Data Science in Healthcare: Beyond the Hype

Michael Becker
Penn Medicine

Agenda

- Hype
- Reality
- Solutions
- Future

The Hype

The Hype

Artificial Intelligence

IBM's Watson is better at diagnosing cancer than human doctors

By **IAN STEADMAN**

11 Feb 2013

The Hype

Artificial Intelligence

IBM's Watson is better at diagnosing cancer than human doctors

By **IAN STEADMAN**

11 Feb 2013

TECH INDUSTRY

IBM's Watson may provide a shortcut to treating cancer

The tech giant's cognitive computer system will help oncologists with the data-intensive work of identifying mutations in DNA and finding specific treatments.

BY **SAMANTHA RHODES** / JUNE 29, 2016 3:15 AM PDT

The Hype

Artificial Intelligence

IRM's Watson is better at

Frank Chen
@withfries2

Follow

Geoff Hinton: "we should stop training radiologists right now, in 5 years #deeplearning will have better performance" #mkt4intel

12:18 PM - 27 Oct 2016 from Toronto, Ontario

159 Retweets **164** Likes

CH INDUSTRY

n may provide a
treating cancer

ive computer system will help
ta-intensive work of identifying
d finding specific treatments.

ES / JUNE 29, 2016 3:15 AM PDT

The Hype

Artificial Intelligence

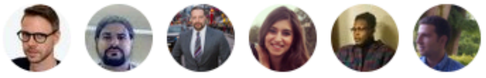
IRM's Watson is he

Frank Chen
@withfries2

Geoff Hinton: "we should sto
radiologists right now, in 5 y
#deeplearning will have bett
#mkt4intel

12:18 PM - 27 Oct 2016 from Toronto, Ontario

159 Retweets **164** Likes



Andrew Ng
@AndrewYNg

Follow

Should radiologists be worried about their jobs? Breaking news: We can now diagnose pneumonia from chest X-rays better than radiologists.
stanfordmlgroup.github.io/projects/chexn…

3:20 PM - 15 Nov 2017 from Mountain View, CA

1,432 Retweets **2,399** Likes

114 1.4K 2.4K

ES / JUNE 29, 2016 3:15 AM PDT

#PHLAI 159 164 @beckerfuffle

The Reality

The Reality

Emergent Tech ▸ Artificial Intelligence

Watson can't cure cancer ... or all the stuff that breaks IT projects

University spends $62m on AI trial, gets the usual trials that come with failure

By Richard Chirgwin 20 Feb 2017 at 06:56 18 SHARE ▼

The Reality

Emergent Tech ▸ Artificial Intelligence

Watson can't cure cancer ... or all the stuff that breaks IT projects

U[illegible] $62m on AI trial, gets the usual

tr

By

BRIEFING • IBM

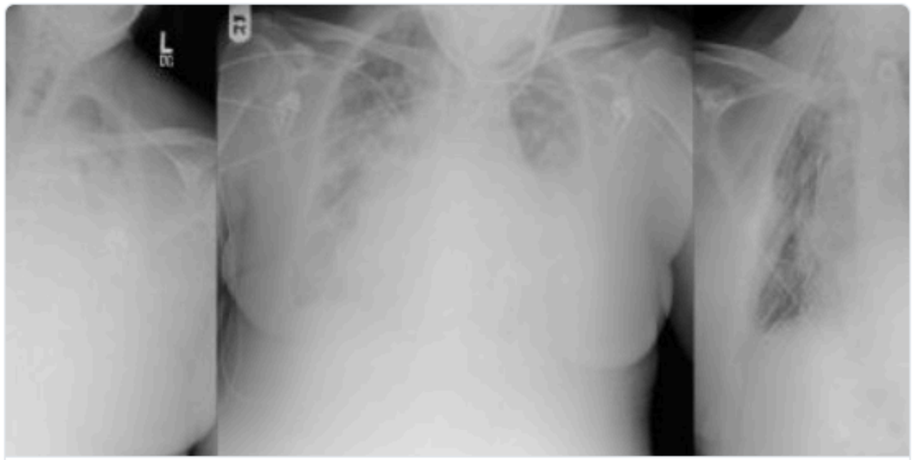
Some Cancer Treatment Recommendations From IBM's Watson Were Unsafe, Report Finds

The Reality

Luke Oakden-Rayner
@DrLukeOR

Follow

I've spent several weeks exploring the ChestXray14 dataset, and I have some serious concerns with it.



Exploring the ChestXray14 dataset: problems

A couple of weeks ago, I mentioned I had some concerns about the ChestXray14 dataset. I said I would come back when I had more info, an...

lukeoakdenrayner.wordpress.com

7:57 PM - 17 Dec 2017

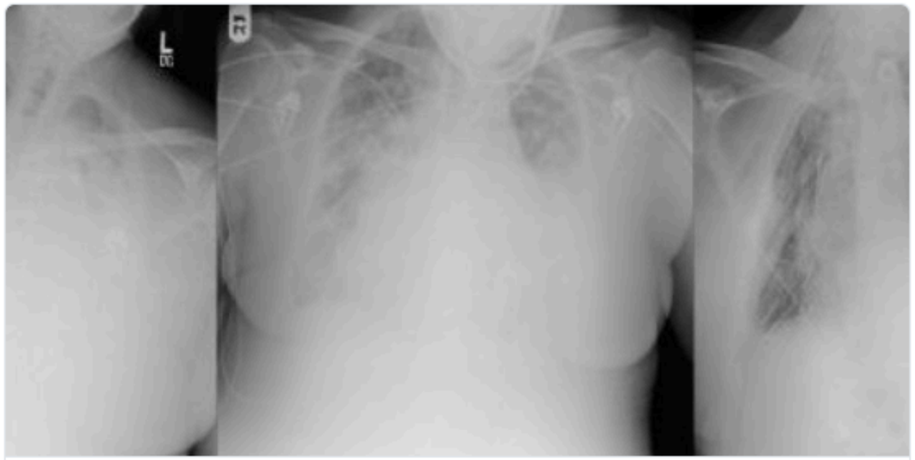
Tweets **885** Likes

The Reality

Luke Oakden-Rayner
@DrLukeOR

Follow

I've spent several weeks exploring the ChestXray14 dataset, and I have some serious concerns with it.



Exploring the ChestXray14 dataset: problems

A couple of weeks ago, I mentioned I had some concerns about the ChestXray14 dataset. I said I would come back when I had more info, an...

lukeoakdenrayner.wordpress.com

7:57 PM - 17 Dec 2017

885 Likes

Luke Oakden-Rayner
@DrLukeOR

Follow

My claim: In its current form it is not fit for training #medical #AI systems, and research on the dataset cannot generate valid medical claims without convincing additional justification.

