February 5, 2024

HARMONIZING HEALTH AND AI: *NAVIGATING INNOVATION AND ETHICS*

Tamiko Eto

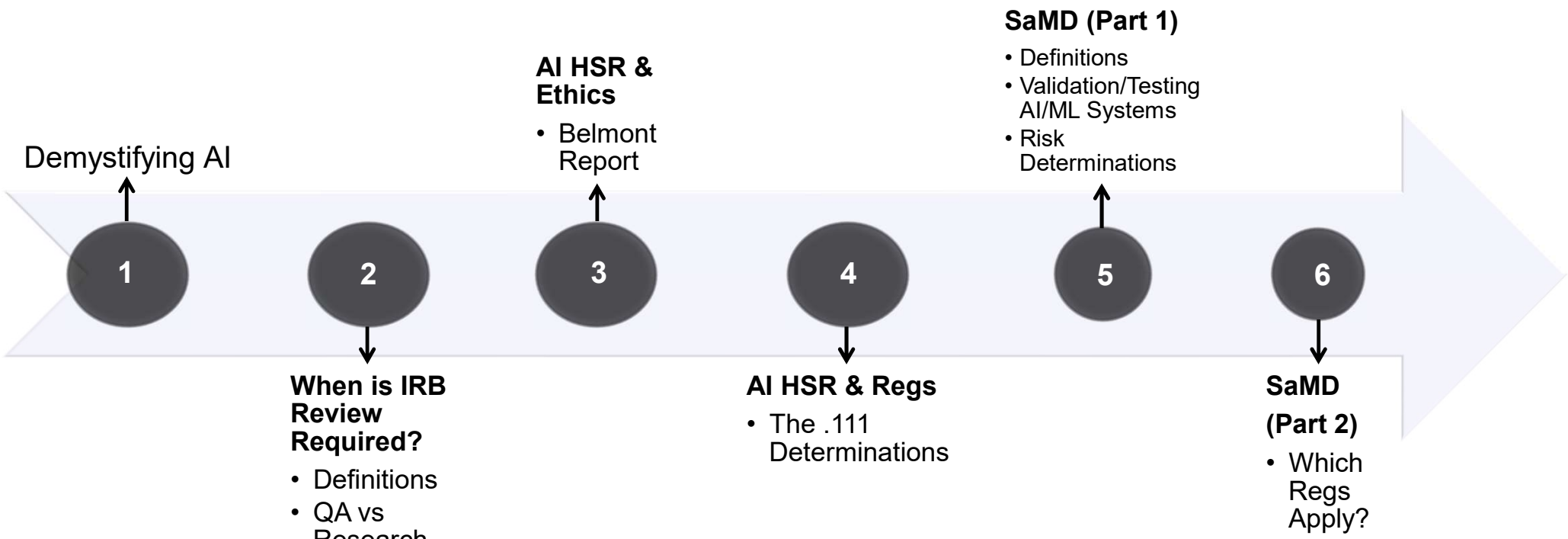
Director, *Research Operations,*
Human Research Protection Program (HRPP) and
Institutional Review Board (IRB)
Mayo Clinic

