Clinical Applications of Human Language Technology: Opportunities and Challenges

Mark Liberman

University of Pennsylvania

http://ling.upenn.edu/~myl

We infer a lot from the way someone talks: personal characteristics like age, gender, background, personality; contextual characteristics like mood and attitude towards the interaction; physiological characteristics like fatigue or intoxication. Many clinical diagnostic categories have symptoms that are manifest in spoken interaction: autism spectrum disorder, neurodegenerative disorders, schizophrenia, and so on.

The development of modern speech and language technology makes it possible to create automated methods for diagnostic screening or monitoring. More important is the fact that these diagnostic categories are phenotypically diverse, representing (sometimes apparently discontinuous) regions of complex multidimensional behavioral spaces. We can hope that automated analysis of large relevant datasets will allow us to do better science, and learn what the true latent dimensions of those behavioral spaces are. And we can hope for convenient, inexpensive, and psychometrically reliable ways to estimate the efficacy of treatments.

I'll present some suggestive preliminary results, and discuss future research opportunities as well as the existing barriers to progress.

Adapted from a presentation at a 2/26/2015 National Research Council workshop:

"Statistical Challenges in Assessing and Fostering the Reproducibility of Scientific Results"

The context:

The past 50 years
have seen enormous quantitative changes
in the efficiency and reproducibility
of speech and language research,
thanks to advances in digital technology.

The near future will bring even larger changes –
not only quantitative changes in productivity and scale ,
but also qualitative changes in the nature of our research,
enabled by new (semi-)automatic methods.

New sources of data
and new methods of automated analysis
are opening up vast new territories of linguistic research.

We can easily acquire and manage new sources of linguistic data
that are several orders of magnitude bigger than old ones.

Because new methods can do old tasks several orders of magnitude more efficiently,
it's increasingly easy to explore these new datasets in old ways.

We can also easily experiment with completely new approaches to analysis and modeling.

And these new methodologies are rapidly spreading
into all the fields that study speech, language, and communicative interaction,
from poetics, sociology, and politics to psychology and neuroscience.

Unfortunately,
biomedical and psychological research practices
are (for the most part) 20-30 years behind the times
in ways that seriously harm research.

So let's take a brief side trip to explain what I mean by that,
based on a presentation to a workshop on
"Statistical Challenges in Assessing and Fostering the Reproducibility of Scientific Results"

Committee on Applied and Theoretical Statistics (CATS),
Board on Mathematical Sciences and their Applications,
National Academy of Sciences

February 26-27, 2015

What is “Human Language Technology”?

The term HLT was coined by DARPA program managers in the 1990s

Inputs might be
- audio or video that includes people communicating
- texts
- complex semi-structured collections of recordings and texts
- data structures representing (perhaps evolving) understanding

Outputs might be
- computer-created audio/video streams
- texts
- data structures representing (perhaps evolving) understanding

Relevant technological capabilities include speech recognition, machine translation, information retrieval and information extraction, summarization, question answering, optical character recognition, speaker identification, language identification, sentiment analysis, etc.

Today, HLT is more and more widely used.

There are three secrets to its success:

1. Cheap fast digital hardware
2. Ubiquitous digital networking
3. A research management technique
   developed in the 1980s
   and applied increasingly widely since then

The first two driving forces are obvious,
but the third involves some semi-obscure intellectual history.

First, a series of HLT failures:

In the 1960s and 1970s, there were many projects focused on machine translation, natural-language interaction with databases, speech recognition, and speech synthesis.

Based on human-coded rules or constraints,
these generally worked to some extent – but they were

- limited in scope,
- brittle under even modest changes in context,
- expensive to create and port,
- and generally disappointing in performance anyhow.

Many influential people concluded
that these endeavors were hopeless.

A leader in pushing this narrative was John Pierce.

Pierce supervised the team that built the first transistor,
and oversaw development
of the first communications satellite.

So his opinion carried considerable weight.


ASSOCIATES

John Pierce, “Whither Speech Recognition?”, JASA 1969:

“… a general phonetic typewriter is simply impossible unless the typewriter has an intelligence and a knowledge of language comparable to those of a native speaker of English.”

“Most recognizers behave, not like scientists, but like mad inventors or untrustworthy engineers. The typical recognizer gets it into his head that he can solve ‘the problem.’ The basis for this is either individual inspiration (the ‘mad inventor’ source of knowledge) or acceptance of untested rules, schemes, or information (the untrustworthy engineer approach). . . .”

“The typical recognizer ... builds or programs an elaborate system that either does very little or flops in an obscure way. A lot of money and time are spent. **No simple, clear, sure knowledge is gained.** The work has been an experience, not an experiment.”

Tell us what you really think, John . . .

"We are safe in asserting that speech recognition is attractive to money. The attraction is perhaps similar to the attraction of schemes for turning water into gasoline, extracting gold from the sea, curing cancer, or going to the moon. One doesn't attract thoughtlessly given dollars by means of schemes for cutting the cost of soap by 10%.
To sell suckers, one uses deceit and offers glamor."

"It is clear that glamor and any deceit in the field of speech recognition blind the takers of funds as much as they blind the givers of funds.
Thus, we may pity workers whom we cannot respect."

But in 1985, DARPA restarted HLT support

Some smart program managers, led by Charles Wayne, had an idea.

They designed a speech recognition research program that

- protects against "glamour and deceit"
  because there is a well-defined, objective evaluation metric
  applied by a neutral agent (NIST)
  on shared data sets; and
- and ensures that "simple, clear, sure knowledge is gained"
  because participants must reveal their methods
  to the sponsor and to one another
  at the time that the evaluation results are presented

Needed: Published data and well-defined metrics

David Pallett, "Performance Assessment of Automatic Speech Recognizers”, *J. of Research of the National Bureau of Standards*, 1985:

Definitive tests to fully characterize automatic speech recognizer or system performance cannot be specified at present. However, it is possible to design and conduct performance assessment tests that make use of widely available speech data bases, use test procedures similar to those used by others, and that are well documented. These tests provide valuable benchmark data and informative, though limited, predictive power. **By contrast, tests that make use of speech data bases that are not made available to others and for which the test procedures and results are poorly documented provide little objective information on system performance.**

“Common Task” structure

- A detailed task definition and “evaluation plan”
  developed in consultation with researchers
  and published as the first step in the project.
- Automatic evaluation software
  written and maintained by NIST
  and published at the start of the project.
- **Shared data:**
  Training and “dev(elopment) test” data
  is published at start of project;
  “eval(uation) test” data is withheld
  for periodic public evaluations

Not everyone liked it

Many Piercians were skeptical:
"You can't turn water into gasoline,
no matter what you measure."

Many researchers were disgruntled:
"It's like being in first grade again --
you're told exactly what to do,
and then you're tested over and over ."

But it worked.

Why did it work?

1. The obvious: it allowed funding to start
   (because the projects were glamour-and-deceit-proof)

   and to continue
   (because funders could measure progress over time)

Why did it work?

2. Less obvious: it allowed project-internal hill climbing
   - because the evaluation metrics were automatic
   - and the evaluation code was public

   This obvious way of working was a new idea to many!
   … and researchers who had objected to be tested twice a year began testing themselves every hour…

Why did it work?

3. Even less obvious: it created a culture
   (because researchers shared methods and results
   on shared data with a common metric)

Participation in this culture became so valuable
that many research groups joined without funding

What else it did

The *common task method* created a positive feedback loop.

When everyone's program has to interpret the same ambiguous evidence,
ambiguity resolution becomes a sort of gambling game,
which rewards the use of statistical methods,
and led to the flowering of "machine learning".

Given the nature of speech and language,
statistical methods need the largest possible training set,
which reinforces the value of shared data.

Iterated train-and-test cycles on this gambling game are addictive;
they create "simple, clear, sure knowledge",
which motivates participation in the common-task culture.