TL:DR

- Compared to human visual assessment, the labels in the ChestXray14 dataset are inaccurate, unclear, and often describe medically unimportant findings.
- These label problems are *internally consistent* within the data, meaning models can show "good test-set performance", while still producing predictions that don't make medical sense.
- The above combination of problems mean the dataset as defined currently is **not fit** for training medical systems, and research on the dataset cannot generate valid medical claims without significant additional justification.
- Looking at the images is the basic "sanity check" of image analysis. If you don't have someone who can understand your data looking at the images when you build a dataset, expect things to go *very* wrong.
- Medical image data is *full* of stratifying elements; features than can help learn pretty much anything. Check that your model is doing what you think it is, every step of the way.
- I will be releasing some new labels with the next post, and show that deep learning *can* work in this dataset, as long as the labels are good enough.

8:16 PM - 17 Dec 2017

68 Retweets 188 Likes

The Reality

the dataset as defined currently is **not fit** for training medical systems

The Reality

Lior Pachter
@lpachter

Follow

Now today more data comes out: stanfordmlgroup.github.io/competitions/m… Guess what… the “deep learning” method sucks. It’s much worse than the radiologists. No way @AndrewYNg is trusting his own diagnoses to this algorithm. /3

	Radiologist 1	Radiologist 2	Radiologist 3	Model
Elbow	0.850 (0.830, 0.871)	0.710 (0.674, 0.745)	0.719 (0.685, 0.752)	0.710 (0.674, 0.745)
Finger	0.304 (0.249, 0.358)	0.403 (0.339, 0.467)	0.410 (0.358, 0.463)	0.389 (0.332, 0.446)
Forearm	0.796 (0.772, 0.821)	0.802 (0.779, 0.825)	0.798 (0.774, 0.822)	0.737 (0.707, 0.766)
Hand	0.661 (0.623, 0.698)	0.927 (0.917, 0.937)	0.789 (0.762, 0.815)	0.851 (0.830, 0.871)
Humerus	0.867 (0.850, 0.883)	0.733 (0.703, 0.764)	0.933 (0.925, 0.942)	0.600 (0.558, 0.642)
Shoulder	0.864 (0.847, 0.881)	0.791 (0.765, 0.816)	0.864 (0.847, 0.881)	0.729 (0.697, 0.760)
Wrist	0.791 (0.766, 0.817)	0.931 (0.922, 0.940)	0.931 (0.922, 0.940)	0.931 (0.922, 0.940)
Overall	0.731 (0.726, 0.735)	0.763 (0.759, 0.767)	0.778 (0.774, 0.782)	0.705 (0.700, 0.710)

3:01 PM - 24 May 2018

Retweets **328** Likes

The Reality

Lior Pachter
@lpachter

Follow

Now today more data comes out:
stanfordmlgroup.github.io/competitions/

	Radiologist 1	Radiologist 2	Radiologist 3	Model
Elbow	0.850 (0.830, 0.871)	0.710 (0.674, 0.745)	0.719 (0.685, 0.752)	0.710 (0.674, 0.745)
Finger	0.304 (0.249, 0.358)	0.403 (0.339, 0.467)	0.410 (0.358, 0.463)	0.389 (0.332, 0.446)
Forearm	0.796 (0.772, 0.821)	0.802 (0.779, 0.825)	0.798 (0.774, 0.822)	0.737 (0.707, 0.766)
Hand	0.661 (0.623, 0.698)	0.927 (0.917, 0.937)	0.789 (0.762, 0.815)	0.851 (0.830, 0.871)
Humerus	0.867 (0.850, 0.883)	0.733 (0.703, 0.764)	0.933 (0.925, 0.942)	0.600 (0.558, 0.642)
Shoulder	0.864 (0.847, 0.881)	0.791 (0.765, 0.816)	0.864 (0.847, 0.881)	0.729 (0.697, 0.760)
Wrist	0.791 (0.766, 0.817)	0.931 (0.922, 0.940)	0.931 (0.922, 0.940)	0.931 (0.922, 0.940)
Overall	0.731 (0.726, 0.735)	0.763 (0.759, 0.767)	0.778 (0.774, 0.782)	0.705 (0.700, 0.710)

3:01 PM - 24 May 2018

Retweets **328** Likes

The Reality

Lior Pachter
@lpachter

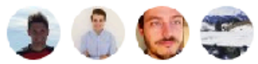
Now today more data c

stanfordmlgroup.github

	Radiologist 1
Elbow	0.850 (0.830, 0.871)
Finger	0.304 (0.249, 0.358)
Forearm	0.796 (0.772, 0.821)
Hand	0.661 (0.623, 0.698)
Humerus	0.867 (0.850, 0.883)
Shoulder	0.864 (0.847, 0.881)
Wrist	0.791 (0.766, 0.817)
Overall	0.731 (0.726, 0.735)

3:01 PM - 24 May 2018

Retweets **328** Likes



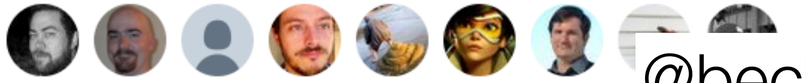
Lior Pachter
@lpachter

Follow

You know what else sucks? When high profile machine learning people oversell their results to the public. It leaves everyone worse off… because how can us mere mortals publish a paper if we haven't rendered an entire profession obsolete with our results? /4

