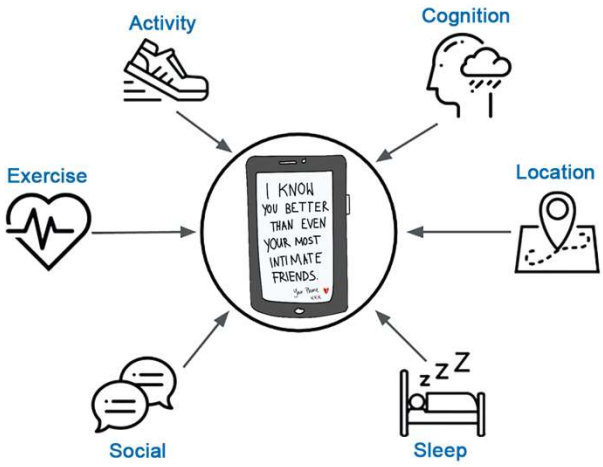
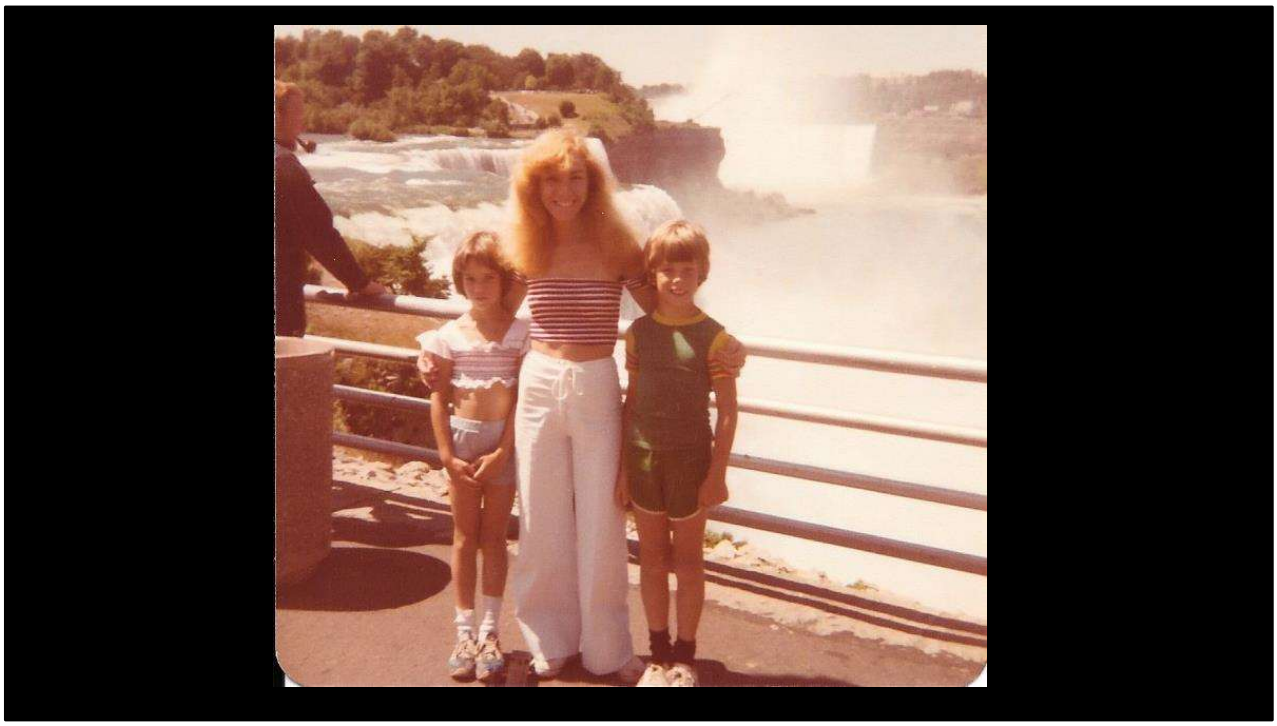
Smart Digital Therapeutics for SUD: Promise, Progress, and Pitfalls

John J. Curtin, PhD
University of Wisconsin - Madison



Addiction Research Center
UNIVERSITY OF WISCONSIN – MADISON

jjcurtin@wisc.edu



When I was still relatively young, my mom lost control of her car when it hit a patch of ice two days before Christmas. Her car spun off the road, her seat collapsed, and she experienced a traumatic brain injury. She survived but she never recovered even to the point of recognizing my sister and me again. She required full-time care for the rest of her life.

My dad is a good man. He devotedly stayed home for 25 years rather than work so that he could care for my mom at home rather than commit her to an institution.

My dad is a good man who he has struggled with his use of alcohol for his entire adult life. At times, it was in the background of our lives. At other times, it was quite severe. He is almost 80 years old now and he has never received any treatment.

It breaks my heart, but it is not surprising. I expect that many of you have similar stories about family and friends who didn't receive the treatment they needed.

We have a mental health crisis in the U.S. and it is a crisis of ****unmet**** high need because our delivery of mental healthcare is deeply flawed.

In 2019, prior to the pandemic, more than half of the 52 million Americans with an active mental illness did not receive ****any**** treatment. ****More than half****!

And for those suffering with a substance use disorder - like my dad - it was worse still. ****9 out of 10 without any treatment****

And these are not just upsetting statistics, they are real people.

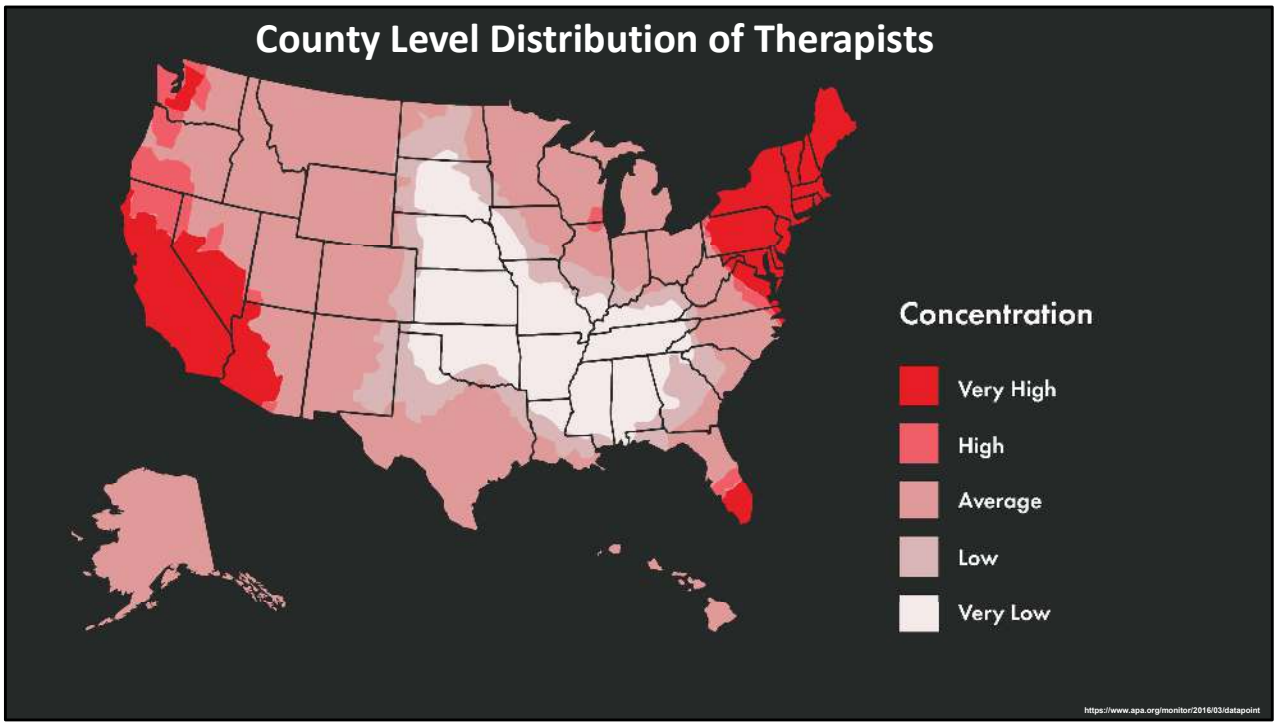
One of them was Victor Kittleson, the **kindhearted and sentimental** brother of one of my graduate students, who died from an opioid overdose here in Minnesota during the pandemic.

This failure to treat is even more troubling for vulnerable groups. Black and LatinX adults receive mental healthcare services at only half the rate of whites.

And similar mental healthcare disparities exist for people living in rural communities and for those with lower incomes.

Access , **acceptability**, and **availability** - these are the factors that undermine our mental healthcare system.

While caring for my mom at home, my dad's **access** to mental healthcare was limited by its high cost without health insurance through a job.



But geography also impacts access.

For example, consider the ****access**** to mental healthcare for the farmer in rural Kansas, when more than 90% of all psychologists and psychiatrists and 80% of social workers work exclusively in metropolitan areas and predominately on the coasts.

But even if my dad had had access to treatment, it likely wouldn't have been ****acceptable**** to him. Like many men of his generation, asking for help from others and sharing personal problems wasn't his strong suit.

But ****even our family**** never discussed it. It was the elephant in the room. Making ourselves vulnerable to therapists and to each other is hard.

And it is harder still because of the stigma that surrounds mental illness even today.

Mental healthcare services are often not ****available**** when we need them most. Many well-regarded therapists have long wait lists that can delay the start of treatment for months.

And once we make it off the wait list, treatment typically involves weekly, monthly, or even less frequent appointments with a therapist. But our mental health needs aren't limited to these pre-scheduled appointments.