The “Common Task Method”

. . . has become the standard research paradigm
in experimental computational science:

- Published training and testing data
- Well-defined evaluation metrics
- Techniques to avoid over-fitting
  (managerial as well as statistical)

Domain: ***Algorithmic analysis of the natural world.***

Over the past 35 years, variants of this method
have been applied to many other problems:

machine translation, speaker identification, language identification, parsing, sense disambiguation, information retrieval, information extraction, summarization, question answering, OCR, sentiment analysis, image analysis, video analysis, … , etc.

The general experience:

1. Error rates decline by a fixed percentage each year,
   to an asymptote depending on task and data quality
2. Progress usually comes from many small improvements;
   improvement by 1% is a reason to break out the champagne.
3. Shared data plays a crucial role – and is re-used in unexpected ways.
4. Glamour and deceit have mostly been avoided.

Versions of the “common task” model are now routinely used,
outside of any sponsored projects,
by both academic and industrial researchers –
and similar “challenges” are routinely created
by technical societies and ad hoc groups.

But there are several important areas that are lagging far behind,
both in technology and in methodology:

- Clinical applications
- Educational applications
- Legal applications

A cultural change in those fields is long overdue!

A trivial example of where we are (or could be):

In June 2014, I participated in a workshop discussion of *tonogenesis*
(A historical change in Chinese, Vietnamese, Thai etc.
where consonant manner distinctions turn into tone differences)

The anatomy, physiology, and physics of voicing distinctions in speech
naturally produce differences in f0.
This has been observed in isolated examples,
but there seemed to be no systematic study in the literature.

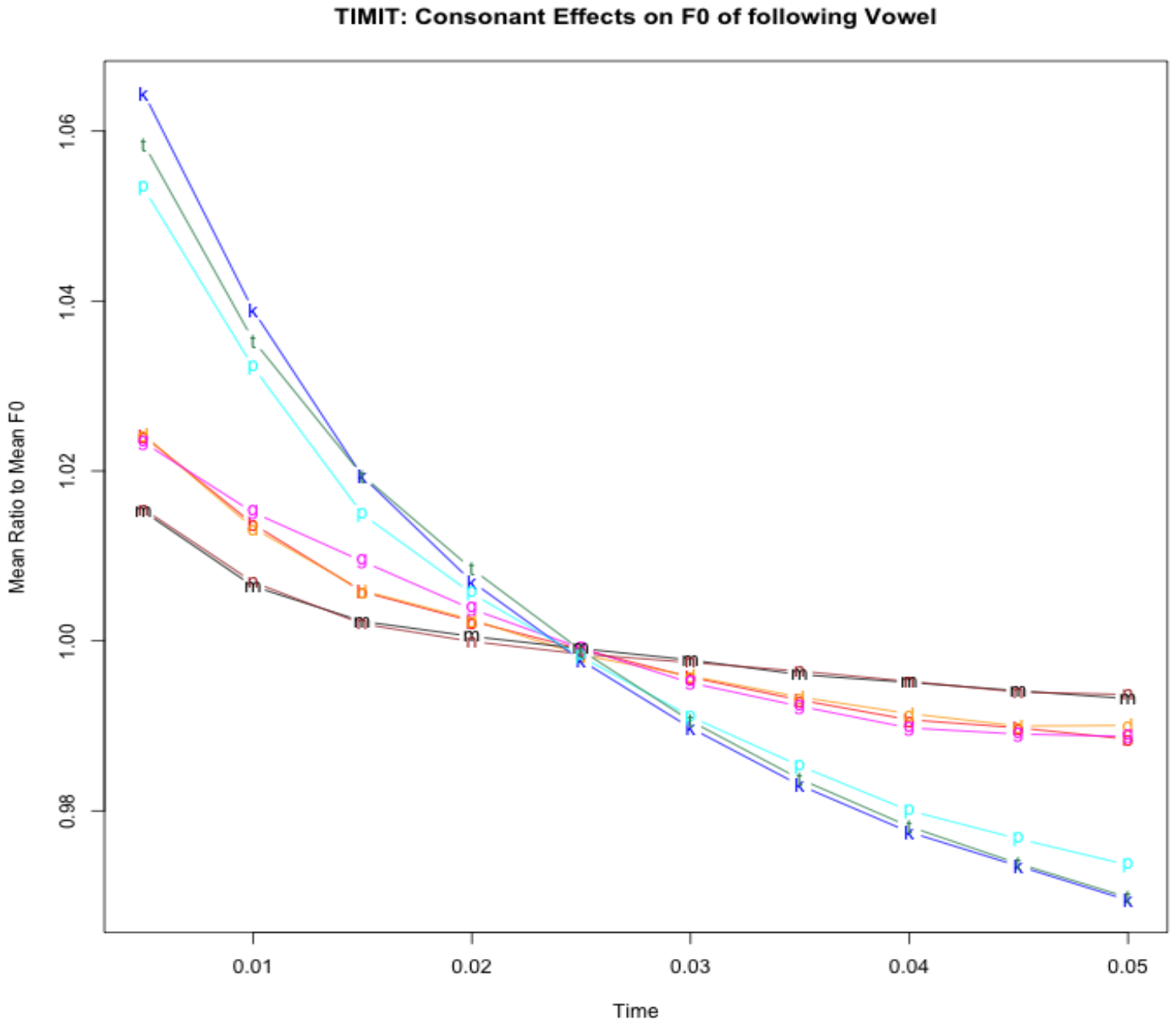
So over breakfast the next morning,
I checked out syllable-initial consonants in the TIMIT corpus.

Method – write a script to

1. Pitch track all 6300 TIMIT sentences, creating f0 estimates every 5 msec
2. Select all syllables beginning with p,t,k b,d,g or m,n
3. Pull out the first 50 msec of voiced speech per syllable
4. Reject all sequences containing discontinuities
5. Normalize the f0 vectors as the ratio relative to each vector's mean

$$y = x/mean(x)$$

6. Plot the mean of all normalized vectors for each initial consonant


TIMIT: Consonant Effects on F0 of following Vowel
Mean Ratio to Mean F0
1.06
1.04
1.02
1.00
0.98
0.01
0.02
0.03
0.04
0.05
Time

In the old days, this would have been several years of work
(which is presumably why no one did it…)

In 2014, I could do it in an hour or so,
while consuming a bowl of cereal and several cups of coffee,
using a laptop computer and a page or so of code.

But major challenges remain.

Many annotation problems remain substantially unsolved, such as diarization of real-world conversations.

And there are kinds of data that are not generally available, or not available at all.

I'll focus an important area of inadequate data:

Recordings of clinical interviews,
neuropsychological tests,
and similar things.

There are policies, laws, and ethical concerns
that require such recordings to be treated in a special way,
and are widely (but falsely) believed
to make cross-site sharing impossible.

Why do we want such recordings for research,
and why do we want to share them?

Because speech and language are provide key behavioral markers,
cheaper and less invasive
than brain imaging, blood tests, or genomic tests,
but also often diagnostically more useful.

And more important, many (most?) relevant problems
are “phenotypically diverse”, in ways that matter –
meaning that we really don’t understand them very well.

With enough data and enough research,
we can hope to find the true latent dimensions
of the relevant behavioral space(s).

But a single site rarely has enough data,
and no single research team is likely to find the answers.

We need to pool data across sites,
and we need a community of researchers
working together to understand it.

As exemplified in the the earlier example,
even small and limited datasets
can yield promising results from simple techniques.

This motivates a serious effort
to find ways to share clinical speech and language data
in consistent and accessible ways
on a large scale.

Example: "Autism Spectrum Disorder"

It's clear that Autism is not a "spectrum", i.e. a single dimension,
but rather a space, with many dimensions –

It's a space that we all live in,
with some corners that have been medicalized
because they cause serious life problems.