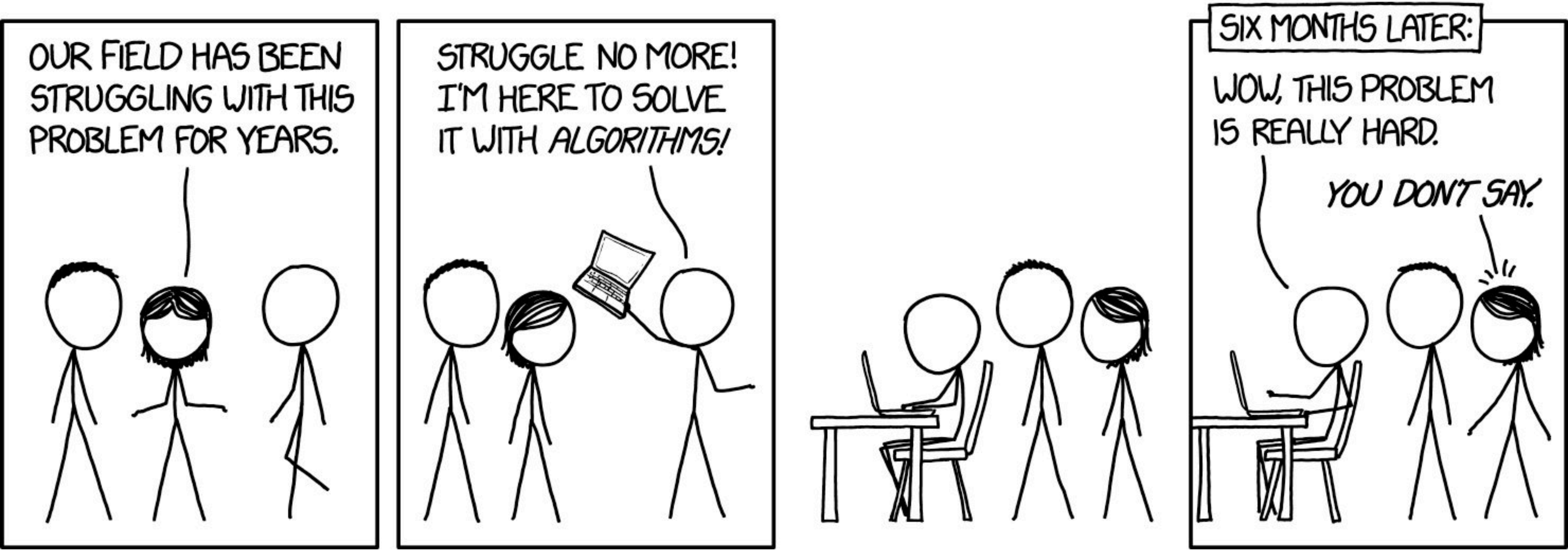
3:01 PM - 24 May 2018

124 Retweets **501** Likes



#PHLAI

@beckerfuffle


OUR FIELD HAS BEEN STRUGGLING WITH THIS PROBLEM FOR YEARS.
STRUGGLE NO MORE! I'M HERE TO SOLVE IT WITH ALGORITHMS!
SIX MONTHS LATER:
WOW, THIS PROBLEM IS REALLY HARD.
YOU DON'T SAY.

Eric Topol 
@EricTopol

Follow

—How good is #AI for prediction in medicine?
—We don't know.
All 12 papers I've summarized are in silico, retrospective, many w/ statistical methodologic issues, and we've yet to see a prospective validation study in a real world clinical environment.

Prediction	N	AUC	Reference
In-hospital mortality, unplanned readmission, prolonged LOS, final discharge diagnosis	216,221	0.93* 0.75+ 0.85#	Rajkomar et al, Nature NPJ Digital Medicine, 2018
All-cause 3-12 month mortality	221,284	0.93^	Avati et al, arXiv, 2017
Readmission	1,068	0.78	Shameer et al, Pacific Symposium on Biocomputing, 2017
Sepsis	230,936	0.67	Horng et al, PLOS One, 2017
Septic shock	16,234	0.83	Henry et al, Science, 2015
C. Difficile infection	256,732	0.82++	Oh et al, Infection Control and Epidemiology, 2018
Developing diseases	704,587	range	Miotto et al, Scientific Reports, 2018
Diagnosis	18,590	0.96	Yang et al, Scientific Reports, 2018
Dementia	76,367	0.91	Cleret de Langavant et al, J Internet Med Res 2018
Alzheimer's Disease (+ amyloid imaging)	273	0.91	Mathotaarachchi et al, Neurobiology of Aging, 2017
Mortality after cancer chemotherapy	26,946	0.94	Elfiky et al, JAMA Open, 2018
Disease onset for 133 conditions	298,000	range	Razavian et al, arXiv, 2016

AUC-area under the curve, a metric of accuracy, LOS-length of stay, N-Number of patients (training + validation datasets), *-in-hospital mortality, +-unplanned readmission, #-prolonged LOS, ^-all patients, ++-for U of Michigan site