Would a therapist have been available to my dad at his moments of greatest need - when he lost a job due to downsizing, or shortly after my mom's accident, or on the many dark mornings when he woke up with his hands shaking and had to decide if he was going to drink again to steady them?

Access, acceptability, and availability - these issues are undermining the treatment capacity of our mental healthcare system and leaving millions without treatment when they need it most.

Fortunately, digital therapeutics, and in particular, digital therapeutics **delivered on smartphones and made smarter still by personal sensing technologies**, are now emerging to target these very same three issues.

But let me pause for a moment. I want to be very clear on one point before we move forward. I do not believe or hope that digital therapeutics will replace human therapists.

Therapists will always be needed for what they do uniquely well. We simply need more than they can provide alone. Digital therapeutics can provide that "more."

So what are digital therapeutics?

Digital therapeutics are software programs or "apps" that are designed to prevent, manage, or treat disease, including mental illness.

Digital therapeutics are delivered to patients on their smartphones and this is the key to their accessibility and availability.

Today, 85% of adults in the U.S. own smartphones. And equally important, ownership is similarly high regardless of race, ethnicity, income, and geography.

Most of us now carry these pocket-sized, powerful computers with us everywhere we go.

And its this widespread use of smartphones that allows digital therapeutics to provide support
* 24 hours a day,
* 7 days a week,
* every day of the year,
* regardless of where we live

Some of the best examples of these digital therapeutics have been developed to target substance use disorders. These apps include multiple supports for patients during their treatment and recovery.

For example, if you need formal treatments, they have you covered. The apps include cognitive behavioral therapy and mindfulness-based relapse prevention

If you need peer support, the apps include discussion forums with other patients.

The apps can also help you locate self-help groups like AA or NA in your community.

The apps can help you track your symptoms over time and your symptoms can even be shared with your therapist if you opt to do so through the apps' clinician dashboard.

And these are just a few examples of the many supports that are possible to provide to patients with digital therapeutic apps.

Of course, all of this would be meaningless if digital therapeutics were not effective.

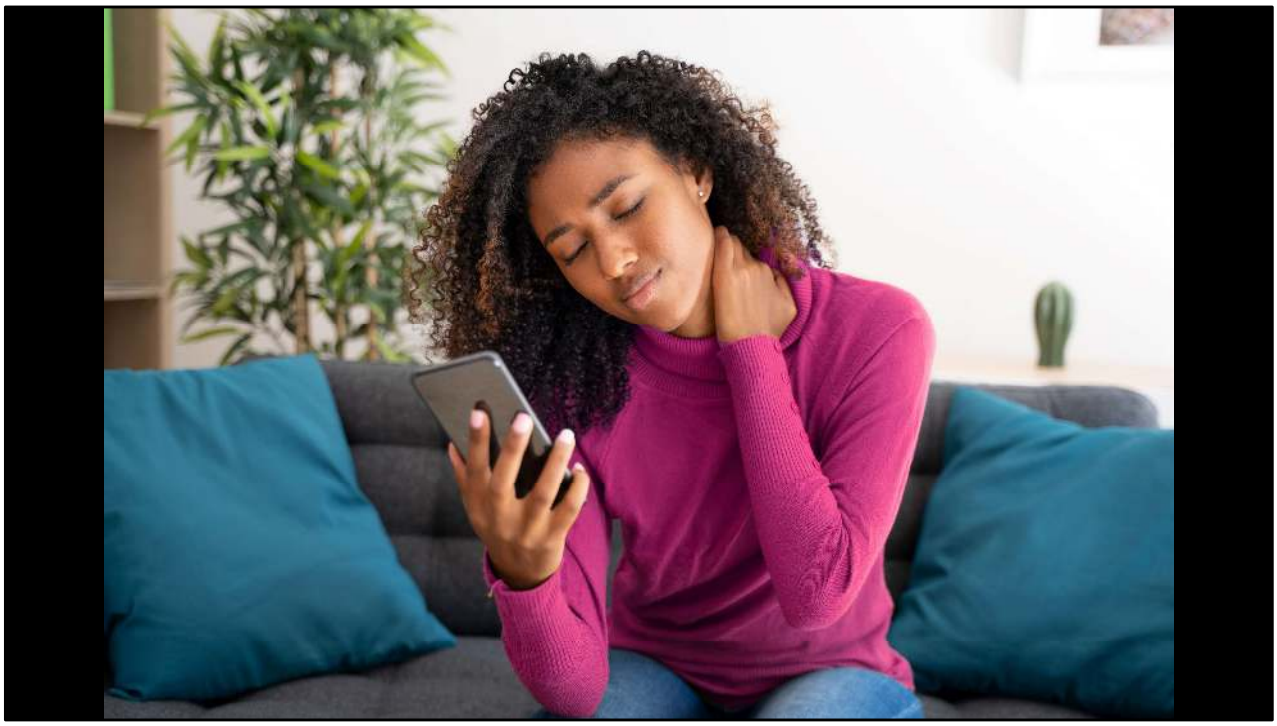
But they are.

For example, patients with substance use disorders who use a digital therapeutic have almost double the odds of being abstinent from alcohol or other drugs.

These increases in abstinence from using digital therapeutics are observed not only when compared to patients on wait lists, who have yet to gain access to treatment but also when digital therapeutics are added on top of traditional treatments for substance use disorders.

And these benefits are durable - they have been documented up to 12 months after the start of treatment.

[PAUSE]



This is a big deal. The magnitude of these benefits are meaningful already, even when we only think about **a single patient** using the app.

But their true power is in their scale, when the **benefits from these apps are multiplied** because they are provided simultaneously to **millions of people in need** and at relatively low cost.

OK, so lets call the apps I have described to you so far, the beta version of digital therapeutics. Their power comes from easy, 24/7 access to their many supports - their treatments, tools, and services.

But this is also their Achilles heel. As the patient using these apps, you now have to tackle difficult questions like:

* When should I use them?
* For how long?
* Which of their many supports are best for me?
* And which are best for me ****right now****, at this moment in time?

"Could you predict not only who might be at greatest risk for relapse …

… but precisely when that relapse might occur …

… and how best to intervene to prevent it?"

My research team became interested in these issues when my colleague Dave Gustafson, the developer of a leading digital therapeutic for substance use disorders, approached us with a simple question. He asked...

> "Could you predict not only who might be at greatest risk for relapse
> but ****precisely when**** that relapse might occur and
> ****how best to intervene**** to prevent it"

Dave had just completed a large study demonstrating the effectiveness of his app. However, he also noticed many of the people who relapsed hadn't used the app in the days leading up to that relapse. And others who had relapsed hadn't used the specific supports in the app that he would have thought would be most effective for them.

He believed that the benefits of his app could be increased if the app knew the person well enough to recognize when they were at greatest risk for relapse and if it was smart enough to recommend the specific supports that would be most effective for them at that moment in time to prevent that relapse.

And I agree with him. The next wave of digital therapeutics, lets call them ****smart digital therapeutics****, must learn to know us better as individuals, not just patients with the same crude diagnosis and same treatment needs at all times.

And these apps will do this through the use of built-in artificial intelligence algorithms that are powered by ****personal sensing****.

Now you may not have heard the term "personal sensing" before, you have almost certainly seen it in action.

I'm a running nut, and for me, ads for trail running shoes, the latest running backpacks, or the newest fancy water bottles follow me around everywhere.

Last week, I installed a new euchre app on my phone so that I could improve my card playing for our next couples night out. As Midwesterners, I expect many of you realize how important this for the coming cold months!
As part of the installation process, the app explicitly asked to track my location and communications so it could serve me up better ads.

Currently, personal sensing uses our personal data **almost exclusively to target ads at us** to sell us things. But we hope to empower people to use personal sensing and their own personal data to improve their mental healthcare instead.

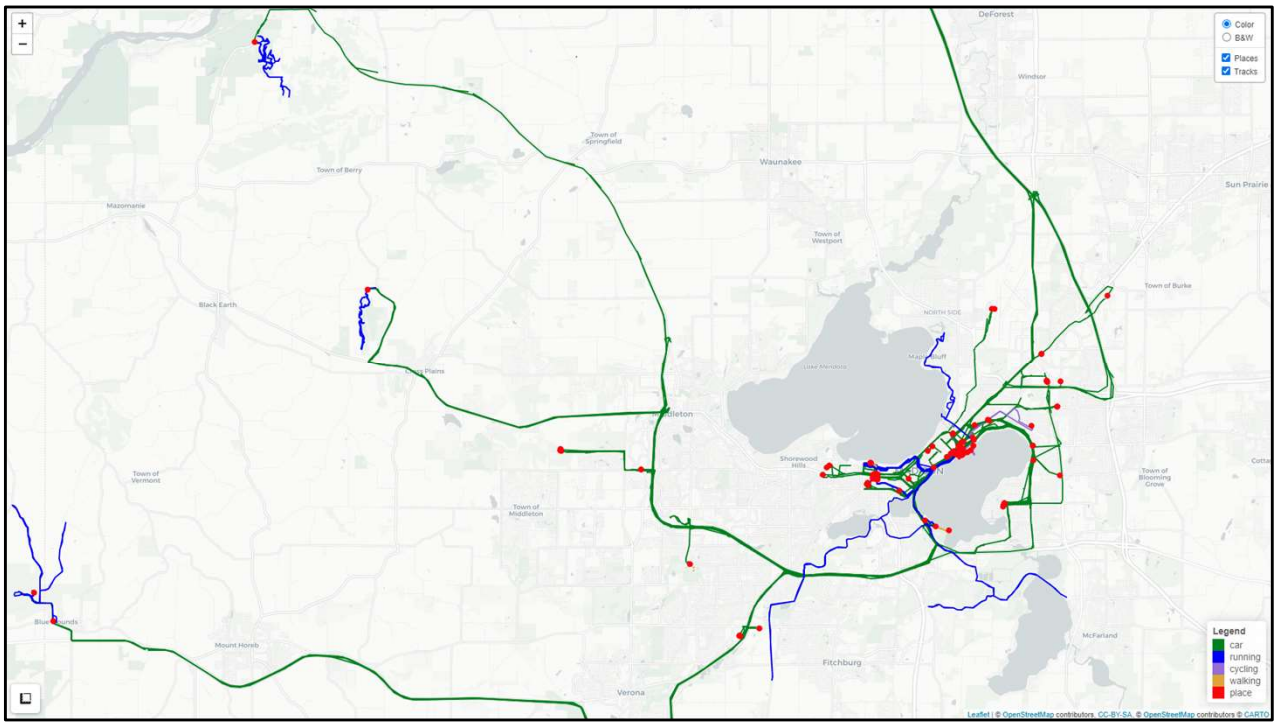
Personal sensing has been supercharged by smartphones.