Is there suitable digital data Out There?

Yes –

for instance, the Autism Diagnostic Observation Schedule (ADOS)

is a standard diagnostic tool,

consisting of a multi-part structured interview

which is video recorded and scored from the video,

with a half a dozen scoring rubrics for of the ~12 segments.

For diagnosis, the multiple scores are added up and thresholded.

O(1,000,000) ADOS recordings are Out There.

An ADOS recording DVD is stored in the patient's folder,
along with many other tests.

We've begun a collaboration
with the Center For Autism Research
at Children's Hospital of Philadelphia,
which has O(3000) such recordings.

We selected an initial set of ~100 interviews,
including interviews with neurotypical controls
and with adolescents with other diagnoses such as ADHD.

We did some preliminary work
to persuade the hospital's Institutional Review Board
that it was both possible and worthwhile
to share 20-minute ADOS audio segments for research purposes

-- with appropriate safeguards.

Analysis of this small pilot corpus (~33 hours)
suggests that every sensible linguistic measurement
shows some interesting signal.

We hope to persuade other clinical centers
to join us in creating a much larger collection.

As Bob Schultz, CAR's director, said:

"With ten thousand interviews,
maybe we could figure out what's really going on."

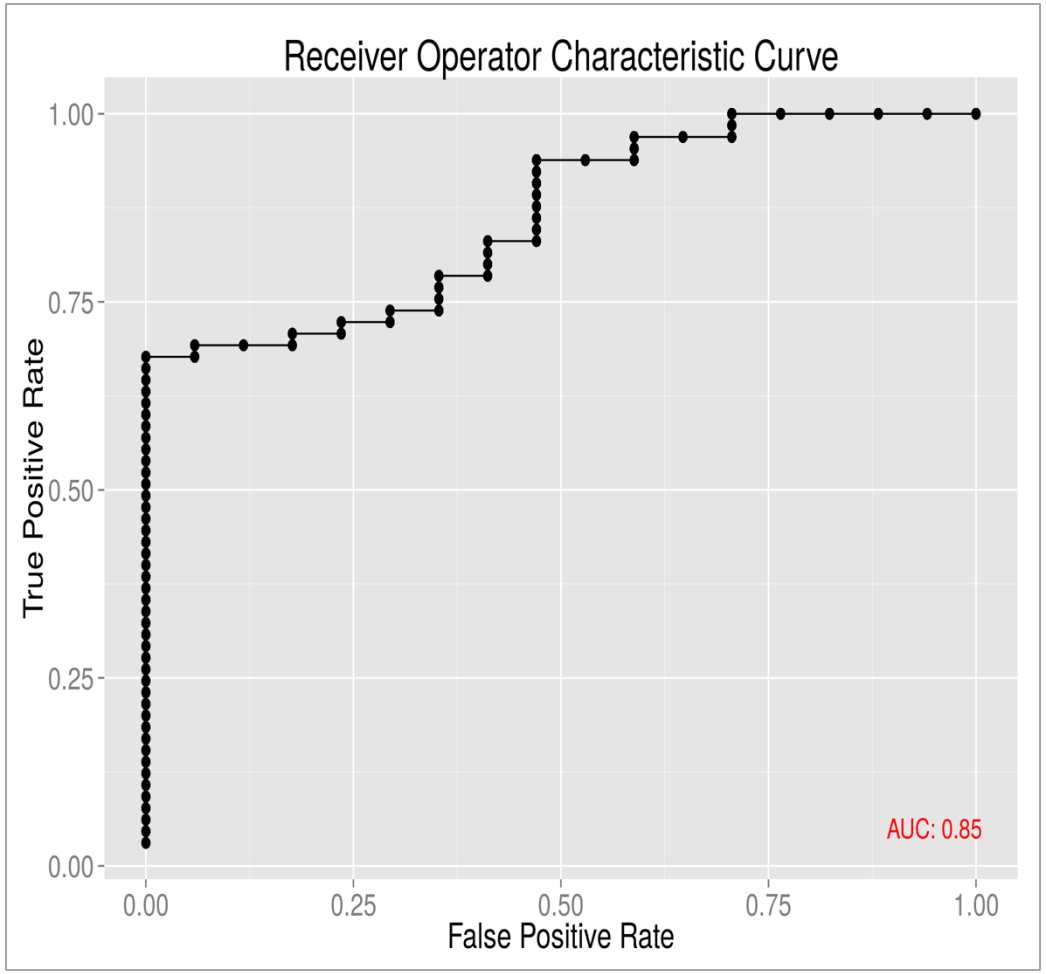
Simple bag-of-words classification – worked better than expected, as usual:

Naïve Bayes, weighted log-odds ratios

Leave-one-out cross validation

Correctly classified
68% of ASD participants
and 100% of typical participants

AUC=85%

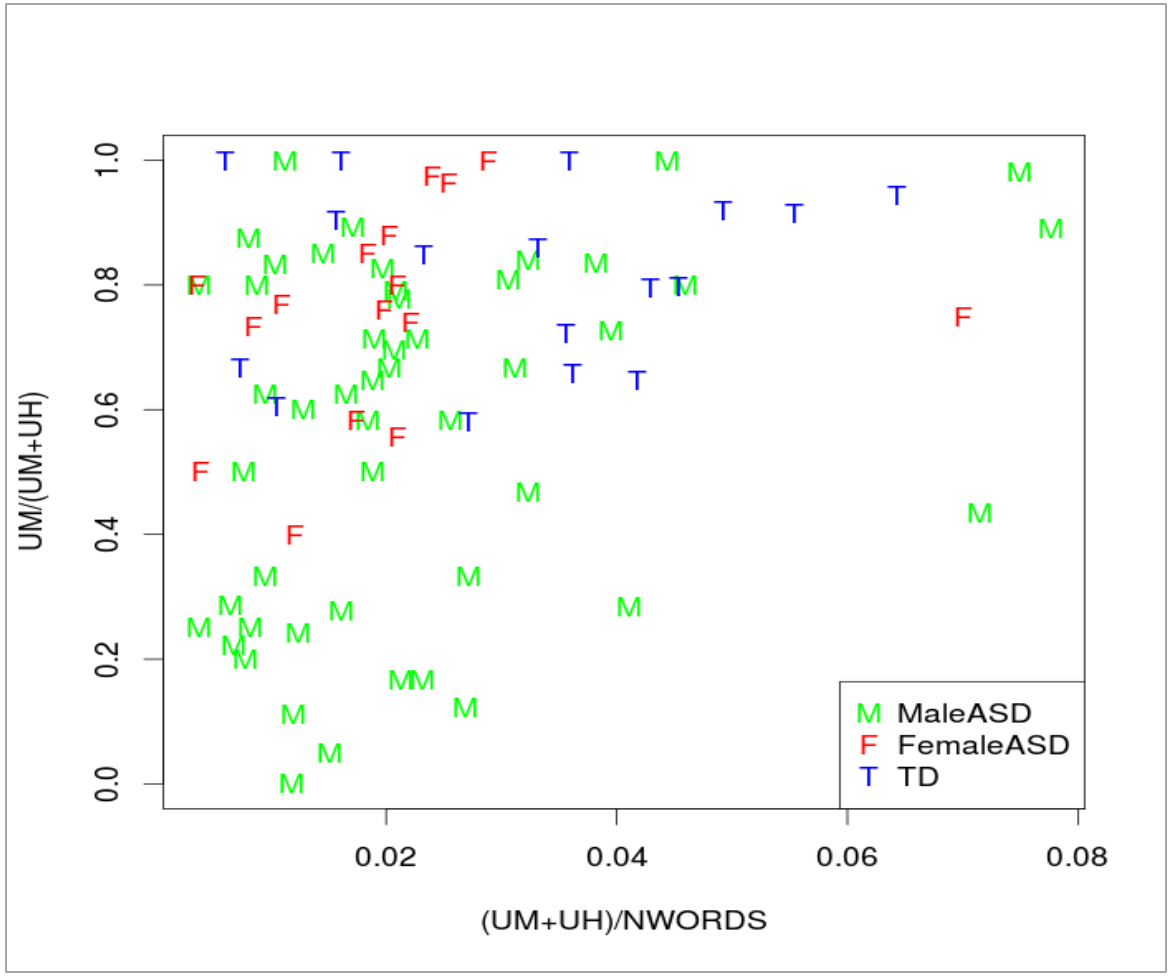


Rates of *UM* production in ASD and TD groups: ***um/(um+uh))***

ASD group: *UM* was 61% of their filled pauses (CI: 54%-68%)

TD group: *UM* was 82% of their filled pauses (CI: 75%-88%)

Minimum value for the TD group was 58.1%, and 23 of 65 participants in the ASD group fell below that value.

