, health equity, and COVID-19. How did those topics come about?

The topics comes from questions that evolved over time as different researchers are curious about what's possible with synthetic data. The loading of data also influences what questions can be asked. Based on our Authority to Operate (ATO), we have access to the VA data warehouse. So, questions and research need to be connected to that resource and that has helped us define things over time.

Being in a pandemic right now, there are increasing curiosities about the impact of telehealth and how it is influencing health and well-being. Could we compare that to other modalities of care? Do we see that telehealth might have a positive impact on overall wellness and reduce in-person visits?

Q2

What is the latest update with implementing MDClone?

We are implementing the MDClone data lake now. We are also building out the data library, our questions, and identifying what information is needed to answer them. We are planning to give access to super users this spring. Data will continue to be ingested into the platform to ensure that it is current.

One of the things that is revolutionary about the MDClone platform is to almost immediately, without needing to be a data scientist, be able to wonder what other data is needed, return to the home page, and add it to the original query - allowing many iterations.

> “We are very excited about the opportunity for users to use the platform with VHA data and to see where they find value to understand what additional data is needed in the data lake to ask additional questions.”

- Dr. Amanda Purnell
Clinicial Data Specialist, VHA

Q3

What are challenges in creating the MDClone data lake?

This is a chronic problem in every single health care's data. There are missing values, uncertainty in exactly how a particular table is defined and how it relates to other tables, and disparate data sets. We are having thoughtful conversations around what this data means to somebody who is not already embedded and transform it into a model that makes sense.

We think about our life and other people's lives in a chronological timeline. Many health care data sets are developed with many relationships and are not built out in an event timeline model. So, we are transforming the data into this new model, which will allow for different kinds of analysis, including synthetic data generation.

"The chronological timeline feature is one of the benefits of MDClone. There are a lot of ways to query and access data but how they can work with that output file can be very difficult. MDClone is designed to be used by a wide range of uses – adopting an "any question, any person, any time" model."

- Arifa Nathoo,
Customer Success Partner, MDClone

Q4

How do you see MDClone's data making an impact on the implementation of Cerner at VA, if at all?

MDClone can be a testing environment for what kinds of information would be valuable to embed in the health record and in what ways, and to whom. You don't want to immediately change the health records for millions of users, so MDClone allows for safe exploration in subsections and building it out from there, before going into production.

We are not directly affected by the Cerner implementation; however, we are able to inform how data can be collected by providing this synthetic data space for people to try out a new method and understand how exactly it affects a health record – without affecting any health records.

The Cerner implementation is capturing information at the facility level and that information gets rolled up to the national-level warehouse that we are connected to. As Cerner gets deployed across more facilities, the information is captured in a format that will match how we need it and is pulled into the warehouse, which is where MDClone also pulls data.

Q5

At SimLEARN, we look at how we can leverage resources within our portfolio to support learning. MDClone and Cerner interoperability can be used for testing. Is there an opportunity for training?

Training is one of the strongest use cases.

When first discussing this project with many stakeholders, one of the first topics that came up was how MDClone would be fantastic for the thousands of medical trainees who come through the VA and want to work on a quality improvement project but might not be able to quickly get access and make their way through the IRB.

While MDClone does contain protected health information (PHI), it provides a safe environment because the synthetic data generated for the user does not contain PHI. In addition to medical trainees, students can test an idea, develop a hypothesis, and perhaps pilot a quality improvement to see the impact without having to go through a lengthy IRB process to get access to the data. We can easily involve more users, such as short-term staff, and engage them in improving the quality of care too.

Q6

When you proposed the idea of synthetic data or MDClone to the VA, how was it received and how did you present that?

Initially and now, there continues to be education about the difference between synthetic data and de-identified data.

De-identified data is original data with unique identifiers removed. Synthetic data is artificial data derived from original data but does not contain identifiable information.

For those who understood the distinction, the next question was people's concerns about privacy and realism. There's a belief that if data is completely private, it won't be very useful, and if it's so realistic, then it can't possibly be private. So, as I began speaking across VHAIE, it was about the possibility of synthetic data being both realistic but not perfect – because that would violate privacy, and that it has clinical utility – but does not share Veteran identifiers.