8:33 AM - 29 Jul 2018

617 Retweets **922** Likes 

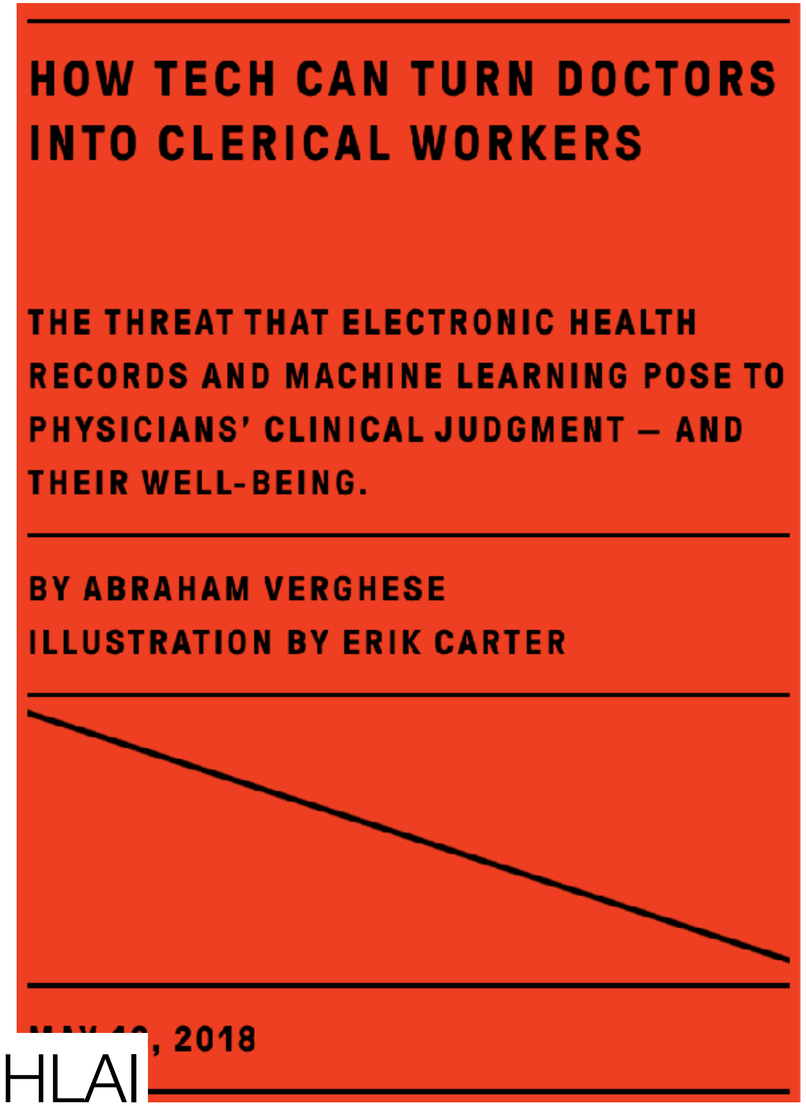
No published data

Eric Topol
@EricTopol

Follow

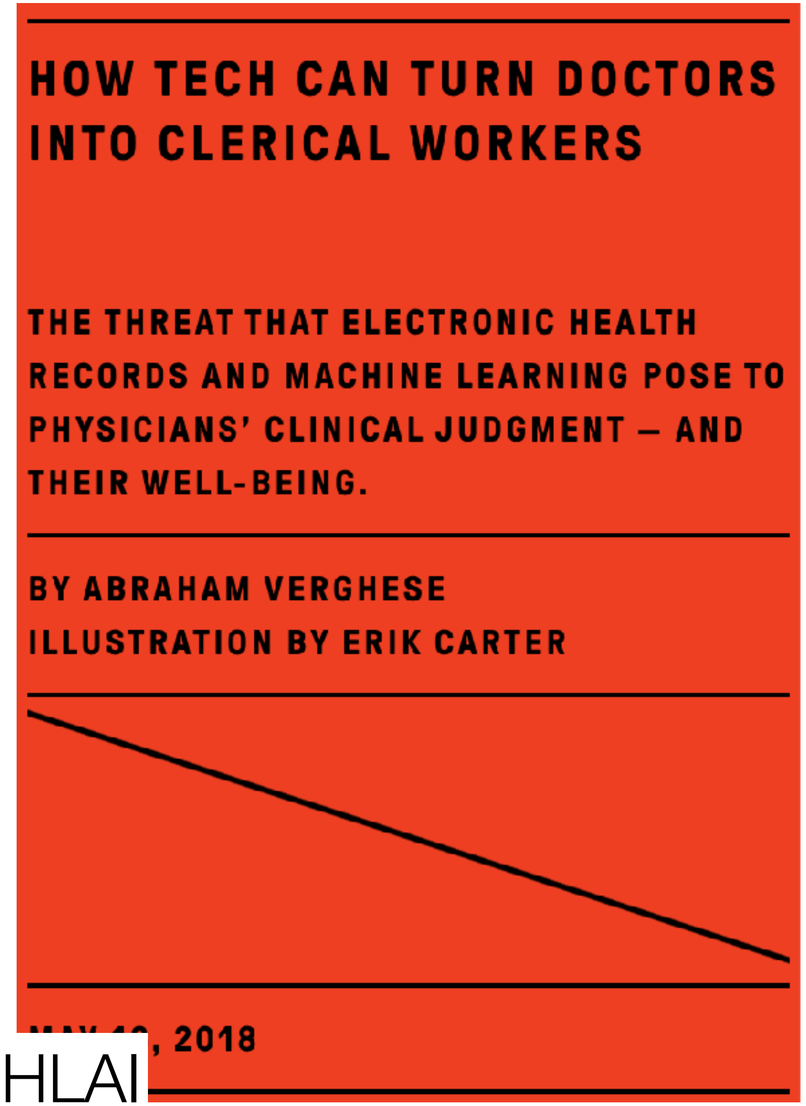
The 1st @US_FDA approved deep learning #AI device for 1°care today: diagnosis of diabetic retinopathy w/o ophthalmologists.
fda.gov/NewsEvents/New …
But no published data. Website claims 87% sensitivity, 90% specificity from 900 patient trial.
+ @SGottliebFDA's upbeat #AI tweets.

Burnout



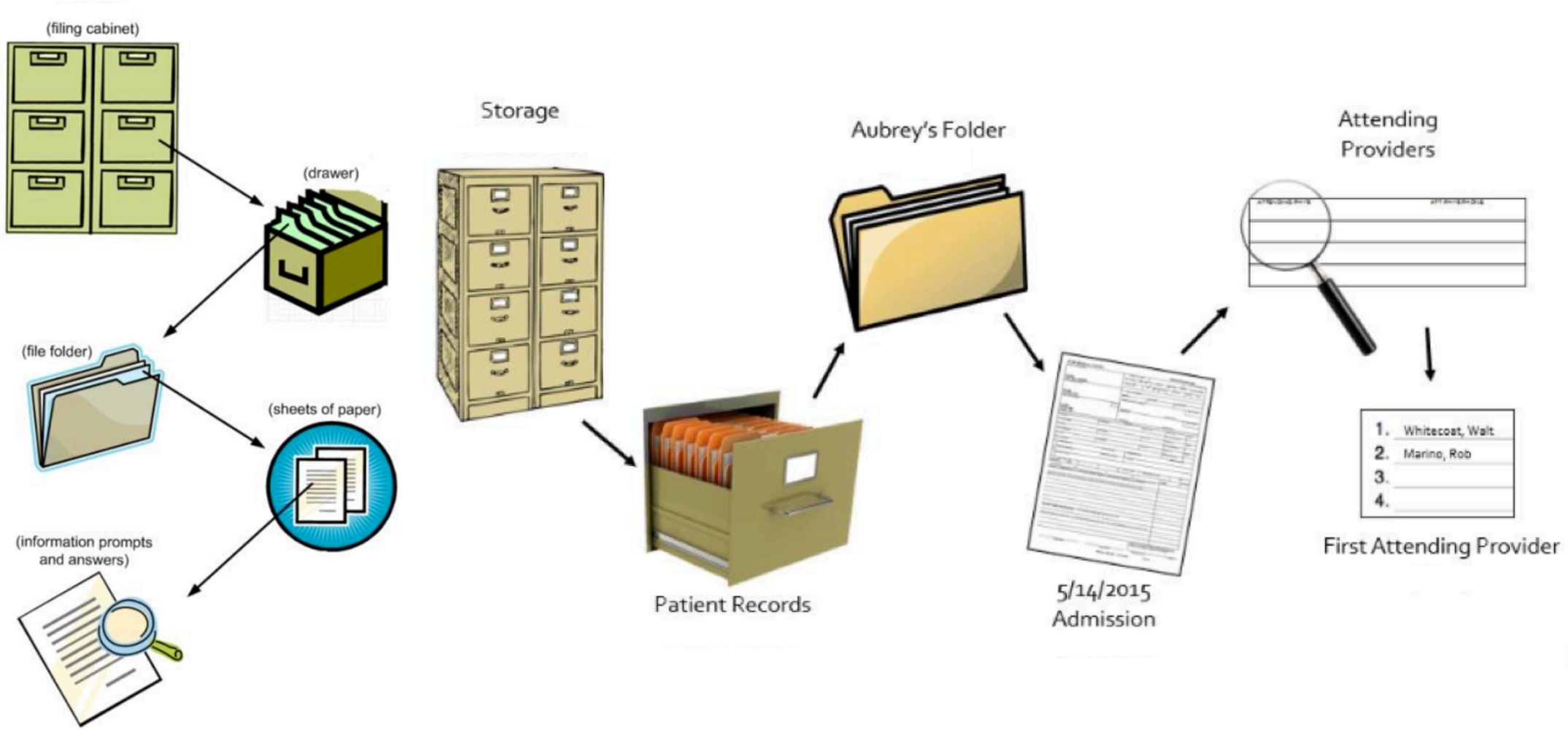
By some estimates, **more than 50 percent of physicians in the United States have at least one symptom of burnout**, defined as a syndrome of emotional exhaustion, cynicism and decreased efficacy at work. It is on the increase, up by 9 percent from 2011 to 2014 in one national study.