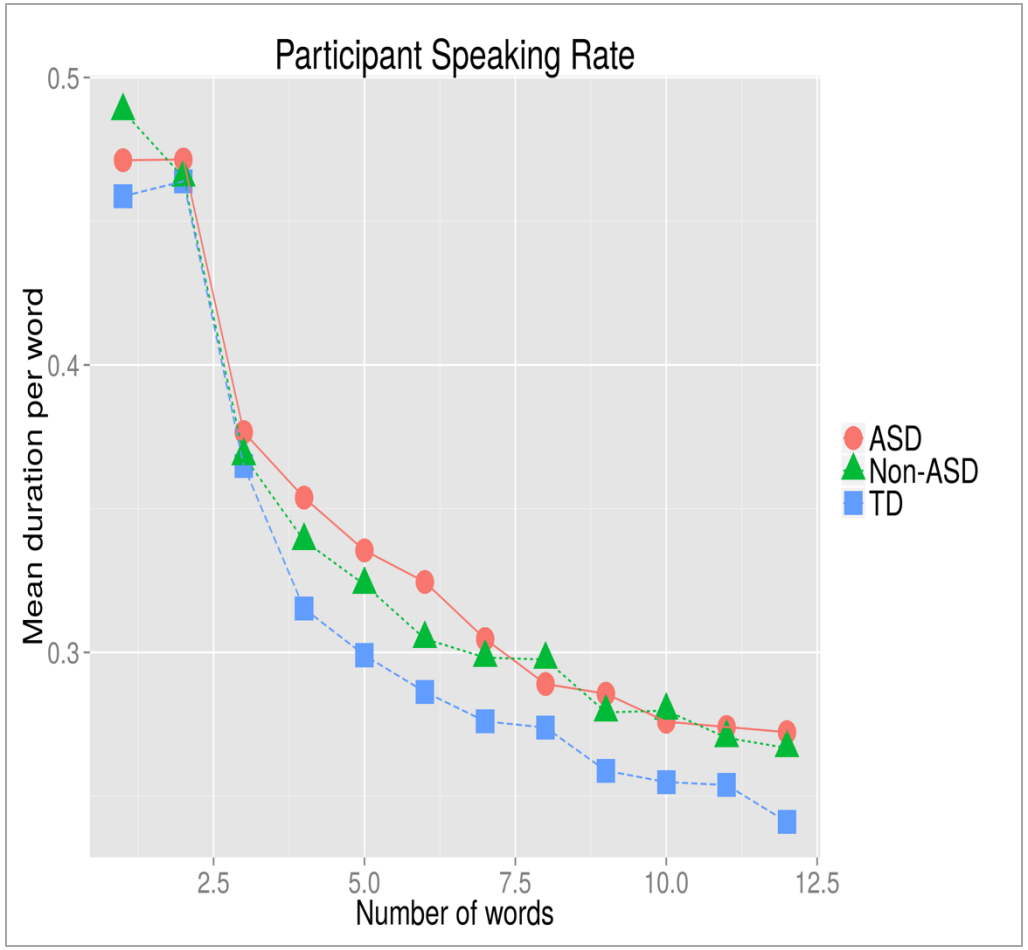


Mean word duration as a function of phrase length:

TD participants spoke the fastest
(overall mean word duration of 376 ms, CI 369-382,
calculated from 6,891 phrases)

Followed by the non-ASD mixed clinical group:
(mean=395 ms; CI 388-401,
calculated from 6,640 phrases)

Followed by the ASD group:
(mean=402 ms; CI: 398-405,
calculated from 24,276 phrases)

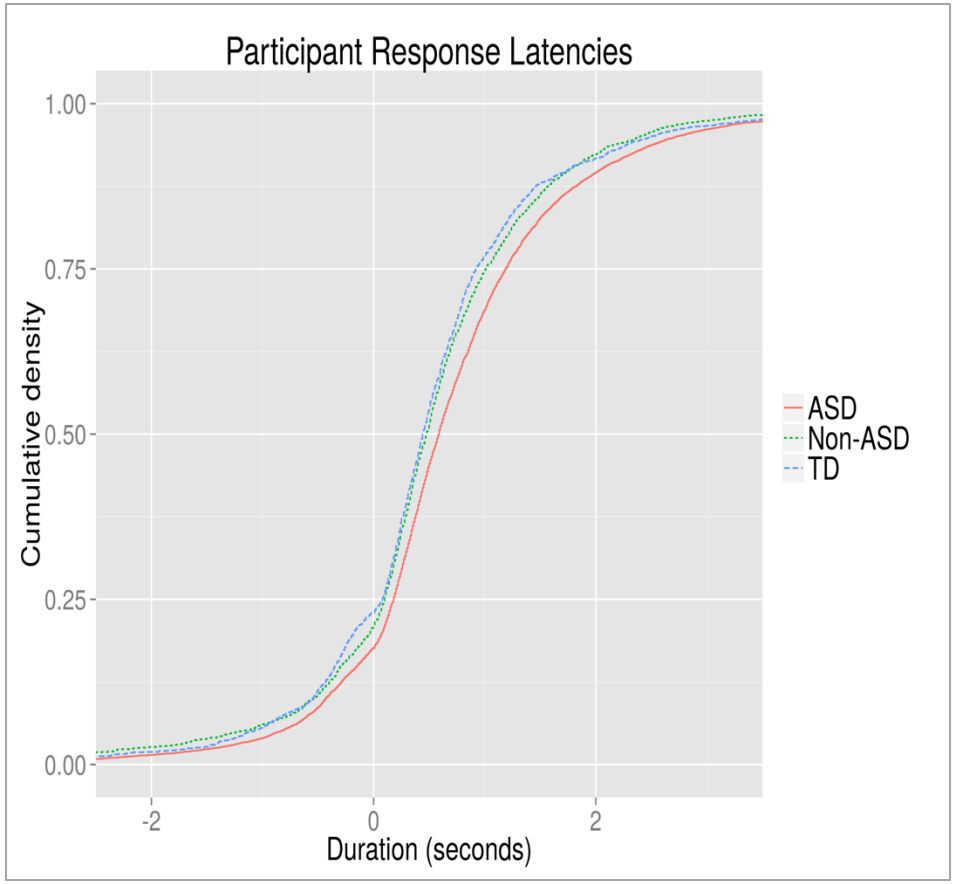


Latency to respond:

Too short = interrupting
speaking over a conversational partner

Too long = awkward silences
interfere with smooth social exchanges

ASD slower than TD

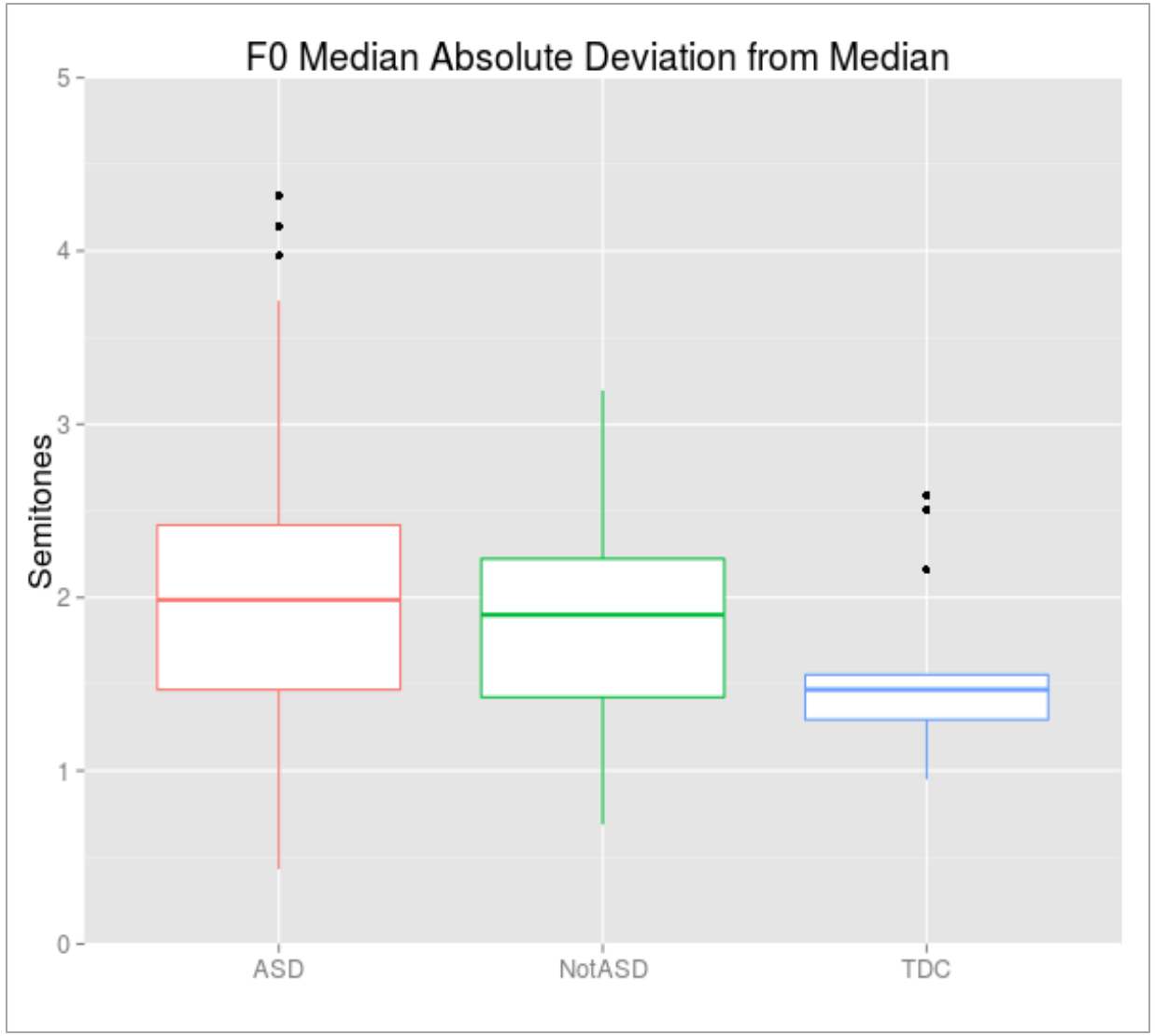


F0 Variation:

Median absolute deviation
from the median (MAD)
Calculated in semitones
relative to speaker's 5th percentile

MAD values are both higher
and more variable
in the ASD
and non-ASD mixed clinical group
compared to the TD group:

ASD: median 1.99 IQR: 0.95
Non-ASD: median 1.95 IQR 0.80
TD: median 1.47 IQR 0.26



. . . and so on . . .

Next steps for ADOS analysis

Expand sample size, enlarge age range, improve specificity

- Multi-site collaboration?
- Downward extension to infancy
- Chart growth to identify points of divergence/targets for intervention

New measures

- New textual and acoustic-phonetic features
- Integration of textual & phonetic features
  - (e.g. dysfluency & pause locations)
- Gesture, gaze, face, posture during conversation
- Other phenotypic data
- Neuroimaging
- Genetics

BUT…

ADOS requires expensive in-person expert collection --

We (also) need scalable automated methods
to collect large and diverse samples.

New ASD Data Collection Initiatives:

Phone bank

- Inexpensive student worker asks ADOS-like questions
- Child and parent language samples, questionnaires, online IQ
- Nationally representative cohort

Computerized Social Affective Language Task (C-SALT)

- Portable self-contained app
- Records language and social affect in schools, clinics, homes
- Controlled recording is conducive to automated approaches (reduces need for transcription)

Simple analysis of audio & transcript from 5-minute unstructured non-clinical conversation:

Table 5: *Classification report of the model.*

Diagnosis	Accuracy	Precision	Recall	F1-score
ASD	0.69	0.80	0.69	0.74
TD	0.83	0.72	0.83	0.77
Average	0.76	0.76	0.76	0.76

Goals and applications:

Support clinical decision-making and improve access

- Low-cost, remote screening
- Direct behavioral observation: record in clinics, integrate into EHR
- Inform identification efforts and assist in differential diagnosis

Identify behavioral markers of underlying (treatable) pathobiology

- Profiles of individual strengths and weaknesses, link to biology
- Personalized treatment planning and improved outcomes
- ****Monitoring and measuring response to interventions****

Give participants and families more information about themselves

- Online feedback
- Monitor growth trajectories

There are many other kinds of datasets
relevant for ASD research –

And many other possible targets for similar research,

for example, the many diverse varieties
of neurodegenerative disorders,
such as Frontotemporal Degeneration,
Parkinsonism, and Alzheimers;

as well as mood disorders, schizophrenia,
and other mental conditions.

Again, in every case that we've looked at,
simple properties of speech and language data
correlate with clinical categories.

We're working with Penn's Frontotemporal Dementia Center
on a dataset of picture-description recordings
from ~1200 patients and elderly controls.

Simple (language-independent) acoustic-phonetic measures
have significant potential value in diagnosis and monitoring.