After sharing this idea for the past few years, it's recently gained traction as people see synthetic data technology improving. Academic publications evaluating synthetic data are available and support what we are trying to do, so more and more potential stakeholders are curious.

Q7

How do you envision the 'perfect implementation' of MDClone, once it is fully matured and integrated?

My goal after implementation is to have a process that enables many different diverse users, including nurse care managers, clinic managers, individual care providers, medical center directors – to name just a few, to easily get feedback about how things are going in the work that they do.

I really believe that MDClone could truly democratize access to data by allowing a variety of people – not just your typical data scientists, but the average user to be curious about the state of what they are doing and how they could do better. They could create small-scale experiments.

It could also make partnerships and innovation projects easier to test and execute. Right now, there is a lengthy timeline for executing various types of partnerships. The reality is that it takes upfront time now to build the infrastructure for that vision to be executed; I maintain that it is possible.

Q8

What features will MDClone have available in the future?

There is the main query tool, which is what is being implemented at the VA right now. A new version is in progress and will expand the functionality by allowing increased customization and making calculations so that the output can provide analytic insights.

Looking ahead at data visualization, where you usually need a data science background, MDClone also includes features that will allow users to use the data in the system with a built-in analysis package and basic-to-medium level of visualization options.

We want to support how a user can “action” what they learned. In addition to research, we promote the use of data for operational changes. For example, you did a query, received data, and conducted analysis. Now you can ask, “What can I do with this? How can I turn this into an action or an intervention?”

When ready, these capabilities could be rolled out at the VA as well.

Q9

New tools and capabilities require training and development. What topics are covered in the MDClone curriculum?

We are installing MDClone Version 5.5 at the VA. There is an online module called VHA Basic Query Tool Training for first-time users. It has an overview about the basic functionality, interface, and data loading and extraction workflow (see below), as well as a training environment where you can practice the skills that you are currently learning and then explore on your own.

Version 6.0 training is being updated and will provide training for advanced users of MDClone. Currently, if a super user wants to learn something, they can get one-on-one help from MDClone.