We use our smartphones to make phone calls and text messages. We often access and post to our social media accounts from our smartphones.

Smartphone-embedded sensors know our moment-by-moment location and activity level. Sensors can even detect other people, or at least their smartphones, in our immediate environment. Our smartphones know when we go to bed and when we rise in the morning.

Personal sensing passively captures all this information and more to understand our recent experiences, preferences, and behaviors. It can be used to predict how we feel right now, and even how we may feel or behave in the future.

Let's take a look at two of the more revealing personal sensing methods that my laboratory is developing to provide you with some intuition about how this works.

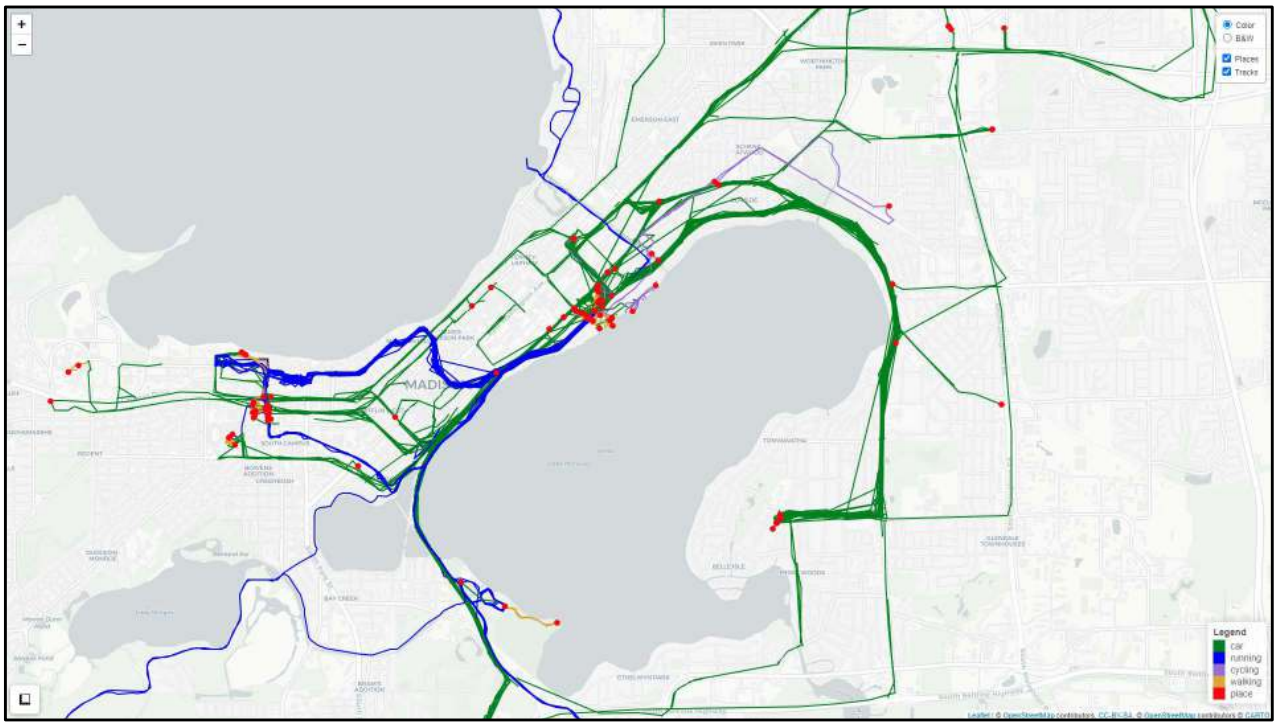


This is a wide view of my moment-by-moment location detected by a digital therapeutic app over a month when we were first experimenting with this sensing method. The app recorded the paths that I traveled, with movement by car in green and running in blue.

The red dots indicate places that I stopped to visit for at least a few minutes.

And although not displayed here, the app knows the days and exact times that I was at each of these locations.

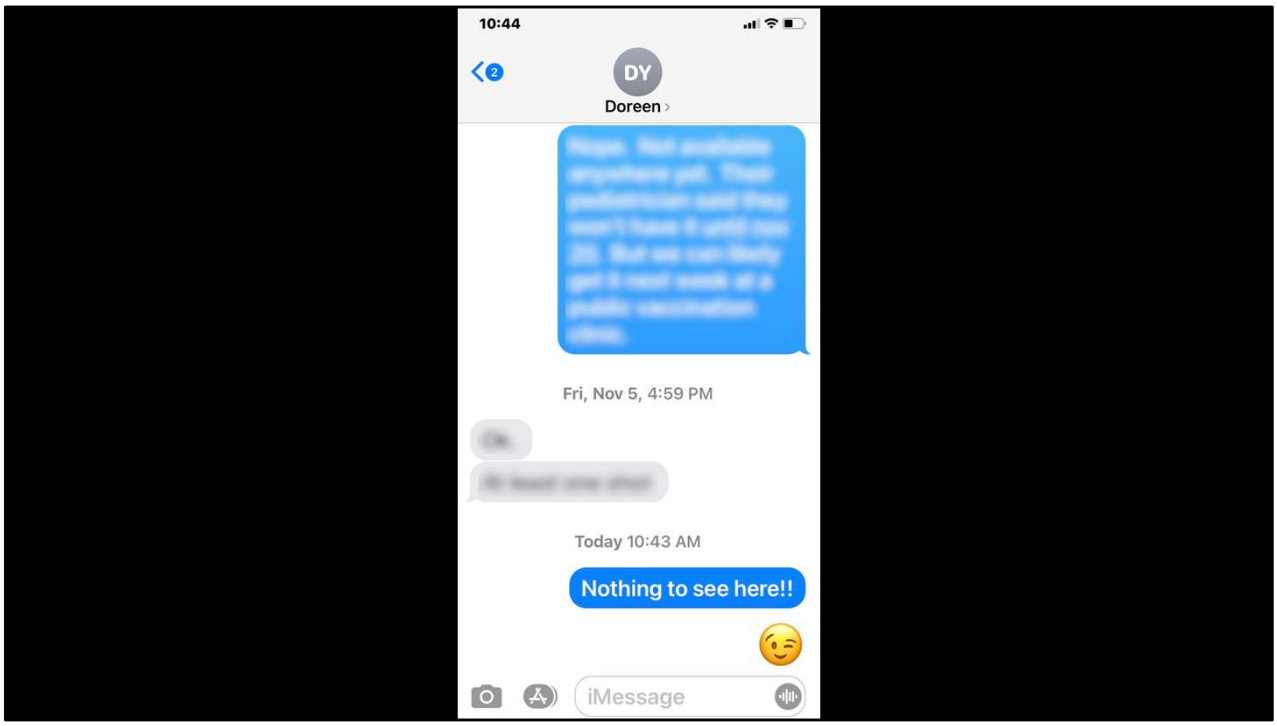
The app can immediately see that I am runner, with long runs leaving from downtown Madison and frequent trail runs on the weekends in the county and state parks to the west and northwest.



Zooming in to the Madison isthmus, the app can see that I drive my children halfway around the lake each morning to their elementary school. And it could detect those stressful mornings when getting my young kids dressed and fed didn't go as planned and we were late, sometimes ****very late****, to school!

The app recorded my daily running commute through downtown Madison to and from my office. From this, it can observe my longs days at the office and also those days that I skipped out.

Looking at the red dots indicating the places I visit, the app can detect the restaurants, bars, and coffee shops where I eat, drink and socialize. It can use public map data to identify these places and make inferences about what I do there.