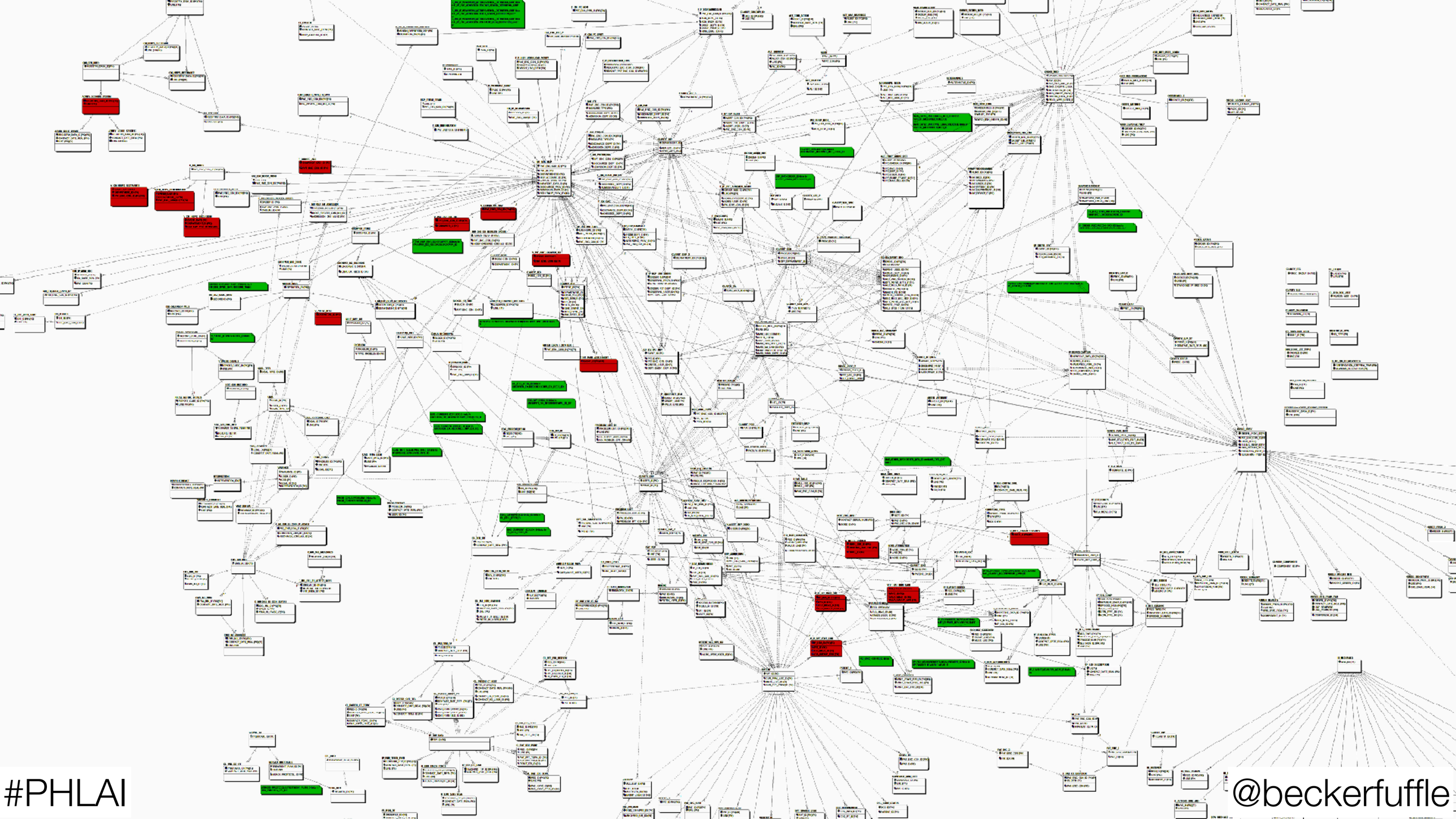
Burnout



a 4,000-key-clicks-a-day problem. The E.H.R. is only part of the issue: Other factors include rapid patient turnover, decreased autonomy, merging hospital systems, an aging population, the increasing medical complexity of patients. Even if **the E.H.R.** is not the sole cause of what ails us, believe me, it **has become the symbol of burnout.**

“Enterprise” Clinical Databases





#PHLAI

@beckerfuffle

Messy human data

Messy human data

- “unable to take temperature, patient just ate ice”

Messy human data

- “unable to take temperature, patient just ate ice”
- “Pt c/o feeling hot, 'like I have a fever’. Deferred temp reassessment for 10min-pt recently sipped ice water”

Messy human data

- “unable to take temperature, patient just ate ice”
- “Pt c/o feeling hot, 'like I have a fever’. Deferred temp reassessment for 10min-pt recently sipped ice water”

> If you're experiencing an insatiable **craving** to eat **ice**, you may have a condition called pica. ... If **ice** is the substance you **crave**, then you may have a type of pica called pagophagia. While there's no single cause of pica or pagophagia, they can occur if you have iron deficiency anemia. May 24, 2018

The Solutions

ActiveProtective

http://www.philly.com/philly/business/small_business/active-protective-airbag-belt-genesis-healthcare-20180316.html

ActiveProtective

http://www.philly.com/philly/business/small_business/active-protective-airbag-belt-genesis-healthcare-20180316.html

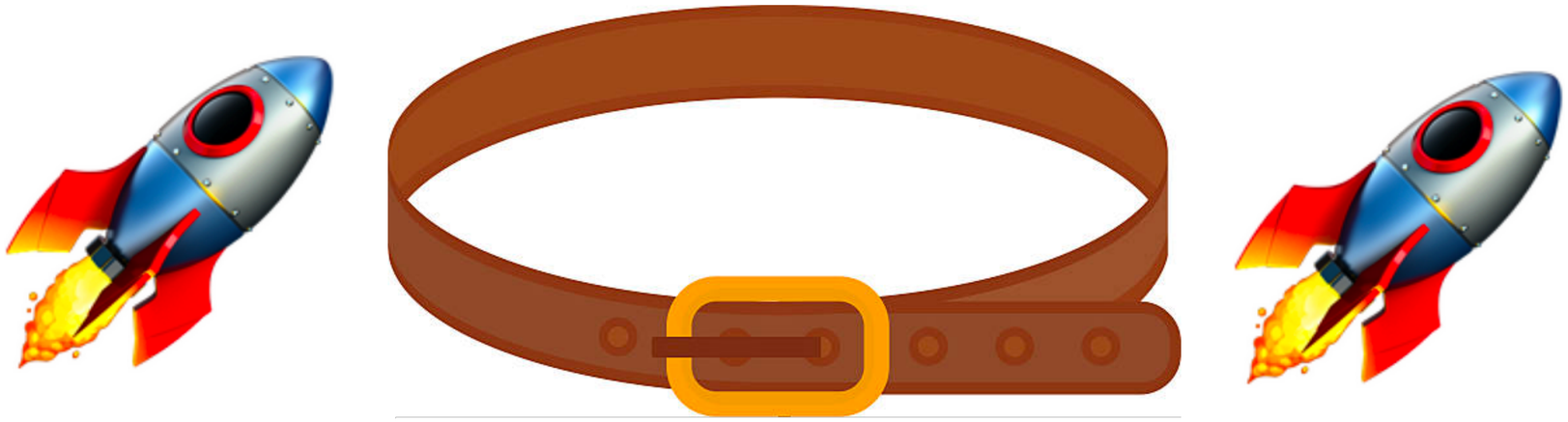
ActiveProtective

http://www.philly.com/philly/business/small_business/active-protective-airbag-belt-genesis-healthcare-20180316.html

ActiveProtective

http://www.philly.com/philly/business/small_business/active-protective-airbag-belt-genesis-healthcare-20180316.html