-

Contact: Eto.Tamiko@Mayo.edu

CARE-Q
Consortium for Applied Research Ethics Quality

Learning Objectives



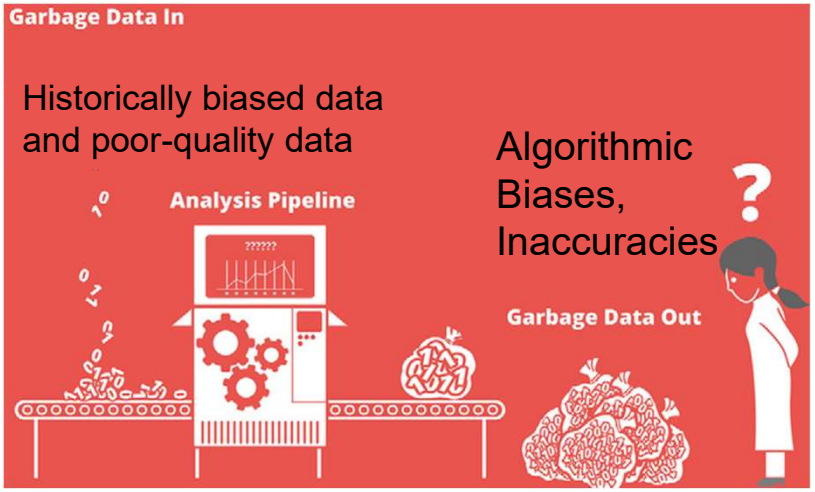
1 DEMYSTIFYING AI/ML

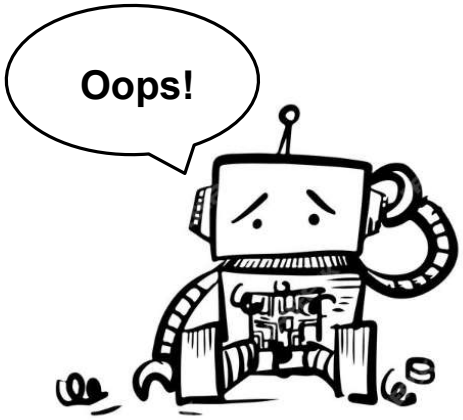
Why Does Any of This Matter?

Opacity

Irreversible Impact

- Hiring, lending, incrimination, misdiagnosis/treatment, etc.)
- No legal pathway for victims





- **Emergent Behavior**
- **Unintended consequences**

A Long History of Regulating AI in the U.S.

AI
Expert Systems
of the 60's

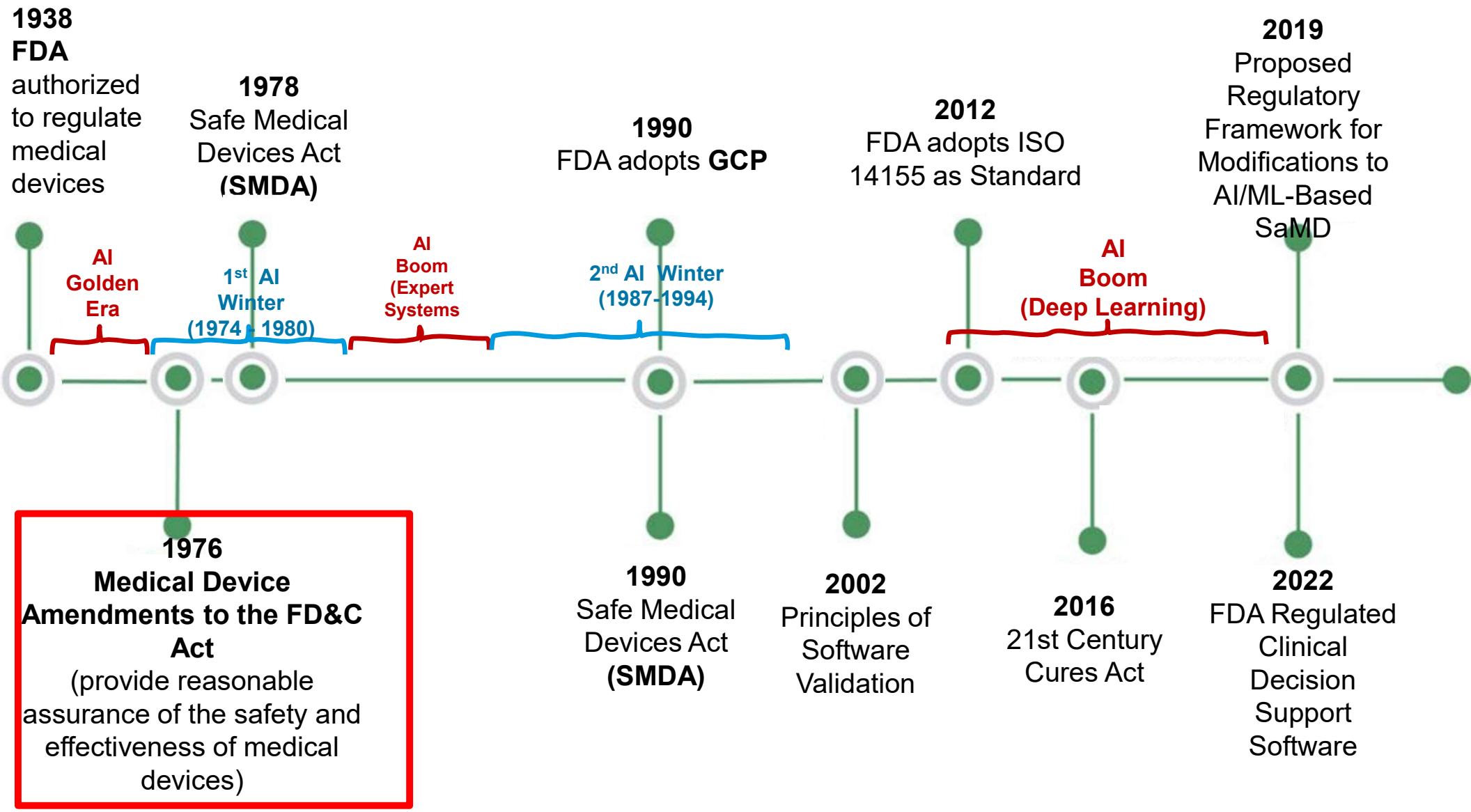
The first AI Clinical Decision-Making Tool