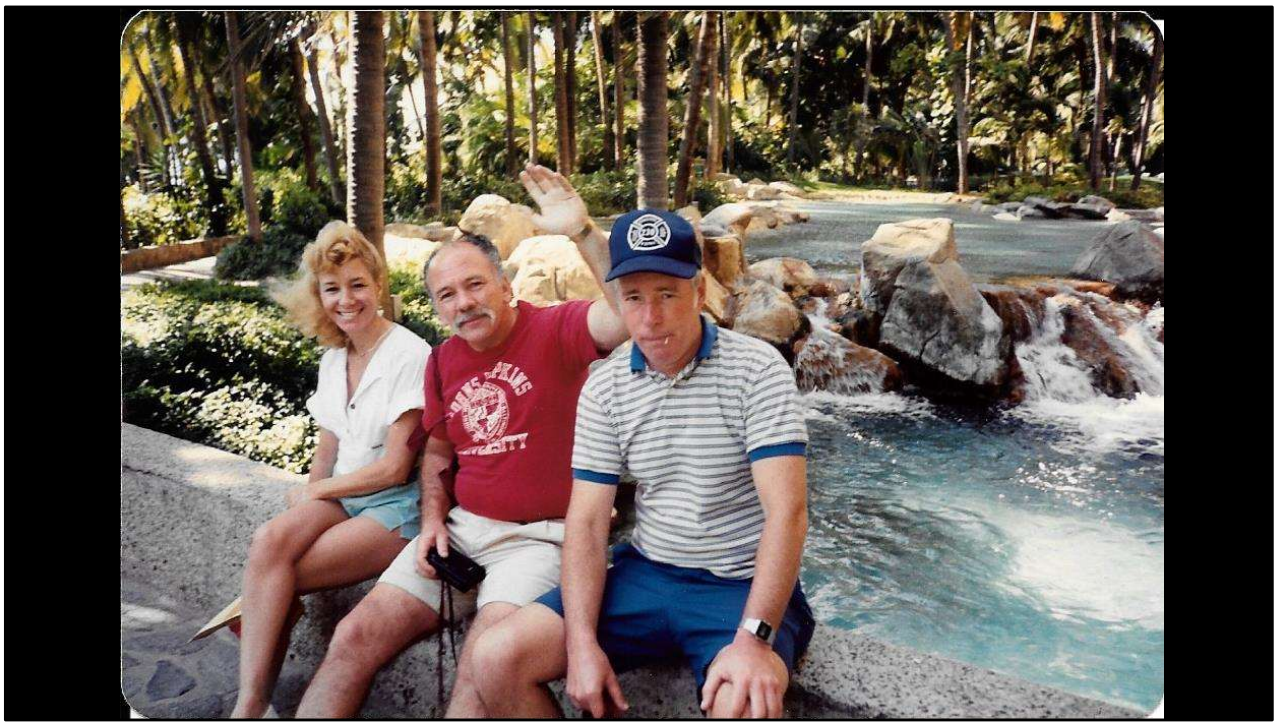
The app also collected my smartphone communications logs and even the content of my text messages.

And no such luck, I don't plan to show you my actual text messages!

But imagine what it learned about me from the patterns of my communications - Who I was calling, when I made those calls, and even the content of what I sent and received by text message.

The app can improve its predictions about us even further by identifying the specific people and places that make us happy or sad or stressed, those that we perceive support our mental health and those who undermine it.

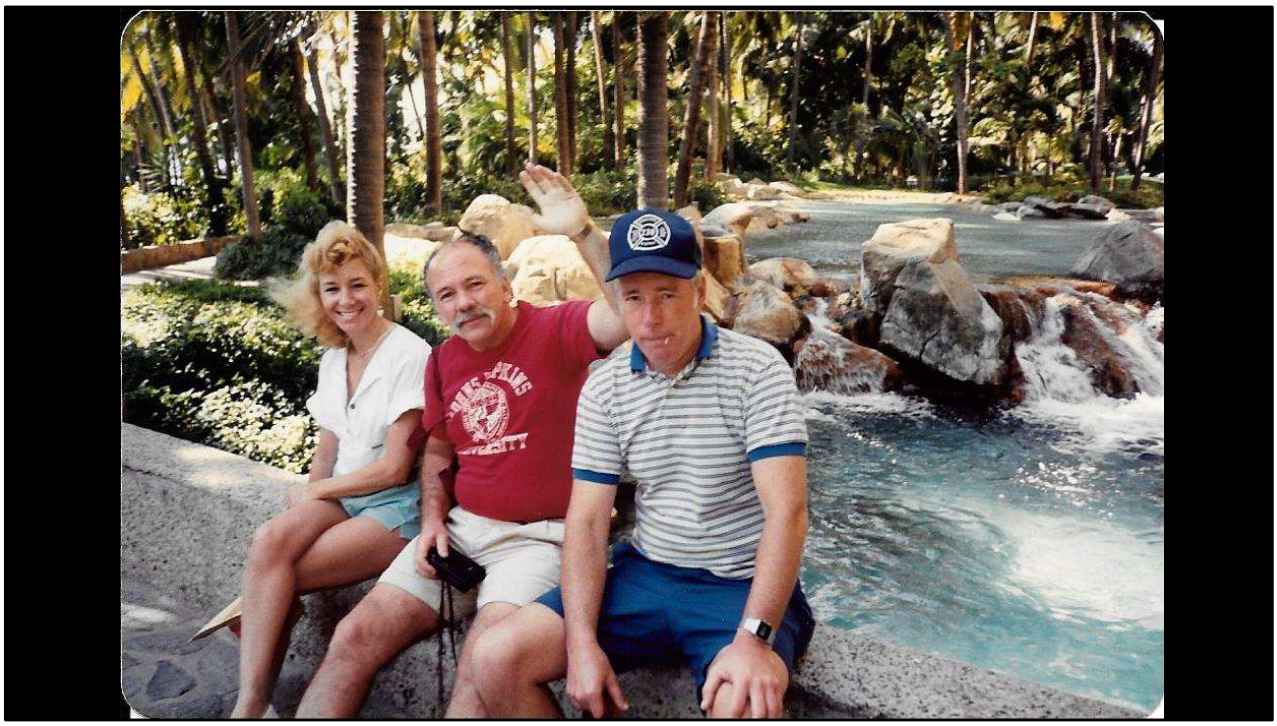
It can gather this information quickly by asking us a few key questions about the people and places it sees us interact with frequently over the first couple of months that we use the app.



For example, if my dad was using this app, it would see that he calls and texts frequently with his close friend, Ed. My dad would report that Ed has been a lifelong source of stability and support.

Given this, the app would know my dad is doing well when he spends time at Ed's house, when they call and text each other to plan activities, when they go for daily walks along the beach by the Long Island Sound.

It could also detect when time spent with Ed abruptly stops each fall because Ed spends his winters in Florida. These months are harder for my dad and he would benefit from more support.



If he was receiving traditional mental healthcare, he might have given permission for the app to share information with his therapist. His therapist might then increase their support of my dad during the months when he is more isolated but direct their support preferentially to other patients when my dad was more stable and supported by his healthy friends and family.

And outside of therapy, the app itself could encourage him to reach out to other supportive family and friends during these months. It could provide him with locations and meeting times for support groups in his community. He could even be assisted to build community in the discussion forums within the app itself.

If the app knew him well, it might even recommend which of these forms of support would be most effective for him.

[PAUSE]

So let me use this framework to now transition to describing how we are taking the first baby steps toward developing **SMART** digital therapeutics for SUD that are powered by sensing and guided by AI or machine learning algorithms to optimize their support of patients during recovery.

I am doing this with at least three goals in mind.
First, I want to give you a sense of how we develop algorithms that use sensing data to support recovery.

Second, I want to give you an early glimpse of what we can already do and what we will eventually be able to do as we get better at this.

And finally, near the end of this presentation, I'd like to shamelessly highlight how we are continuing algorithm development in an ongoing project
for patients with OUD that some of you might be able to help us recruit for if you were interested in partnering with us.

Lapse Prediction in Patients with Alcohol Use Disorder

- ❑ 151 patients with AUD
- ❑ Early in recovery (1-8 weeks)
- ❑ Committed to abstinence throughout study period
- ❑ Followed for up to 3 months
- ❑ Collected active and passive personal sensing data streams

- ❑ **GOAL:** Develop a temporally precise lapse monitoring (prediction) system for patients with AUD


J. Curtin (PI)


D. Gustafson (Co-I)


X. Zhu (Co-I)

NIH National Institute on Alcohol Abuse and Alcoholism

We have recently completed a NIAAA funded project where we collected data from 151 participants who were in early recovery from a moderate to severe alcohol use disorder.

These participants were committed to abstinence at the start of the study and we followed them for up to 3 months, collecting a variety of active and passive personal sensing data streams.

Our first goal with this grant was to develop machine learning algorithms that can generate temporally precise predictions about when future lapses back to alcohol use will occur for patients with AUD.

Personal Sensing Data Streams

- ❑ 4X daily ecological momentary assessments (EMA)
- ❑ Monthly self-report
- ❑ Geolocation (GPS)
- ❑ Cellular communications (voice and text messages)
  - ✓ Meta data
  - ✓ Text message content
- ❑ Sleep sensor (Wake/sleep times; sleep efficiency; wakings; restlessness)

We collected a variety of active and passive personal sensing data streams.

Participants completed brief (7-10 item) ecological momentary assessments or EMAs, 4 times per day

We also have
- more temporally coarse, monthly self reports,
- Moment by moment geolocation,
- Meta data from their cellular communications and the actual content of their text messages,
- and we had sleep sensors in their beds.

We are in the early stages of model building at this point and I will focus today on results from models using only EMA. However, we are actively working with GPS and I'll end with some brief discussion of those preliminary models as well

4x Daily Ecological Momentary Assessments

Current

- ✓ Craving
- ✓ Affect
- ✓ Risky situations
- ✓ Stressful events
- ✓ Pleasant events

Future

- ✓ Risky situations
- ✓ Stressful events
- ✓ Confidence

So let me tell you a bit more about the 4x daily EMA we collected.

On each EMA, participants reported the date and time of any lapses back to alcohol use that they hadn't previously reported.

All of the EMAs also asked them about their current craving, affective valence and arousal, recent risky situations, and recent stressful and pleasant events since their last EMA.