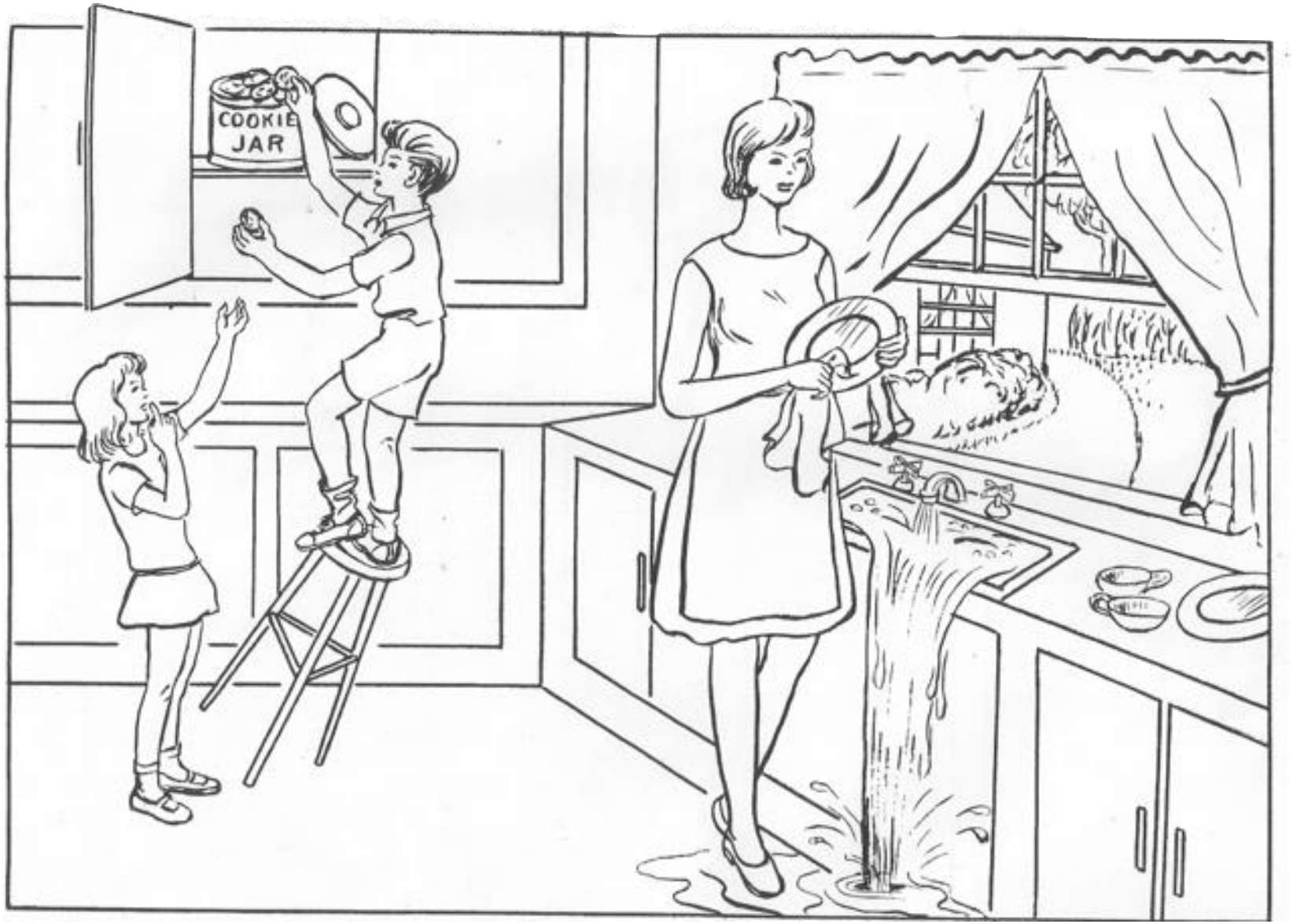
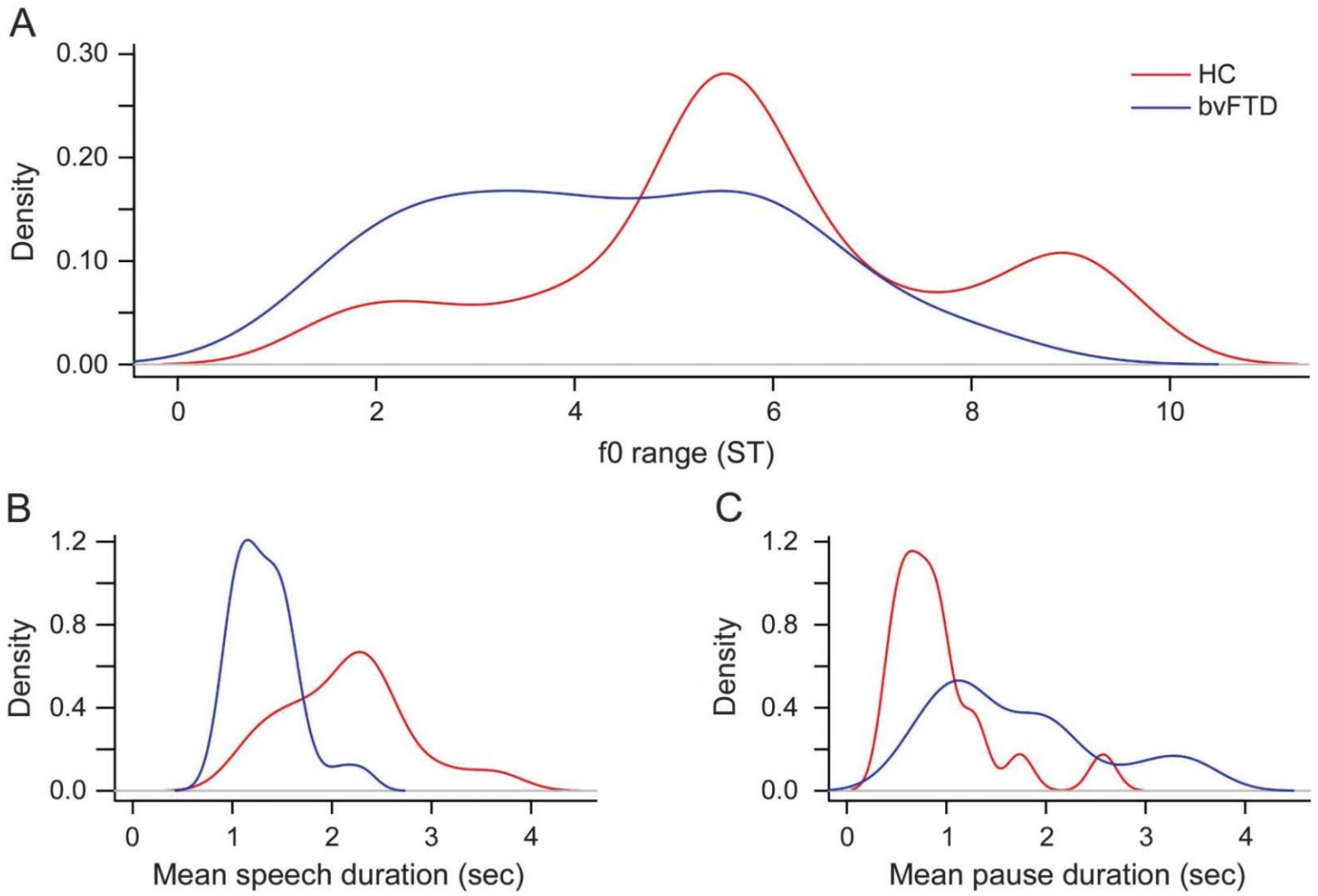
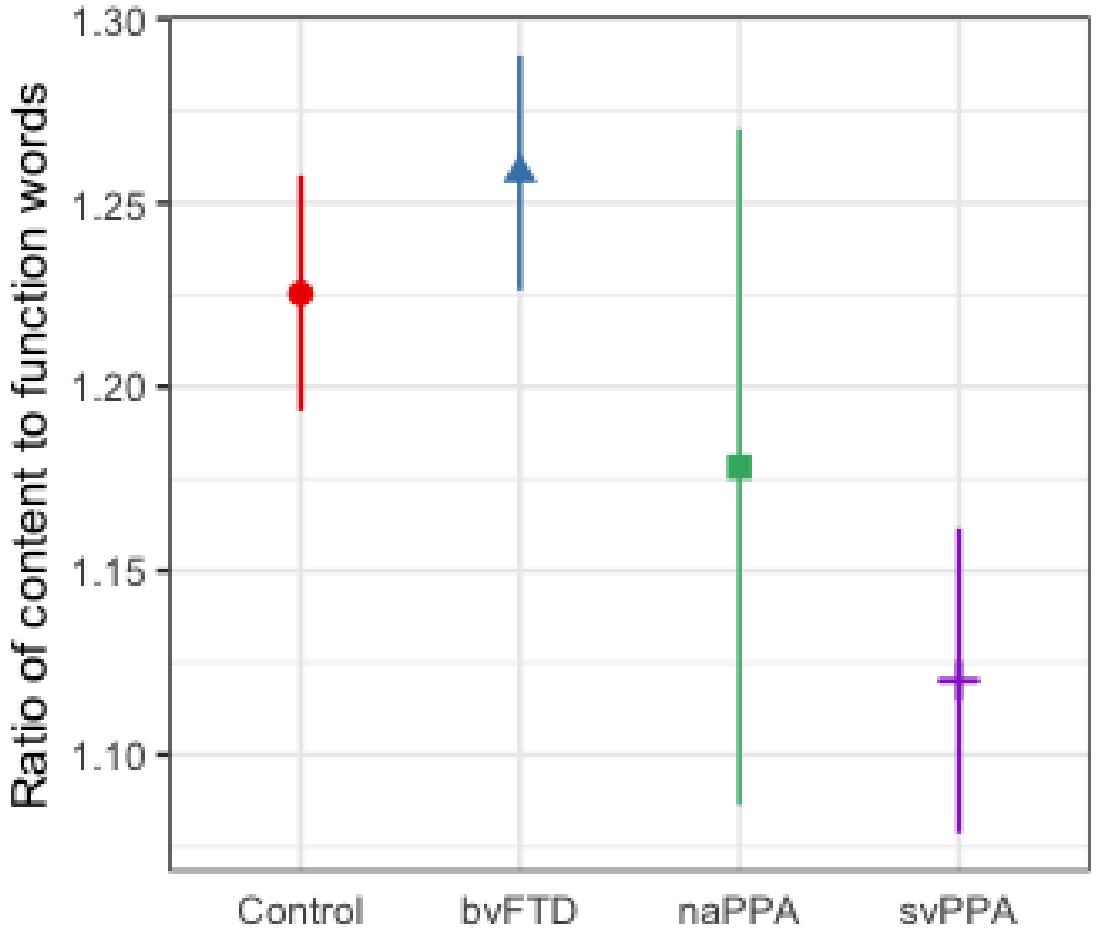