BENEFITS

- Cheaper analysis
- Convenient PC-based tool

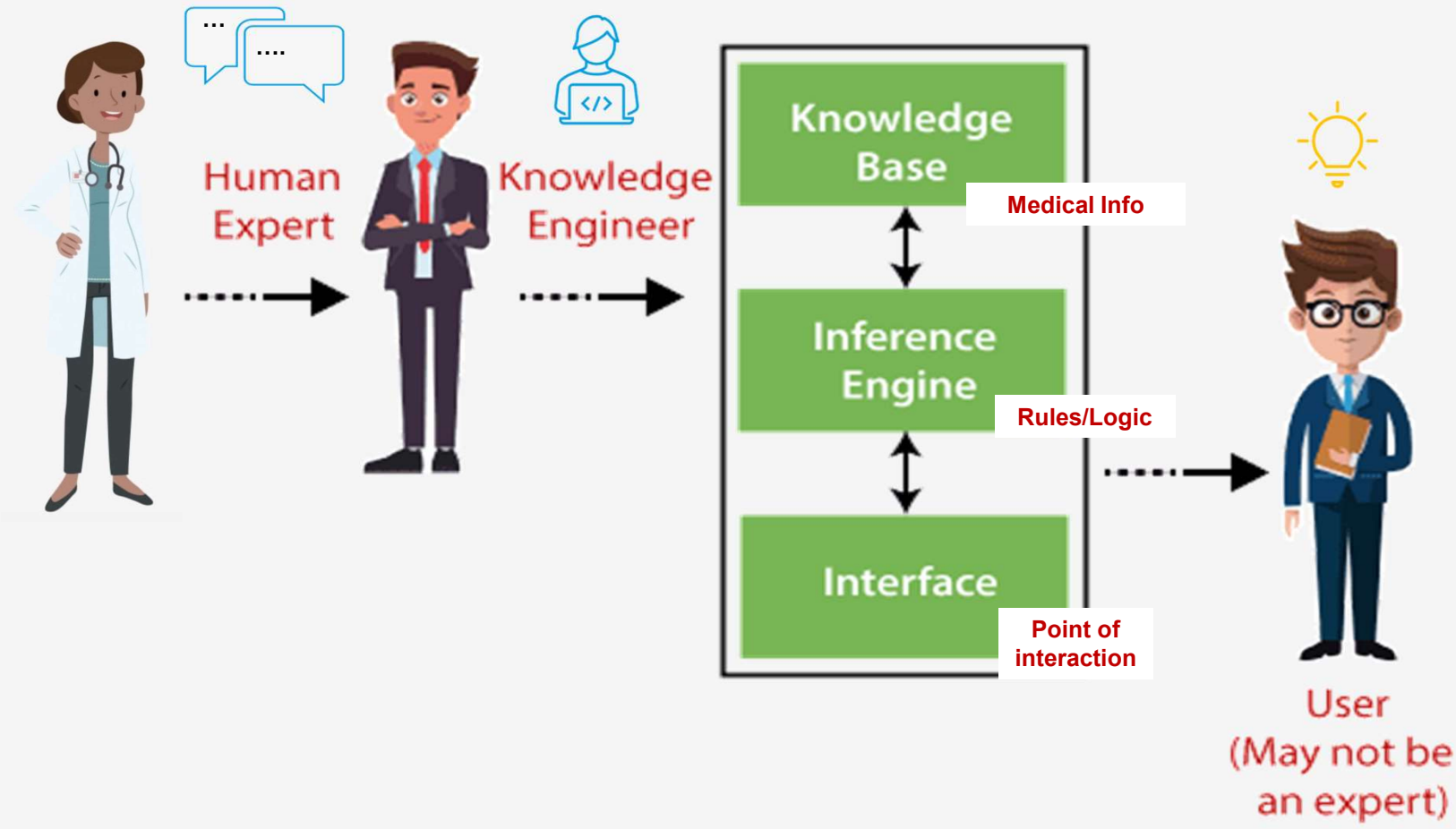
CHALLENGES

- **Liability**
- System integration
- Development reliant on end users
- Output conflicts with original intentions
- Budget constraints