Unobservable system



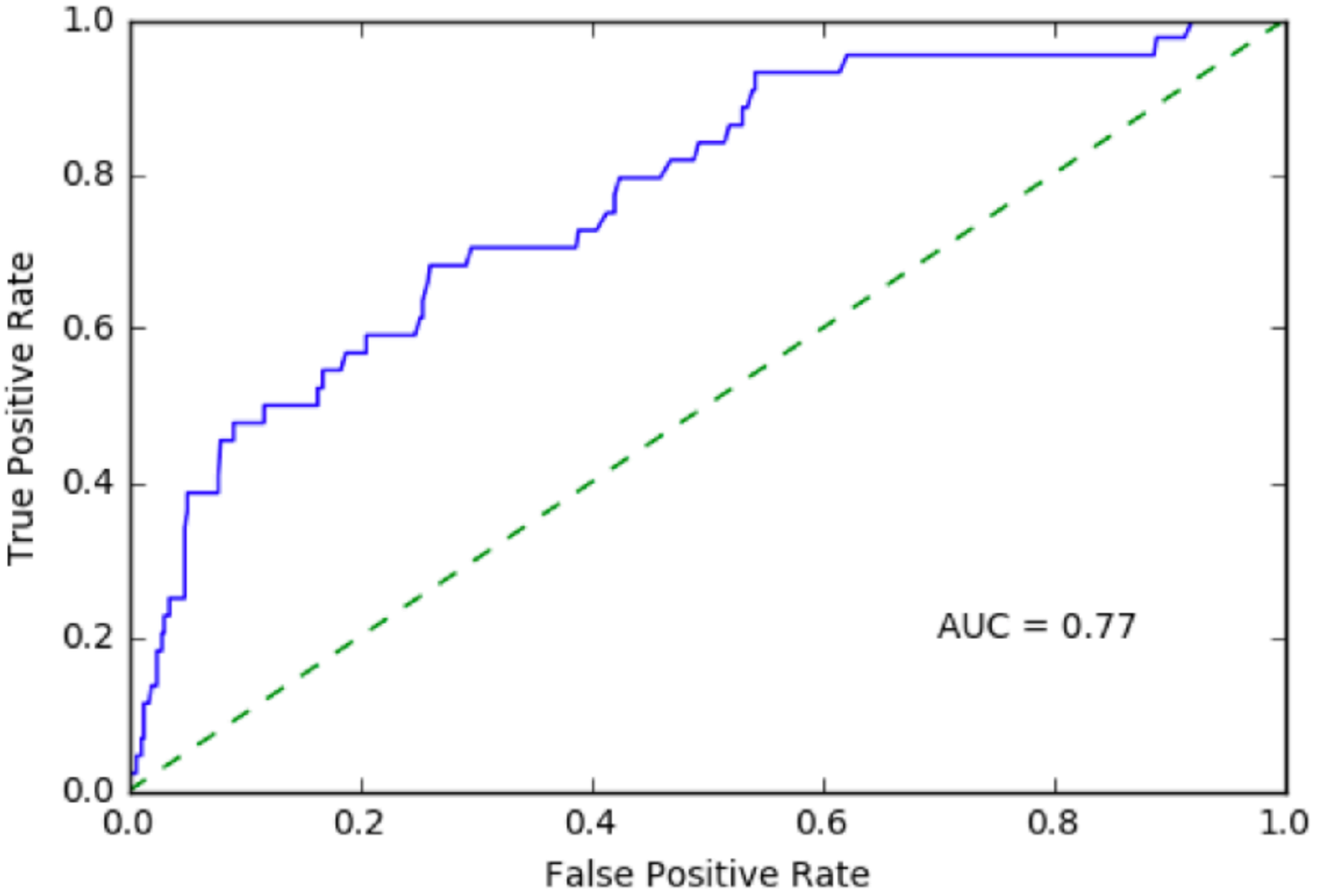
Experiment Design

Airbag System
Completely Observable

Jetpack System
Intervention impacts prediction

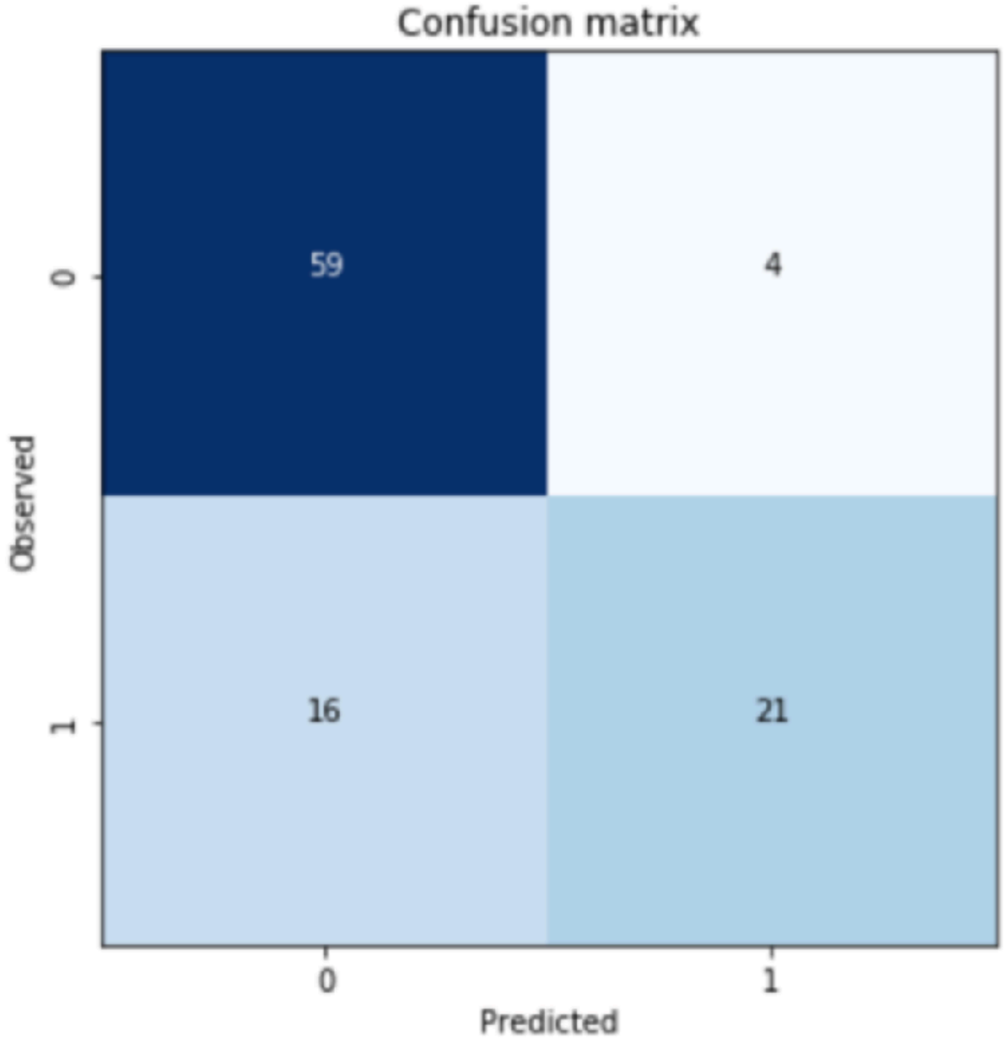
Is your solution an Airbag or a Jetpack? Can you design it like an Airbag?