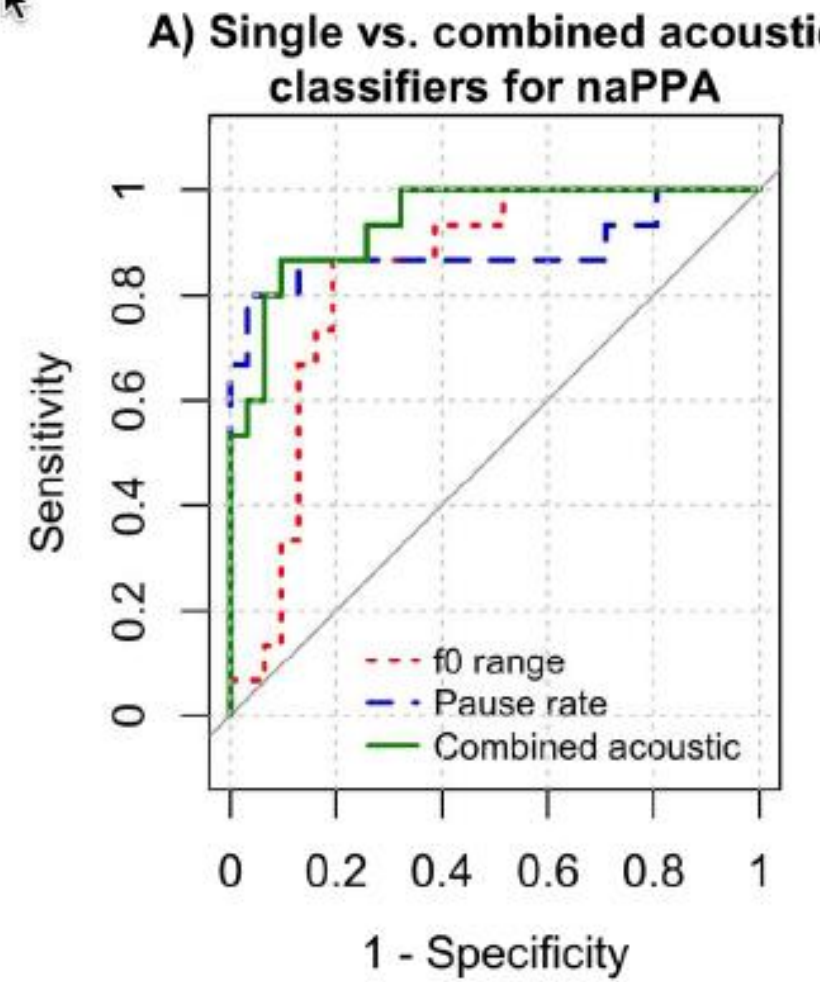
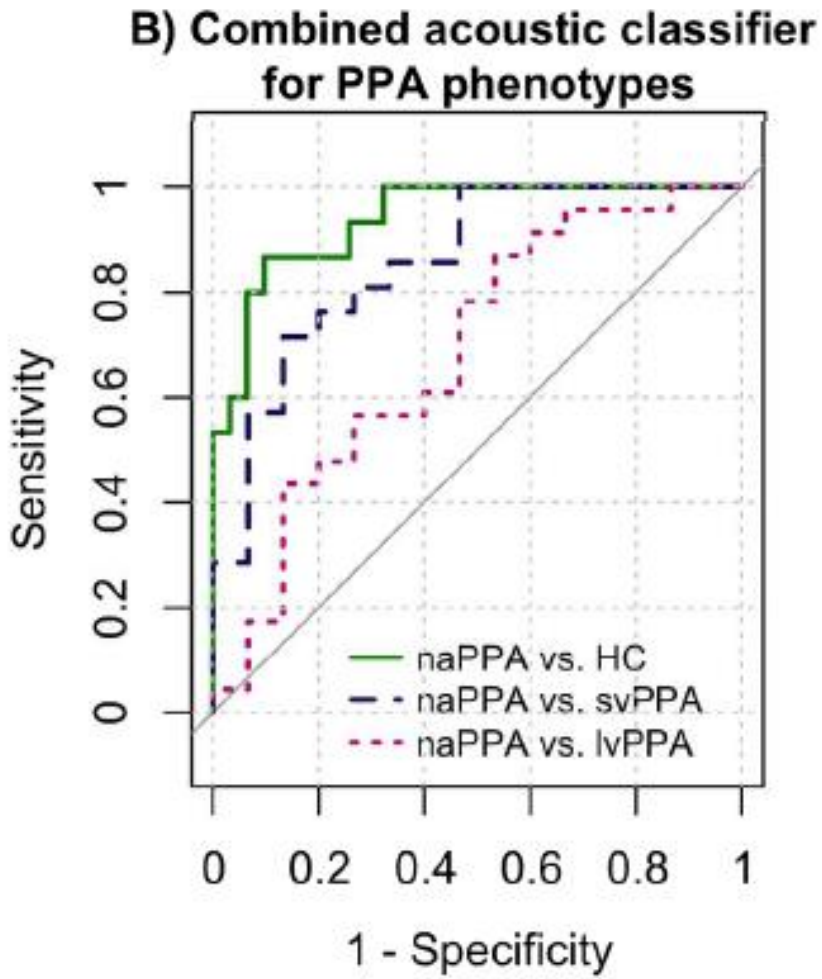