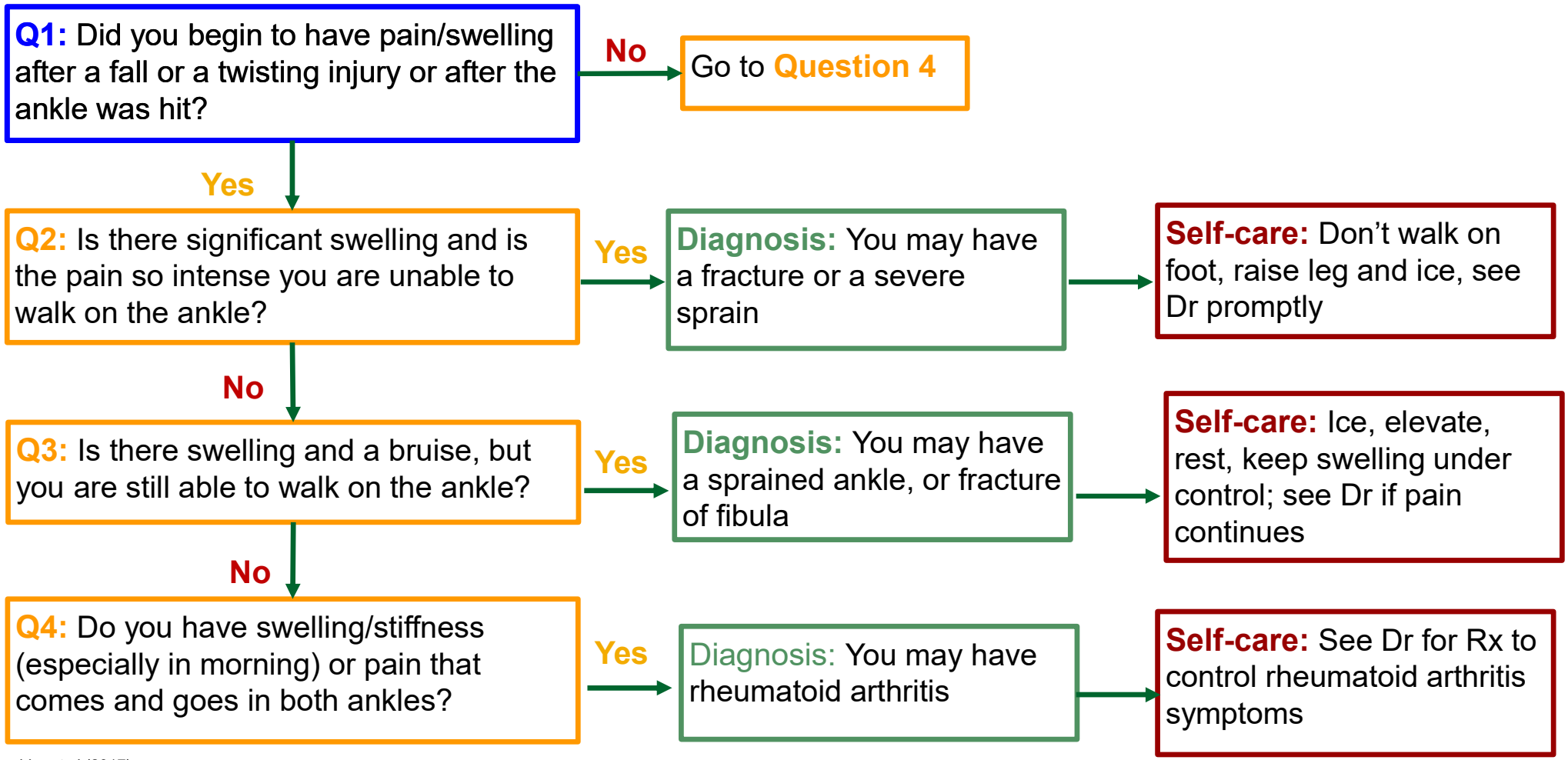

1938
FDA
authorized
to regulate
medical
devices
1978
Safe Medical
Devices Act
(SMDA)
1990
FDA adopts GCP
2012
FDA adopts ISO
14155 as Standard
2019
Proposed
Regulatory
Framework for
Modifications to
AI/ML-Based
SaMD
AI
Golden
Era
1st AI
Winter
(1974 - 1980)
AI
Boom
(Expert
Systems
2nd AI Winter
(1987-1994)
AI
Boom
(Deep Learning)
1976
Medical Device
Amendments to the FD&C
Act
(provide reasonable
assurance of the safety and
effectiveness of medical
devices)
1990
Safe Medical
Devices Act
(SMDA)
2002
Principles of
Software
Validation
2016
21st Century
Cures Act
2022
FDA Regulated
Clinical
Decision
Support
Software

Artificial Intelligence (AI)
vs
Machine Learning (ML)

Expert Systems (Rule Based, Binary, Branch Logic)