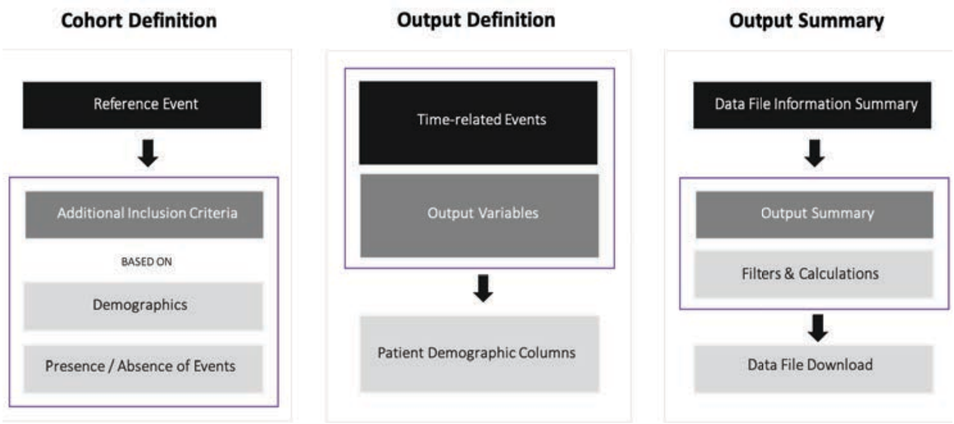
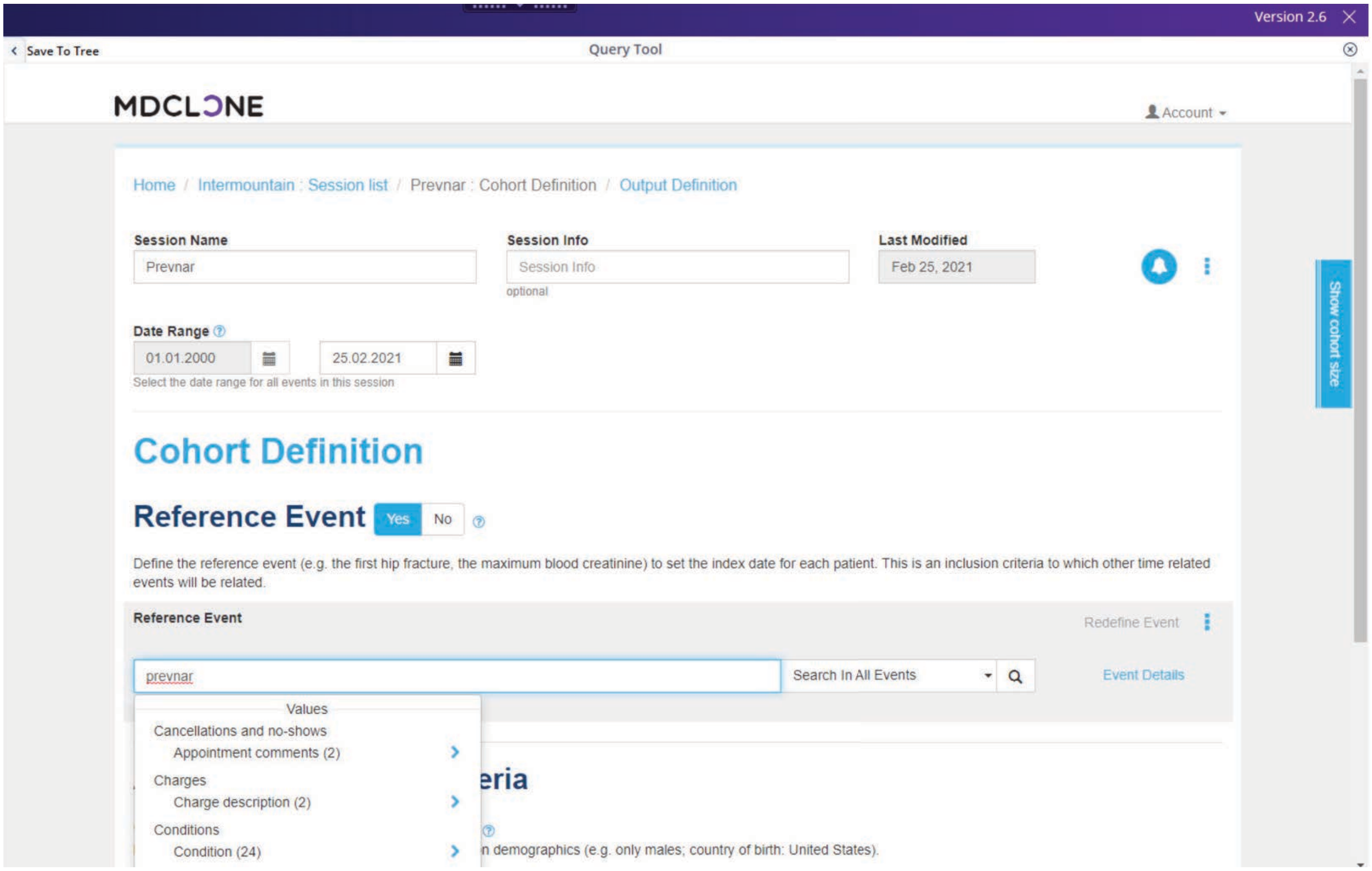


Image: MDClone



Home / Intermountain : Session list / Prevnar : Cohort Definition / Output Definition

Session Name: Prevnar

Session Info: Session Info
optional

Last Modified: Feb 25, 2021

Date Range: 01.01.2000 – 25.02.2021
Select the date range for all events in this session

Cohort Definition

Reference Event

Define the reference event (e.g. the first hip fracture, the maximum blood creatinine) to set the index date for each patient. This is an inclusion criteria to which other time related events will be related.