On the first EMA each day, they also reported any future risky situations and stressful events that they expected in the next week and their confidence that they would remain abstinent.

Feature Engineering

- ✓ **Features based on recent past experiences (12, 24, 48, 72, 168 hours)**
- ✓ **Min, max, and median response (all items)**
- ✓ **History (count) of past lapses (item 1) and completed EMAs (compliance)**
- ✓ **Raw scores and change scores (from baseline/all past responses)**

We used these raw EMAs to engineer about 300 features (or predictors) to use in machine learning models to predict future lapses

We formed features by aggregating EMA items over various past time periods ranging from 12 -168 hours in the past

We calculated mins, maxes and medians for the EMA items in these time periods

We also calculated counts of past lapses and counts of past EMAs completed to index compliance

And we included these scores both in raw form and as change from baselines for the participant based on all their previous responses since the start of the study.

Machine Learning Methods

- ❑ **Predict hour-by-hour probability of future lapse**

- ❑ **Lapse window widths**
  - ✓ **1 hour**
  - ✓ **1 day**
  - ✓ **1 week**

For our purposes today I wont dive deep into the machine learning methods but let me highlight a few high level details

We used these features I just described to make predictions about the hour-by-hour probability of a future lapse. We are developing separate models for three future lapse windows – lapses in the next hour, lapses in the next day, and lapses in the next week.

For example, if I was in recovery from an AUD, I could use these models to generate the probability that I would lapse after this presentation starting at 2pm. One model would generate the probability of a lapse between 2 pm and 3 pm today, the second would predict the probability of a lapse between 2 pm today and 2 pm tomorrow and the third would provide the probability of a lapse at any time between 2 pm today and 2 pm next Thursday.

And of course, all of the models would only use data collected prior to 2 pm today so that they are "**predicting**", in the full sense of the word, into the future and not just demonstrating an association.

And each of these models would update their predictions, on a rolling basis, to provide new, updated probabilities for a future lapse every hour after that for also long as I used

the app.

Machine Learning Methods

- ❑ **Statistical Algorithms**
  - ✓ **ElasticNet GLM (e.g., LASSO, ridge regression)**
  - ✓ **Random Forest**
  - ✓ **XGBoost**
  - ✓ **KNN**

- ❑ **Model Tuning and Performance Evaluation**
  - ✓ **Area under ROC curve (AUC) as primary performance metric**
  - ✓ **Sensitivity, Specificity, Balanced accuracy, Positive predictive value**
  - ✓ **Using grouped (by participant) 10-fold CV**

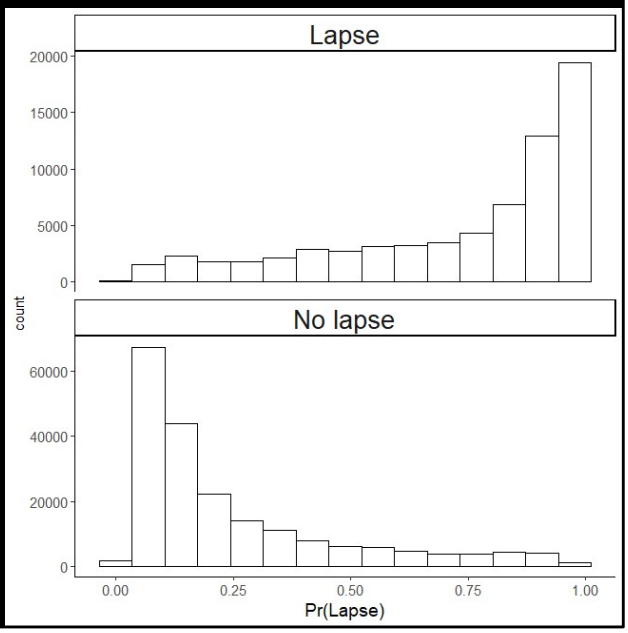
We are evaluating machine learning model configurations that differ by common statistical algorithms.

We are evaluating these models primarily using the area under the ROC curve but we also consider and report a variety of other common metrics.

And, of course, these performance metrics are calculated for **new observations** and **new participants** that the models have never seen and were not trained on by using grouped 10-fold cross-validation.

1 Week: Probabilities for No Lapse and Lapse

- Model predicts probability of lapse in next week for "new" observations in test set
- Can panel predictions for GROUND TRUTH lapse and no lapse observations
- Want high probabilities to be high for true lapses and low for true no lapses



Ok, lets start with the model that provides the coarsest level of temporal specificity – 1 week, and let me take a moment to make the predictions that this machine learning model provides more concrete for you

On the right, you are looking at histograms of the lapse probability predictions that the model makes for all the weeks for all the patients in the test set.

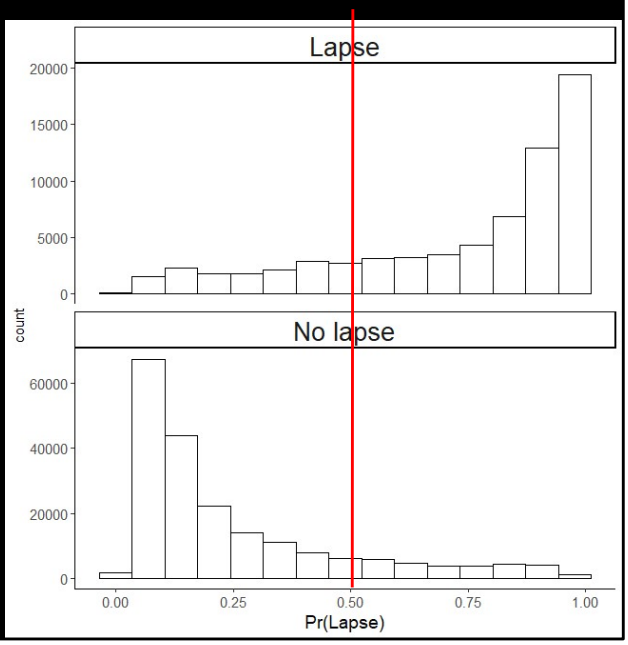
I've paneled these histograms by whether a lapse did or did not happen in reality for each predicted week. The top panel is for weeks that contained lapses and the bottom panel is for weeks with no lapses.

Ideally, you want the predicted probabilities to be very high for weeks where a lapse did indeed occur, and very low for weeks where there was no lapse.

And this is exactly what we see for the **one week** lapse window model

1 Week: Probabilities for No Lapse and Lapse

- Model predicts probability of lapse in next week for "new" observations
- Can panel predictions for GROUND TRUTH lapse and no lapse observations
- Want high probabilities to be high for true lapses and low for true no lapses
- **Need decision threshold for classification (.50 default)**



Now to move from probabilities to actual categorical decisions – in other words, predicting a lapse or no lapse in some specific week, we need a decision threshold. A probability of .50 is often used for this threshold, but as we will discuss later, there are times when we might want to choose other thresholds.

But for now, I will use .5 such that the model will predict "lapse" for all weeks with probabilities > .5 and it will predict "no-lapse" for all weeks with probabilities < .5

Performance Metrics by Lapse Window Width

	1 week	1 day	1 hour
AUC			
Sensitivity	0.79		
Specificity	0.86		
Balanced accuracy	0.82		

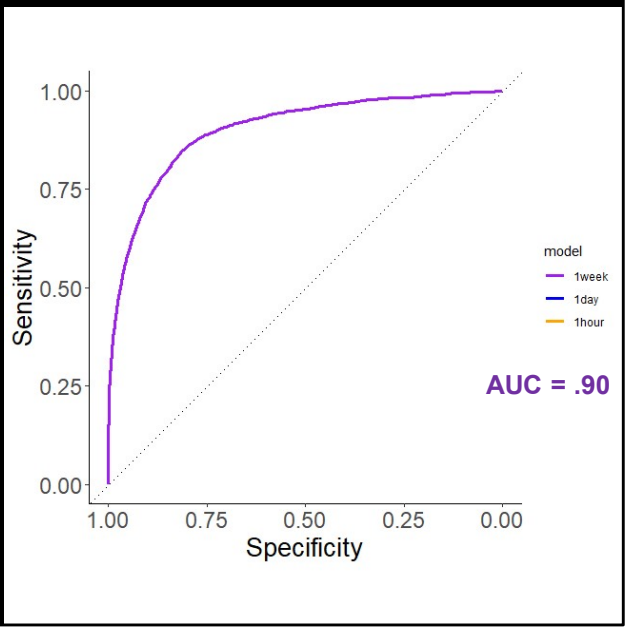
Using this decision threshold, we can now calculate the model's sensitivity, specificity and balanced accuracy which is just the average of these two.

This one week lapse prediction model correctly predicts "lapse" for 79 percent of the weeks that contain a lapse and it correctly predicts "no-lapse" for 86 percent of the weeks that do not include a lapse.

1 Week: ROC Curve

Area under the ROC curve (AUC)

- ✓ **Across all decision thresholds**
- ✓ **~.5 (random) – 1.0 (perfect)**



Coarse rules of thumb for AUC
.70 - .80 are considered fair
.80 - .90 are considered good
≥ .90 are considered excellent

But as I said, depending on the application, we may not want to always use a .5 decision threshold. And this is where the ROC curve and the area under this curve come into play as a performance metric. The ROC curve is a plot of the model's sensitivity by its specificity across all possible decision thresholds.

The area under this curve can range from approximately .5 for a random model (displayed as a dotted line) to 1.0 for a model that predicts perfectly. The better the performance of the model, the more the curve pushes up into the top left corner of the graph that indicates combined high sensitivity and high specificity.

And the AUC for our 1 week model is .90, which is generally considered excellent performance.