Is this a good model?



http://predictivehealthcare.pennmedicine.org/2017/11/07/from-predictions-to-decisions.html

Is this a good model?



http://predictivehealthcare.pennmedicine.org/2017/11/07/from-predictions-to-decisions.html

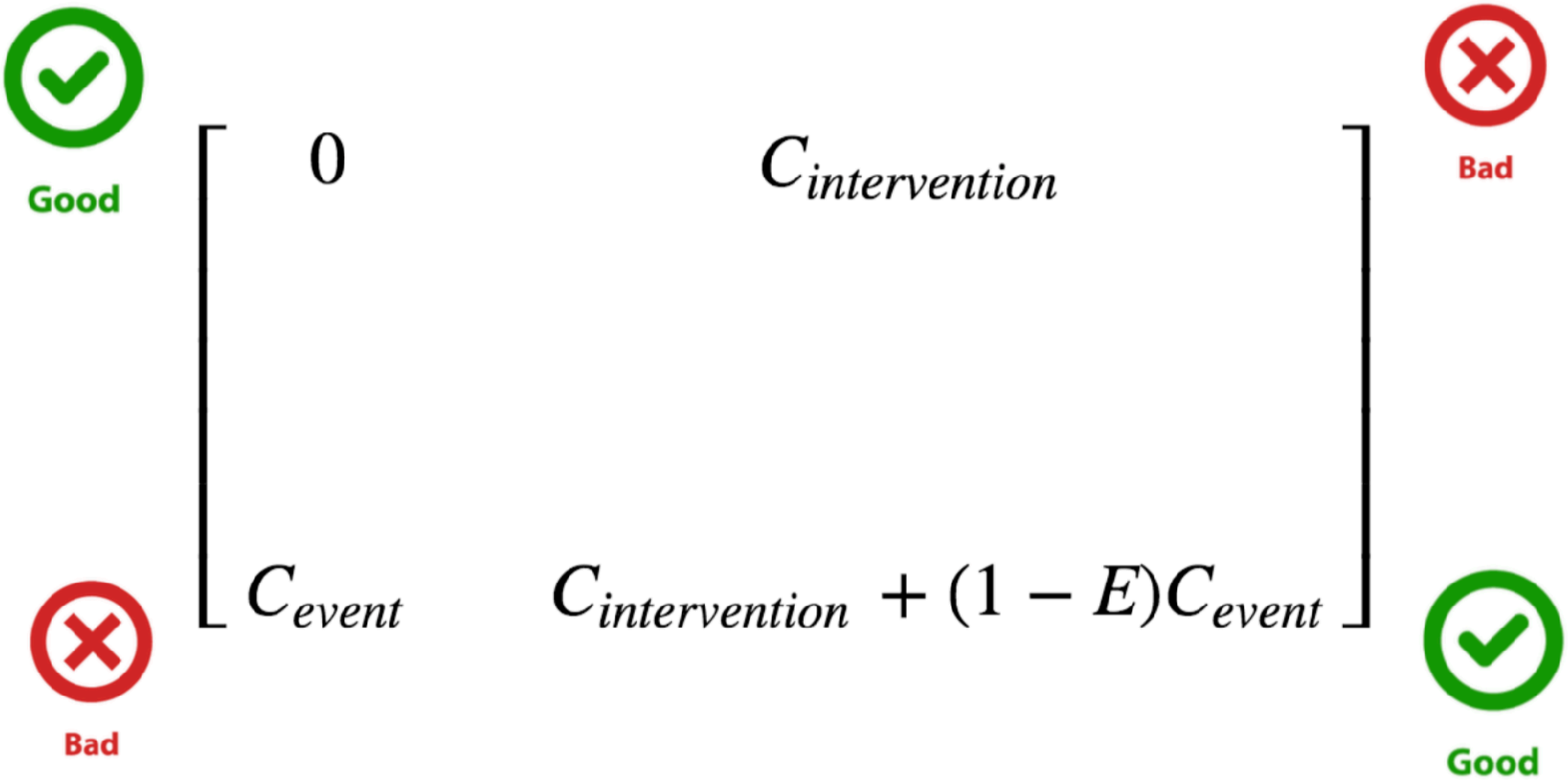
When outcomes are uncertain, the best decision is the one that has the highest *expected* goodness

---

$$\mathbb{E}[Goodness] = \sum_{i \in I} Pr(outcome_i)\ x\ Goodness(outcome_i)$$

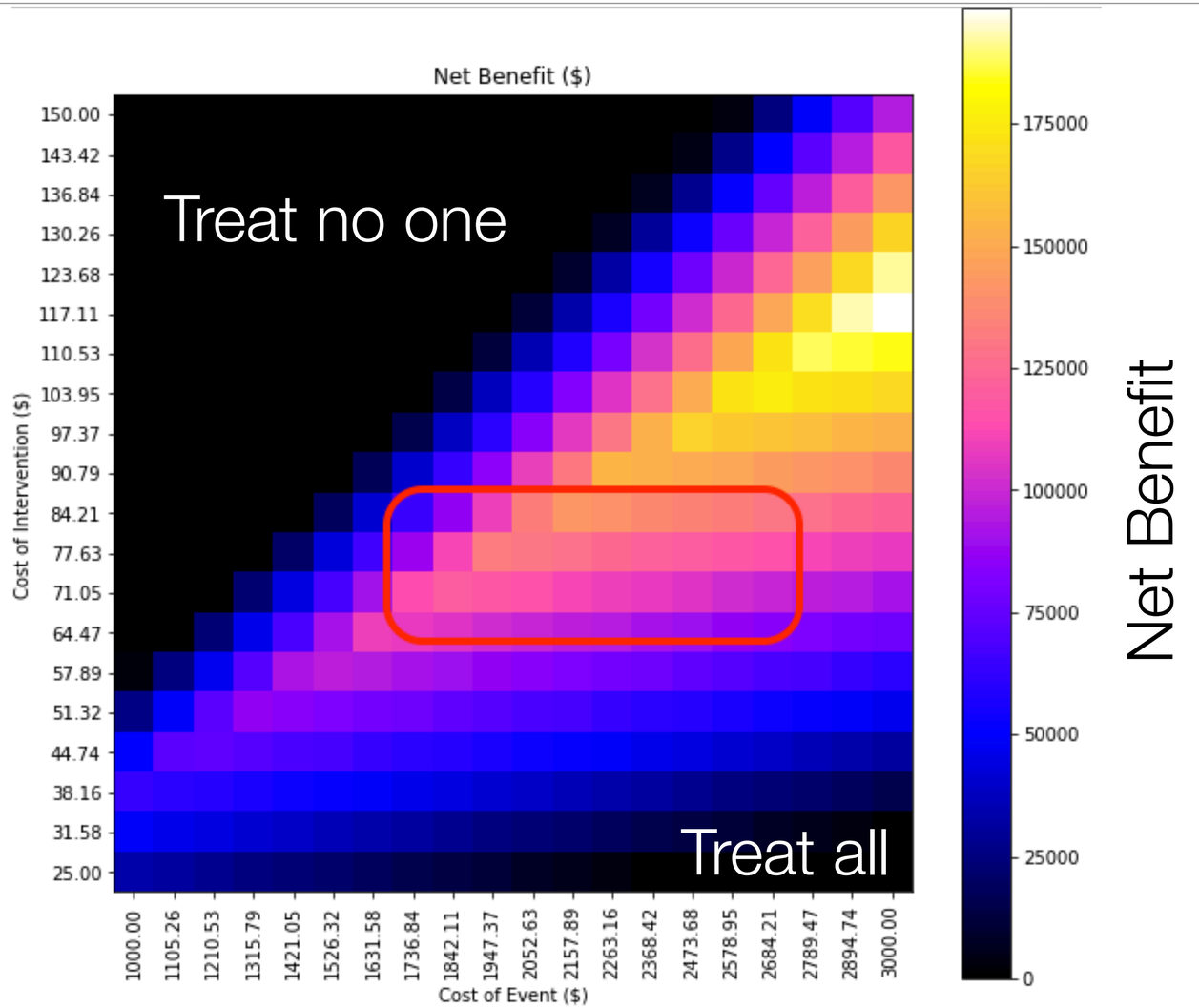
http://predictivehealthcare.pennmedicine.org/2017/11/07/from-predictions-to-decisions.html