COOKIE
JAR

Figure 3 **Speech measures distributions**




Ratio of content to function words
1.30
1.25
1.20
1.15
1.10
Control
bvFTD
naPPA
svPPA


A) Single vs. combined acoustic classifiers for naPPA
Sensitivity
0
0.2
0.4
0.6
0.8
1
f0 range
Pause rate
Combined acoustic
0
0.2
0.4
0.6
0.8
1
1 - Specificity


B) Combined acoustic classifier for PPA phenotypes
Sensitivity
0
0.2
0.4
0.6
0.8
1
naPPA vs. HC
naPPA vs. svPPA
naPPA vs. lvPPA
0
0.2
0.4
0.6
0.8
1
1 - Specificity

The good news:

- Picture descriptions give a lot of diagnostic information (…that's why neurologists use them…)
- The task is quick, and easy to automate (e.g. using a web app)
- So this task could be part of a longitudinal test battery measuring linguistic and cognitive skills across time

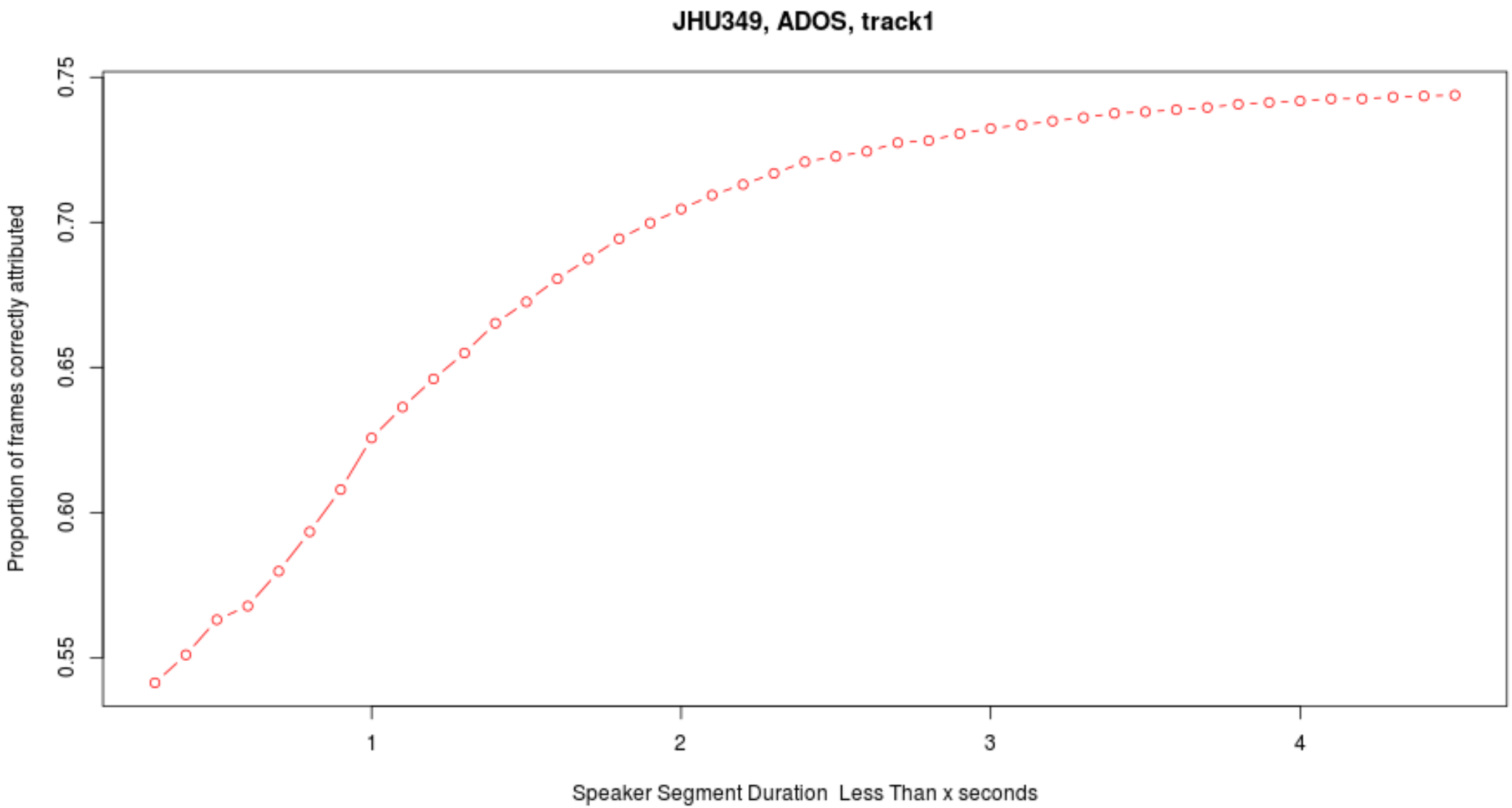
The bad news:

- Repeated description of the same picture is problematic
- There are only a couple of commonly-used pictures
- Even for those, there’s no basis for psychometric norming

Therefore we're planning to

- Create ~50 suitable pictures or short animations
- Get thousands of descriptions of each picture
  to permit psychometrically stable automated measures
- Combine with other standard tasks that can be automated
  - Digit span
  - "Fluency"
  - …etc…

With subject consent for data publication!


JHU349, ADOS, track1
Proportion of frames correctly attributed
0.55
0.60
0.65
0.70
0.75
1
2
3
4
Speaker Segment Duration Less Than x seconds

Interviewer: Okay, are you ready?
Subject: Yeah.
Interviewer: Okay.
Interviewer: Anna Thompson of South Boston,
Interviewer: employed as a scrubwoman in an office building,
Interviewer: reported at the city hall station
Interviewer: that she had been held up on state street
Interviewer: the night before and robbed of fifteen dollars.
Interviewer: She had four little children,
Interviewer: the rent was due, and they had not eaten for two days.
Interviewer: The officers, touched by the woman's story, made up a purse for her.
Interviewer: So I want you to tell me everything you can remember.

Speaker2: Are you ready?
Speaker1: Anna Thompson of South Boston
Speaker2: and Freud as a scrub woman in an office building reported at the city hall station that she had been held up on State Street the night before and Roz of
Speaker1: $15. She had for little
Speaker2: children. The rent was due and they had not eaten for two
Speaker1: days the officers
Speaker2: for the woman Story made
Speaker1: up a purse for her.
Speaker2: Don't you tell me everything you can remember?

Subject: This is bad.
Subject: This poor lady, she didn't have much money.
Subject: She didn't, uh, money to buy food.
Subject: You know, I just can't, I'm not good at this.
Subject: I'm not doing good on this one.

Speaker1: Flemington ice
Speaker1: food

Interviewer: For this next test, I'm going to say a letter of the alphabet,
Interviewer:: and I'd like you to give me as many words that start with that letter as quickly as you can.
Interviewer: So for example, if I say ~B,
Interviewer: you might give me words like bad, bottle, or bed.
Subject: uh huh
Interviewer: However, I don't want you to give me words that are proper names, such as Boston or Bob.
Subject: mhm
Interviewer: And I also don't want you to give me the same on again with a different ending,
Interviewer: such as bake, baking, or baked.
Interviewer: Do you have any questions?
Subject: oh I see, uh huh okay.
Interviewer: Okay?
Interviewer: Well the first letter is ~F.
Interviewer: Give me as many words that start with ~F as quickly as you can.
Interviewer: Go ahead and start.

Speaker1: British Max Plus I'm going to send a letter of the alphabet and I'd like you to give me as many words that start with that letter as quickly as you can. So for example, if I think he was like Dad said the same one again with a different ending baking or they do you have any questions but start with us as quickly as you can so having sex

Subject: Fuel,
Subject: face,
Subject: fuzz,
Subject: um
Subject: {breath}
Subject: {laugh}
Subject: uh floater,
Subject: uh
Subject: flag,
Subject: uh
Subject: film,
Subject: uh fish,
Subject: uh forest,
Subject: uh

Speaker1: You'll face.
Speaker1: pause
Speaker2: border film

The NAS reproducibility workshop was alarming –

There's a crisis of credibility
in many areas of scientific research,
as documented elsewhere before and since:

John Ioannidis, "Why Most Published Research Findings Are False",
PLoS Medicine 8/30/2005.

"Amid a Sea of False Findings, the NIH Tries Reform",
Chronicle of Higher Education 3/16/2015:

> ALS researchers, seeking a cure for Lou Gehrig's disease, went back and reproduced studies on more than 70 promising drugs. They found no real effects.
>
> "Zero of those were replicable," Dr. [Francis] Collins said. "Zero. And a couple of them had already moved into human clinical trials …"