1 Week: Features

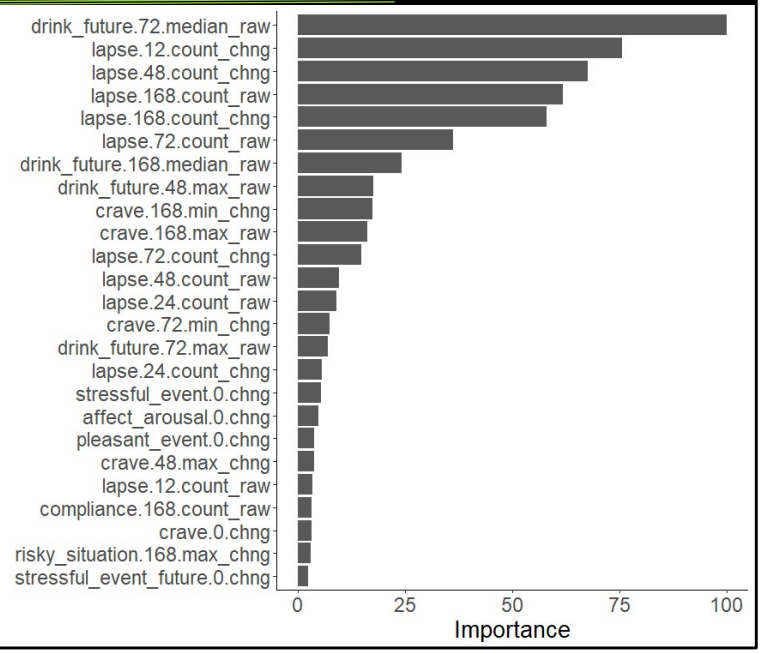
- ❑ **Small number of features are important**

- ❑ **8 EMA items**
- ✓ **Past lapses (item 1)**
- ✓ **Craving (item 2)**
- ✓ **Risky situation (item 3)**
- ✓ **Stressful events (item 4)**
- ✓ **Pleasant events (item 5)**
- ✓ **Affective arousal (item 7)**
- ✓ **Expected future risky situation (item 8)**
- ✓ **expected future drinking (item 10)**

- ✓ **EMA compliance**

- ❑ **Counts, median, mins and maxes are useful**

- ❑ **Both raw score (_raw) and change from baseline (_chng) are useful**



In the spirit of making this model more transparent and interpretable, lets briefly look under the hood at the feature importance for the top 25 features

The plot on the right shows feature names and their associated importance. From this we see a few important characteristics of the model.

First, only a small number of the approximately 300 features contribute meaningfully to predictions. You can see that the importance is already skewing toward 0 even among the first 25 important features

Second, 8 of the 10 original, raw EMA items are represented among these features.

And counts, mins, maxes, and medians, as well as raw and chng scores across all time periods all make contributions.

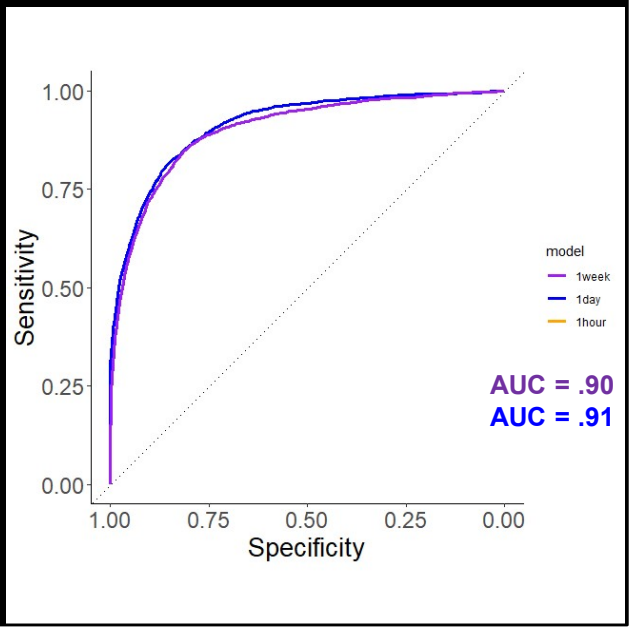
1 Day: ROC Curve

Coarse rules of thumb
.70 - .80 are considered fair
.80 - .90 are considered good
≥ .90 are considered excellent

So we were very encouraged by the performance of this model because it exceeded the performance of the only other published week level lapse prediction model that we were aware of.

But we were also eager to see how well we could do if we required a higher level of temporal precision by developing a day level model.

1 Day: ROC Curve



Coarse rules of thumb
.70 - .80 are considered fair
.80 - .90 are considered good
≥ .90 are considered excellent

And here is the ROC curve and the AUC for the day level model in blue, superimposed on the week level model in purple.

The day level model performed as well as, if not better, than the week level model with an AUC of .91

[PAUSE]

Performance Metrics by Lapse Window Width

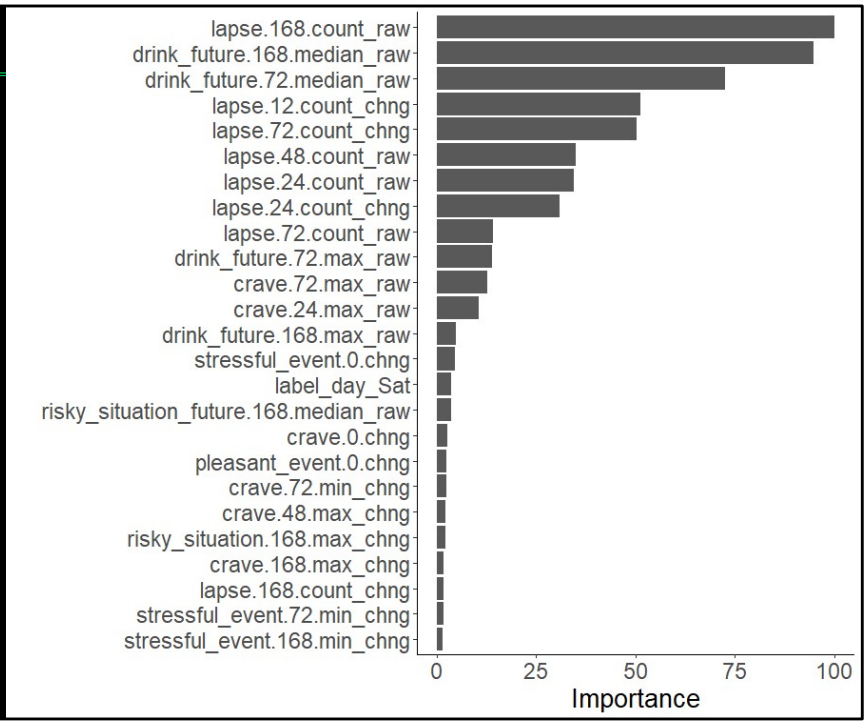
	1 week	1 day	1 hour
AUC	0.90	0.91	
Sensitivity	0.79	0.81	
Specificity	0.86	0.86	
Balanced accuracy	0.82	0.83	

And consistent with this higher AUC, the day level model also had slightly better sensitivity and balanced accuracy, along with comparable specificity to the week level model

[PAUSE]

1 Day: Features

- ❑ **Similar (fewer unique?) EMA items**
- ❑ **Day emerges as new feature (Saturday)**



We can look at the important features for this model too to get a sense of how it works.

The important EMA items are similar to the week model although this day level model seems to depend more on fewer unique EMA items.

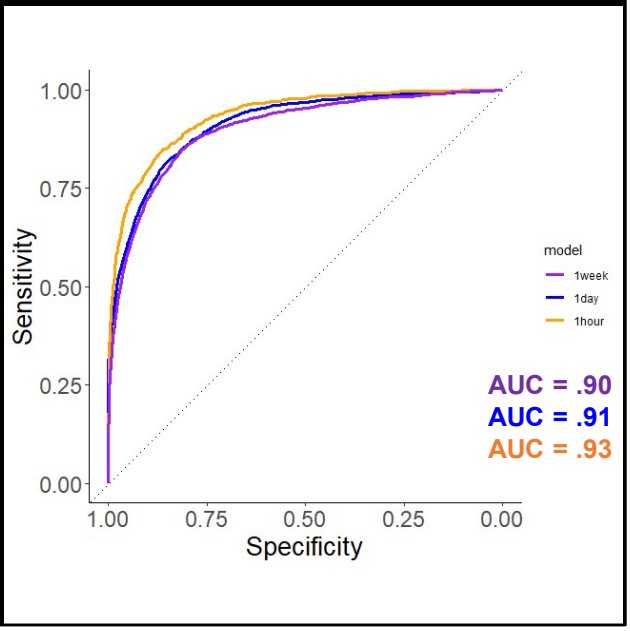
And the **day of the week** for the lapse window emerges as a useful predictor. Not surprisingly, Saturdays have higher probability of lapse than other days of the week and the model can take advantage of this.

1 Hour: ROC Curve

Coarse rules of thumb
.70 - .80 are considered fair
.80 - .90 are considered good
≥ .90 are considered excellent

Given our success with this day level model, we next developed a model with the highest level of temporal precision we could build given how we measured lapses, which was the hour level model.