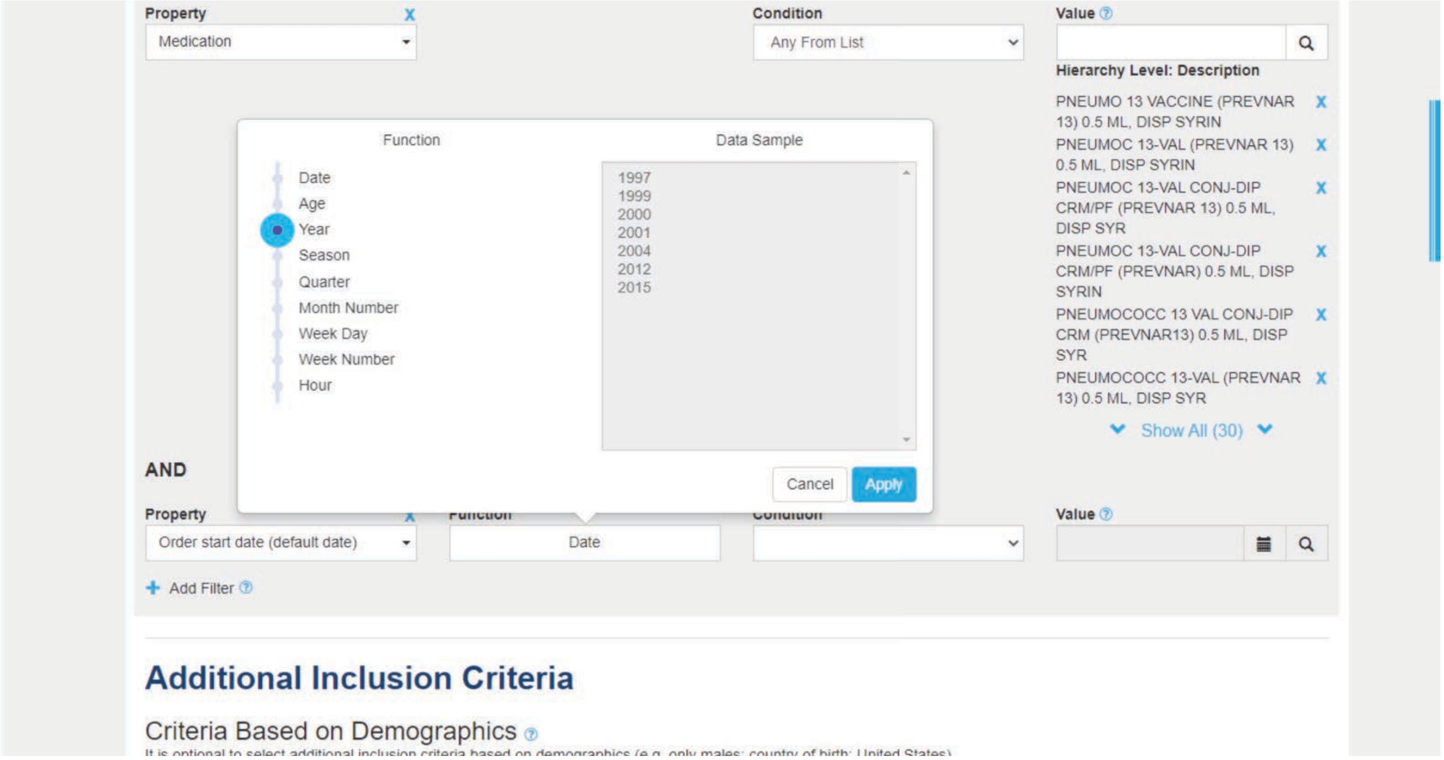
Reference Event

prevnar

Values
Cancellations and no-shows
Appointment comments (2)
Charges
Charge description (2)
Conditions
Condition (24)

eria

n demographics (e.g. only males; country of birth: United States).