Heathcare Example



http://predictivehealthcare.pennmedicine.org/2017/11/07/from-predictions-to-decisions.html

Decision theory



http://predictivehealthcare.pennmedicine.org/2017/11/07/from-predictions-to-decisions.html

If I knew **[information]**
I would do **[intervention]**
to improve **[measurable outcome]**

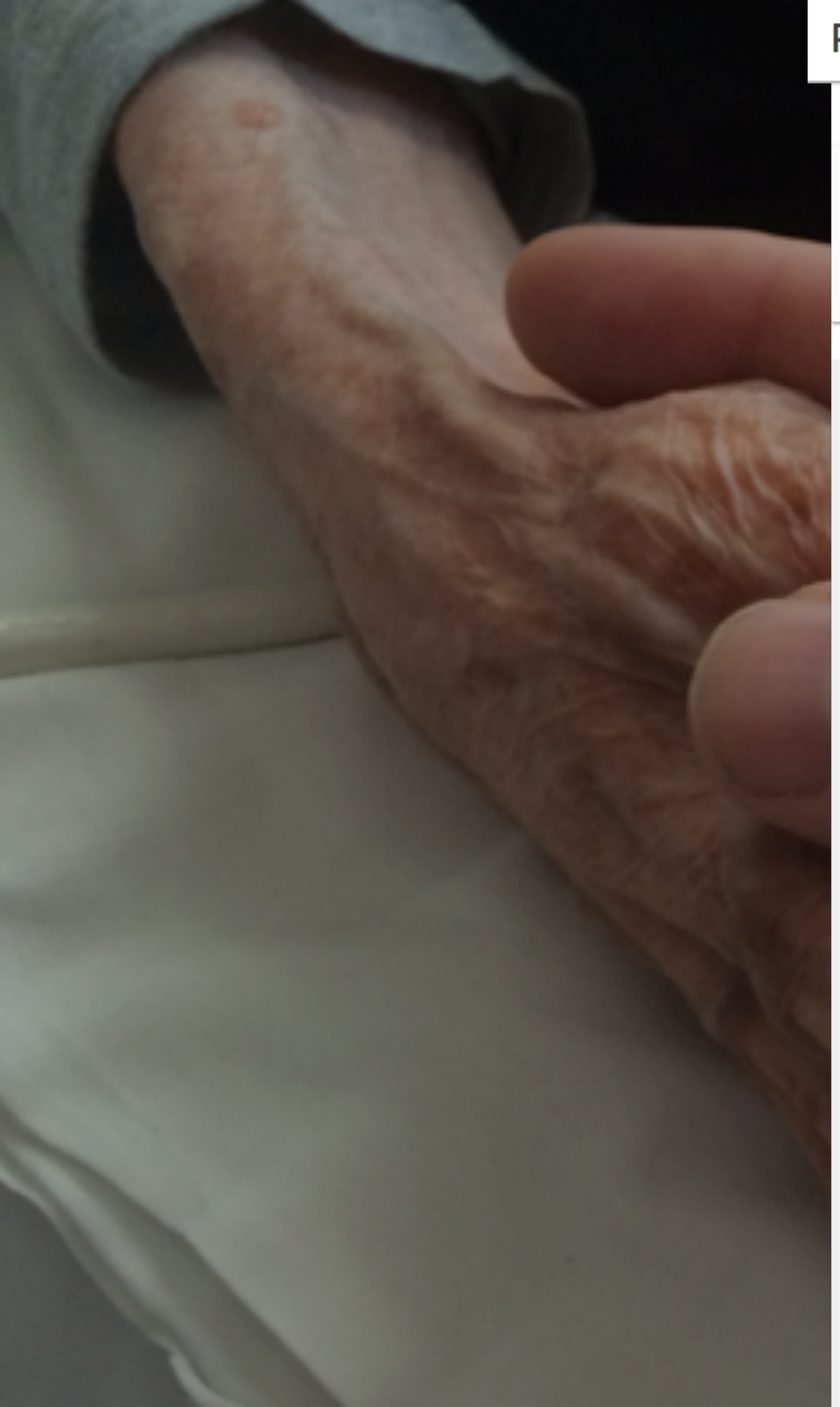
Many patients receive Palliative Care too close to their last days. The delayed delivery undermines the full utility of this service for patients in need.

https://www.pennmedicine.org/news/news-blog/2018/june/palliative-connect-digitizing-the-physicians-intuition-to-prompt-critical-conversations

https://www.pennmedicine.org/news/news-blog/2018/june/palliative-connect-digitizing-the-physicians-intuition-to-prompt-critical-conversations

I would **make sure they receive palliative care** to ensure **the care team understands their goals**

https://www.pennmedicine.org/news/news-blog/2018/june/palliative-connect-digitizing-the-physicians-intuition-to-prompt-critical-conversations



Palliative Connect

Welcome Nina!

Logout

Filter by location

-- Any --

New

Following

DOE, JANE

Age: 00

Location: PAH

Service: Medicine PAH, Medicine 8A

Following

Completed

Declined

- Select a reason -

Add Note

SUBMIT NOTE

View Notes

1 new

View Actions

0 new

DOE, JOHN

Age: 99

Location: HUP

Service: MEDICINE HUP, ONC SOLID

APP

Following

Completed

Declined

- Select a reason -

Add Note

SUBMIT NOTE

View Notes

0 new

View Actions

0 new

The Future

The Future

https://www.intelegain.com/ai-in-healthcare/

“Algorithms don’t heal people, people do.”