1 Hour: ROC Curve



Coarse rules of thumb
.70 - .80 are considered fair
.80 - .90 are considered good
≥ .90 are considered excellent

And it turns out that we can do somewhat better still with hour level predictions. Here the orange curve represents the ROC curve for the hour level model, which had the best AUC yet, .93

[PAUSE]

Performance Metrics by Lapse Window Width

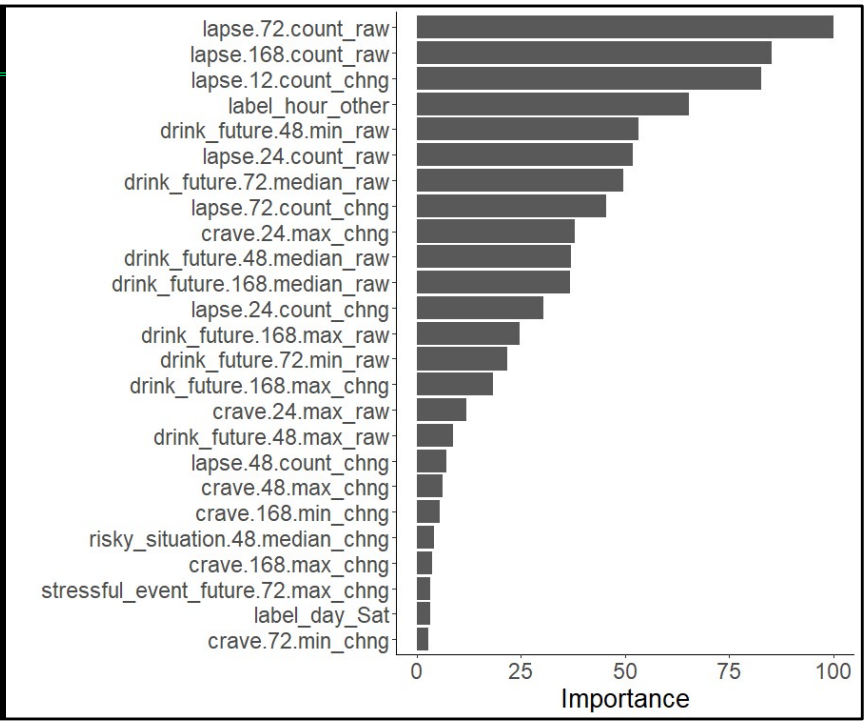
	1 week	1 day	1 hour
AUC	0.90	0.91	0.93
Sensitivity	0.79	0.81	0.84
Specificity	0.86	0.86	0.87
Balanced accuracy	0.82	0.83	0.86

And again, consistent with that AUC, the sensitivity, specificity, and balanced accuracy were higher still for this model.

[PAUSE]

1 Hour: Features

- ❑ **Past reports from previous 0 – 168 hours are still useful**
- ❑ **Still fewer unique EMA items**
- ❑ **Day still matters**
- ❑ **Hour matters (evening/5pm – midnight vs. other)**



Looking at the important features, we see this model uses fewer still unique EMA items (only 5 of the 10 items)

The day associated with the lapse window still matters - hours on Saturday are still higher risk

But now we can also use the hour of the lapse window as a feature too. And again not surprisingly, lapses are higher probability between 5 pm and midnight than at other times of the day across all days

Positive Predictive Value (PPV)

	1 week	1 day	1 hour
AUC	0.90	0.91	0.93
Sensitivity	0.79	0.81	0.84
Specificity	0.86	0.86	0.87
Balanced accuracy	0.82	0.83	0.86
PPV			

Lets return one more time to the performance of these models because I would be remiss if I didn't complicate the story a bit before we close.

Clearly, the sensitivity, specificity, and balanced accuracy of all three of these models is encouragingly high

But what is often missed when we evaluate the performance of machine learning models is their positive predictive value or PPV.

PPV looks at the percentage of positive lapse predictions from the model that are actually true lapses. And PPV is often lower when the positive event, in our case, lapses, is infrequent.

Positive Predictive Value (PPV)

	1 week	1 day	1 hour
AUC	0.90	0.91	0.93
Sensitivity	0.79	0.81	0.84
Specificity	0.86	0.86	0.87
Balanced accuracy	0.82	0.83	0.86
PPV			
	% Lapse: 25.4%	% Lapse: 7.7%	% Lapse: 0.4%

Now we haven't talked about the frequency of lapses across the three lapse windows, but it should be intuitively clear that the percent of weeks that contain a lapse will be higher than the percent of days that contain a lapse, and that the percent of hours that contain a lapse will be lower still.

And this is what we see in our data.

Positive Predictive Value (PPV)

	1 week	1 day	1 hour
AUC	0.90	0.91	0.93
Sensitivity	0.79	0.81	0.84
Specificity	0.86	0.86	0.87
Balanced accuracy	0.82	0.83	0.86
PPV	0.65	0.32	0.02
	% Lapse: 25.4%	% Lapse: 7.7%	% Lapse: 0.4%

Because of this, we see relatively low PPV for our models – and particularly the day and the hour level models.

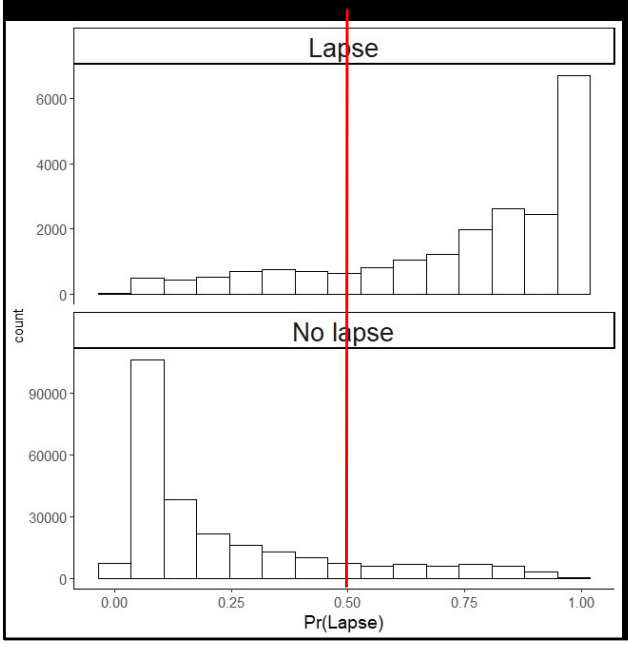
These models catch most lapses (they have hi sensitivity), and they correctly label most no-lapse observations as well (hi specificity). But when you consider their positive predictions, a high percentage of these lapse predictions are false alarms.

This has two big implications.

First, we need to consider the impact of false alarms when using these predictions. It may **not** be problematic to encourage the patient to use their digital therapeutic in cases where the model makes a positive lapse prediction because the cost of a false alarm in that instance is low. More DTx use is likely always good.

But we may **not want to encourage** the use of more costly or burdensome treatments if PPV is low.

Impact of Decision Thresholds on Performance: 1 day

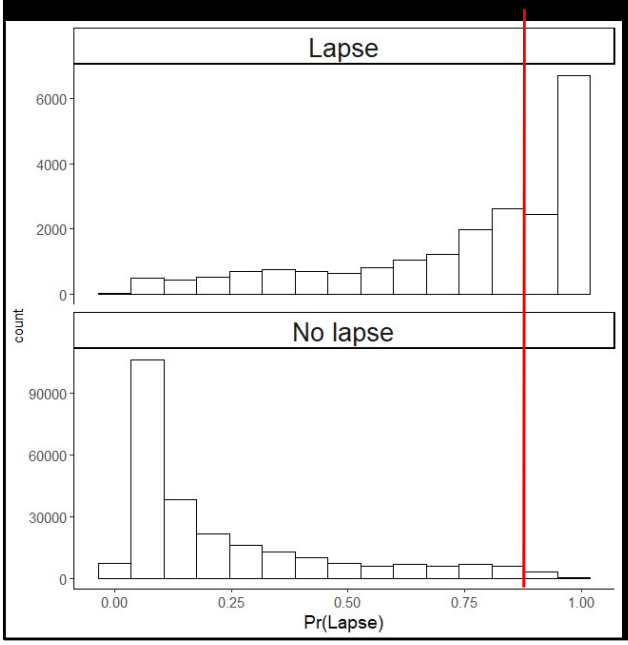


	Threshold = .50	
Sensitivity	**0.81**	
Specificity	**0.86**	
PPV	**0.32**	

Second, if we need higher PPV, we may be able to get it by using a decision threshold that is higher than .5

Here I am showing you the performance of the day level model when we use the default .5 threshold that you saw previously

Impact of Thresholds on Performance: 1 day



	Threshold = .50	Threshold = .90
Sensitivity	**0.81**	**0.40**
Specificity	**0.86**	**0.99**
PPV	**0.32**	**0.83**

But if we move the threshold up to .9 before we call an observation a lapse, we can increase the PPV from .32 to .83.

And this is what we will need to do if we want to recommend more costly treatments or other actions where false alarms may be problematic.

Key Take Home Messages

- ❑ **Relatively high combined sensitivity and specificity**
- ❑ **Comparable performance (AUC) from 1 week down to 1 hour windows**
- ❑ **Will need to adjust decision thresholds to fit how we use the algorithm.**
  - ✓ **Lower PPV OK for low burden or low cost recommendations**
  - ✓ **Higher PPV needed to recommend "costly" interventions or actions**

So to recap our main take-aways

- We were generally encouraged by these preliminary models.

- We've demonstrated that we can get relatively high combined sensitivity and specificity

- We can do this for not only coarse one week windows but also temporally precise windows down to one hour

- But we also recognize that depending on what we will use the predictions for, we may need to adjust decision thresholds to trade off sensitivity for greater positive predictive value.

(Selective) Next Steps

- **GPS and other passively sensed signals**

One of the **key next steps** is to develop models that **rely more on passive sensing** rather than EMA to lower the patient burden of using these systems long term. To this end, we have started to build preliminary models using GPS.