Property: Medication

Condition: Any From List

Value

Hierarchy Level: Description

- PNEUMO 13 VACCINE (PREVNAR 13) 0.5 ML, DISP SYRIN
- PNEUMOC 13-VAL (PREVNAR 13) 0.5 ML, DISP SYRIN
- PNEUMOC 13-VAL CONJ-DIP CRM/PF (PREVNAR 13) 0.5 ML, DISP SYR
- PNEUMOC 13-VAL CONJ-DIP CRM/PF (PREVNAR) 0.5 ML, DISP SYRIN
- PNEUMOCOCC 13 VAL CONJ-DIP CRM (PREVNAR13) 0.5 ML, DISP SYR
- PNEUMOCOCC 13-VAL (PREVNAR 13) 0.5 ML, DISP SYR

Function	Data Sample
Date	1997
Age	1999
● Year	2000
Season	2001
Quarter	2004
Month Number	2012
Week Day	2015
Week Number	
Hour	

AND

Property: Order start date (default date)

Function: Date

Condition

Value

Additional Inclusion Criteria

Criteria Based on Demographics

It is optional to select additional inclusion criteria based on demographics (e.g. only males; country of birth: United States)

Image: MDClone

Q10

Are there things that the broader VA community can do to support you from a partnership perspective or with MDClone development at VA?

Absolutely. The more users that we have that would be interested in testing out the platform and providing us feedback about data they want in our data lake, the better.

In addition to training, another opportunity for synthetic data is partnerships. We have strong partnerships with program offices within VHAIE, such as the Office of Mental Health and Suicide Prevention.

The Office of Connected Care is one of our strongest stakeholders. It is extremely difficult to share information, even in government, and synthetic data can help us partner more effectively with Department of Defense, FDA, the Census, and even academic institutions or industry who might have compelling expertise that we don't have to solve problems. It allows us to wonder about a potential solution before we get into original data.

I want to see if there are more people who want to help us make this tool work. With more diverse users, we get different kinds of feedback from different kinds of stakeholders. I want to be open to the potential for the best and most impactful use case being one I never thought of.

WORKS REFERENCED

- Morrow JD, Foraker RE, Greenberg J. “Driving Real-World Clinical Research with Synthetic Data – Featuring Institute of Informatics, Washington University in St. Louis,” MDClone, watched live webinar on 25 February 2021 and available as recording at https://www.mdclone.com/webinar-library/driving-real-world-clinical-research-with-synthetic-data

- Foraker, Randi E., et al. “Spot the difference: comparing results of analyses from real patient data and synthetic derivatives,” JAMIA Open, Volume 3, Issue 4, December 2020, pp 557-566, doi: 10.1093/jamiaopen/ooaa060, https://academic.oup.com/jamiaopen/article/3/4/557/6032922

- Guo, Aixia, et al. “The Use of Synthetic Electronic Health Record Data and Deep Learning to Improve Timing of High-Risk Heart Failure Surgical Intervention by Predicting Proximity to Catastrophic Decompensation,” Frontiers in Digital Health, 7 December 2020, doi: 10.3389/fdgth.2020.576945, https://www.frontiersin.org/articles/10.3389/fdgth.2020.576945/full

- “How synthetic data will improve Veteran health and care: Using artificial data allows for easy access to reliable data on Veteran health,” VAntage, 7 December 2020, https://blogs.va.gov/VAntage/81908/synthetic-data-improve-veteran-care/

- Johnson, Mark. “The Future of Synthetic Health Care Data,” Forbes, 3 August 2020, https://www.forbes.com/sites/forbestechcouncil/2020/08/03/the-future-of-synthetic-health-care-data/?sh=1f529ee17b93

- “MDClone Partners with VHA Innovation Ecosystem to Provide Better, Smarter, Faster Healthcare to U.S. Veterans: Collaboration Will Enable the Country’s Largest Integrated Healthcare System to Democratize Data to Address Veteran’s Unique Care Needs Related to Suicide Prevention, COVID-19, and Beyond,” MDClone, 16 August 2020, https://www.mdclone.com/news-press/articles/mdclone-partners-with-vha-innovation-ecosystem-to-provide-better-smarter-faster-healthcare-to-us-veterans

- Reiner Benaim A, et al. “Analyzing Medical Research Results Based on Synthetic Data and Their Relation to Real Data Results: Systematic Comparison From Five Observational Studies,” JMIR Med Inform 2020;8(2):e16492; https://medinform.jmir.org/2020/2/e16492; DOI: 10.2196/16492)

SimLEARN
Excellence in Veterans' Healthcare

Thank you for reading!

Please join the SimLEARN EHTI community and share your ideas with us by visiting our Teams Site below:

Visit the EHTI Teams Site