There is clearly predictive signal in the GPS but likely not enough to stand alone as the only features. We are seeing AUCs in the low .7s when we use only GPS to predict future lapses.

However, when we add GPS to the EMA models that I described today, we appear to need fewer unique EMA items to get the same level of model performance. So the addition of GPS may serve to lower patient burden while maintaining model performance.

(Selective) Next Steps

- ❑ GPS and other passively sensed signals
- ❑ **Build models with lead times > 0 hours**

We also need to build models with lead times greater than 0 hours. Remember that for the models we are discussing now, we use all available data up to this moment to predict a future lapse in a window starting right at this moment and into the future. The 0 lead time models, when combined with precise narrow windows, like the one hour window are ideal for recommending just in time interventions within the app. For example, if the model detects you have a high probability of lapsing in the next hour, it might immediately guide you through some urge surfing or recommend that you distract yourself with some engaging games on your phone.

But for other interventions, you might need more lead time. For example, if you needed to schedule time with a therapist or a sponsor or a supportive friend, or you needed to improve the balance of rewarding vs. stressful activities in your life, you need some advance warning to do this to prevent future lapses. So we are working on developing models that will predict future lapses not starting today at 2pm, for example, but starting next Thursday at 2 pm to that you have more time to take preventive steps.

(Selective) Next Steps

- ❑ GPS and other passively sensed signals
- ❑ Build models with lead times > 0 hours
- ❑ **More diversity in training data …**

We are excited by the early performance of these machine learning models BUT I want to be clear that these are preliminary research studies. And they included mostly white participants from our local community in Madison, WI.

Models trained on these participants would be unlikely to work well with black and brown patients or patients from rural communities.

Machine learning models must be trained on diverse samples of patients or their use may exacerbate rather than reduce existing mental healthcare disparities.

We are now collecting data for a NIDA funded project where we are specifically recruiting for racial, ethnic, and geographic diversity across the entire United States.

(Selective) Next Steps

- ❑ GPS and other passively sensed signals
- ❑ Build models with lead times > 0 hours
- ❑ More diversity in training data … **and other SUD outcomes**

In this same project, we have also shifted from a focus on AUD, to instead predicting lapses back to opioid use in patients in recovery from OUD.

This is the NIDA funded project that I mentioned earlier that I wanted to selfishly plug to this audience. We are actively recruiting patients with OUD who have received MOUD for at least a month for this new project.

We are recruiting nationwide and we are looking for treatment sites that want to partner with us. These can be sites in cites like the twin cities but we are also hoping to increase representation of patients in more rural settings as well because we expect that the signals that predict lapse are likely different for people living in cities vs. suburban areas vs. rural areas of the country.

If you are interested in learning more about this, lets talk either during the discussion period or offline later.

(Selective) Next Steps

- ❑ GPS and other passively sensed signals
- ❑ Build models with lead times > 0 hours
- ❑ More diversity in training data … and other SUD outcomes
- ❑ **Use models to improve DTx engagement and clinical outcomes**
  - ✓ **"Clinician-guided" real-time use of DTx based on lapse probability and features**
  - ✓ **How to craft patient feedback to encourage trust in the algorithm**

Of course, accurately predicting future lapses is only useful if these predictions can be used to sustain engagement with treatments and improve clinical outcomes.

We have now started to consider how we can use these predictions to help patients optimize their use of a digital therapeutic.

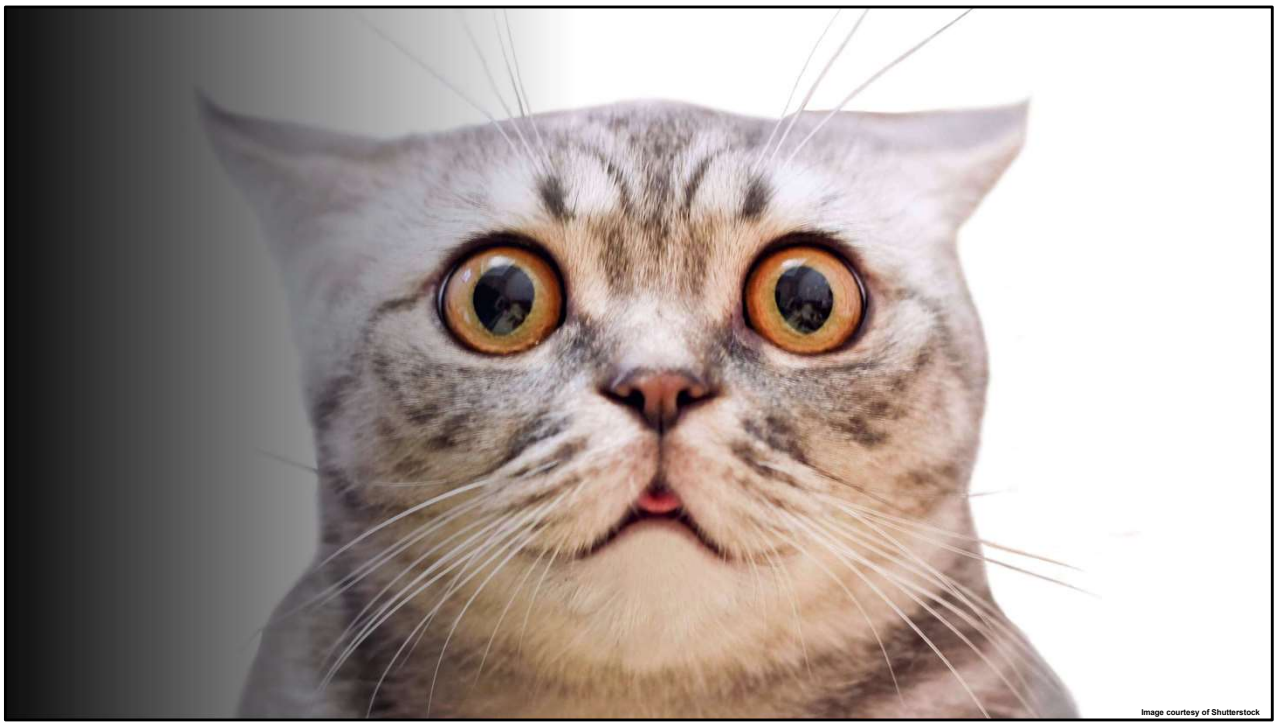
These models may already be able to suggest **when** more use of the DTx is needed because lapse risk is increasing or high but we are also hoping that we can use these models to recommend **which specific tools and supports in the DTx** are best for that patient at that moment in time given their lapse risk probability and the dominant features contributing to the lapse prediction.

We are also sensitive to the fact that these recommendations from the machine learning models will need to be provided to patients in a transparent manner that also encourages them (and their treatment providers) to trust the algorithm and follow its recommendations.

We are preparing a NIAAA grant for February to begin this work and I'd be happy to unpack

and discuss these ideas more during the discussion period.

And finally, I can't end without at least touching on one other elephant in the room. I suspect that at some point during this presentation, perhaps when I was showing you my movement patterns or NOT showing you my text messages, you thought…..



"Holy crap, this is really private information that these apps would collect from me. Who will have access to it and what will they do with it?"

Most of us are too familiar with recent egregious privacy violations including the Facebook Cambridge Analytica scandal and the more recent WhatsApp Pegasus Spyware scandal.

Given this, you might be surprised to hear me say that I am generally optimistic that we will get these privacy issues resolved, at least narrowly in the context of digital therapeutic apps. I'm not making any promises for other apps on your phone or god forbid, Facebook! You're on your own there.....

So here's why I'm optimistic. In the last five years, the FDA has recognized both the potential benefits and risks posed by digital therapeutics.

In response, the FDA has begun to regulate software, including smartphone apps, as it does other medical devices if the purpose of that software is to prevent, manage, or treat disease.

This means that the FDA now evaluates the effectiveness and risks, including privacy risks, of digital therapeutics before clearing them for use with patients.

These FDA policy changes are huge and they begin to situate digital therapeutics squarely within healthcare, where privacy protections have been considered paramount.

Digital therapeutics are here ****today****.

The FDA has already cleared the first two digital therapeutics for substance use disorders.

Our nation's VA Medical Centers have developed digital therapeutics to treat other mental illnesses.

And the VA is providing their apps for free to everyone, not just veterans, through their VA mobile health website. If you or your patients need more care than you are receiving now, download and try these apps.

However, remember that the beta versions of digital therapeutics that are available today are still improving.

But as they get smarter through personal sensing and AI, better mental healthcare is within our reach. Smart digital therapeutics can deliver the right treatments, at the right time, every time, and for all of us.

My dad did not receive the mental healthcare he needed. Neither did Victor Kittleson. With smart digital therapeutics, I hope that we can tell a different story for you, your family and friends, and your patients.

Thank you.

CRediT (Contributor Roles Taxonomy)

This research is highly collaborative and requires many hands.

Kendra Wyant (doctoral student):	data curation, conceptualization, formal analysis, software
Sarah Sant'ana (doctoral student):	data curation, conceptualization
Gaylen Fronk (doctoral student):	conceptualization
John Curtin (faculty):	conceptualization, formal analysis, funding acquisition, methodology, software, supervision
David Gustafson (faculty):	conceptualization
X. Zhu (faculty):	conceptualization, methodology
Susan Wanta (staff)	project administration, software
Candace Lightheart (staff)	data curation, investigation, supervision
Megan Schultz (staff)	data curation, investigation
Jill Nagler (staff)	investigation
Kerry Keiser (staff)	investigation
Chris Gioia (staff)	clinical supervision