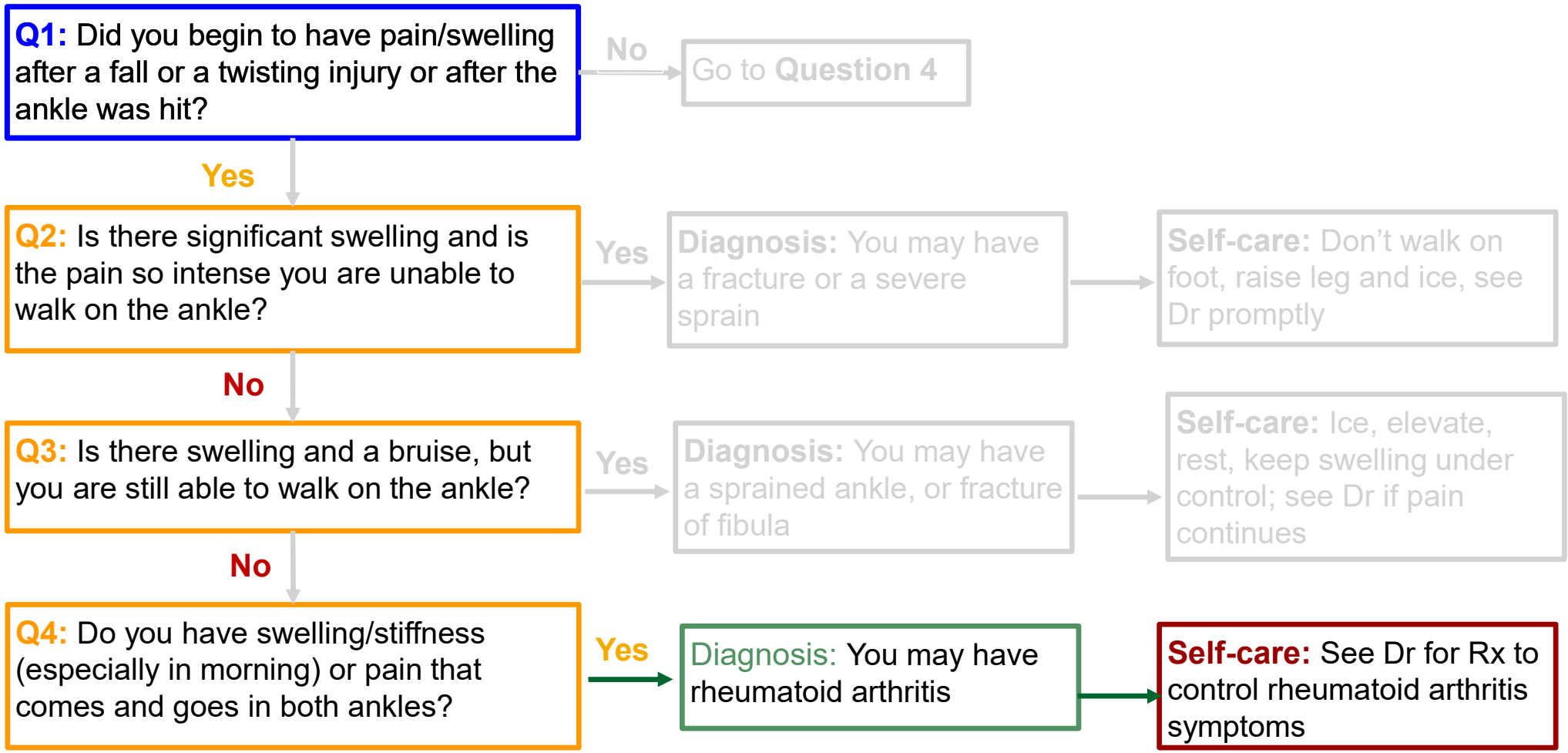


Expert Systems: Rule-Based, Logic Branching



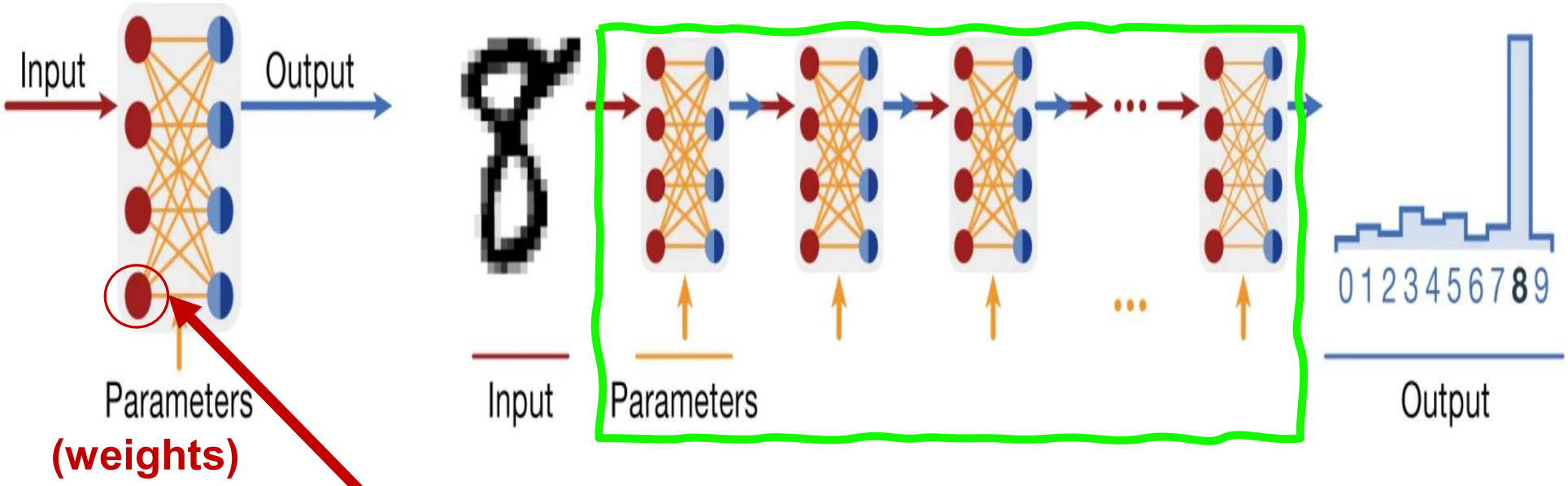
Qwaider et al (2017)

Expert Systems: Rule-Based, Logic Branching



Qwaider et al (2017)

Machine Learning: Statistical Modeling



(patterns that might be important…and what isn't)

2

WHEN IS IRB REVIEW REQUIRED

- ESTABLISHING COMMON DEFINITIONS

- QA VS RESEARCH

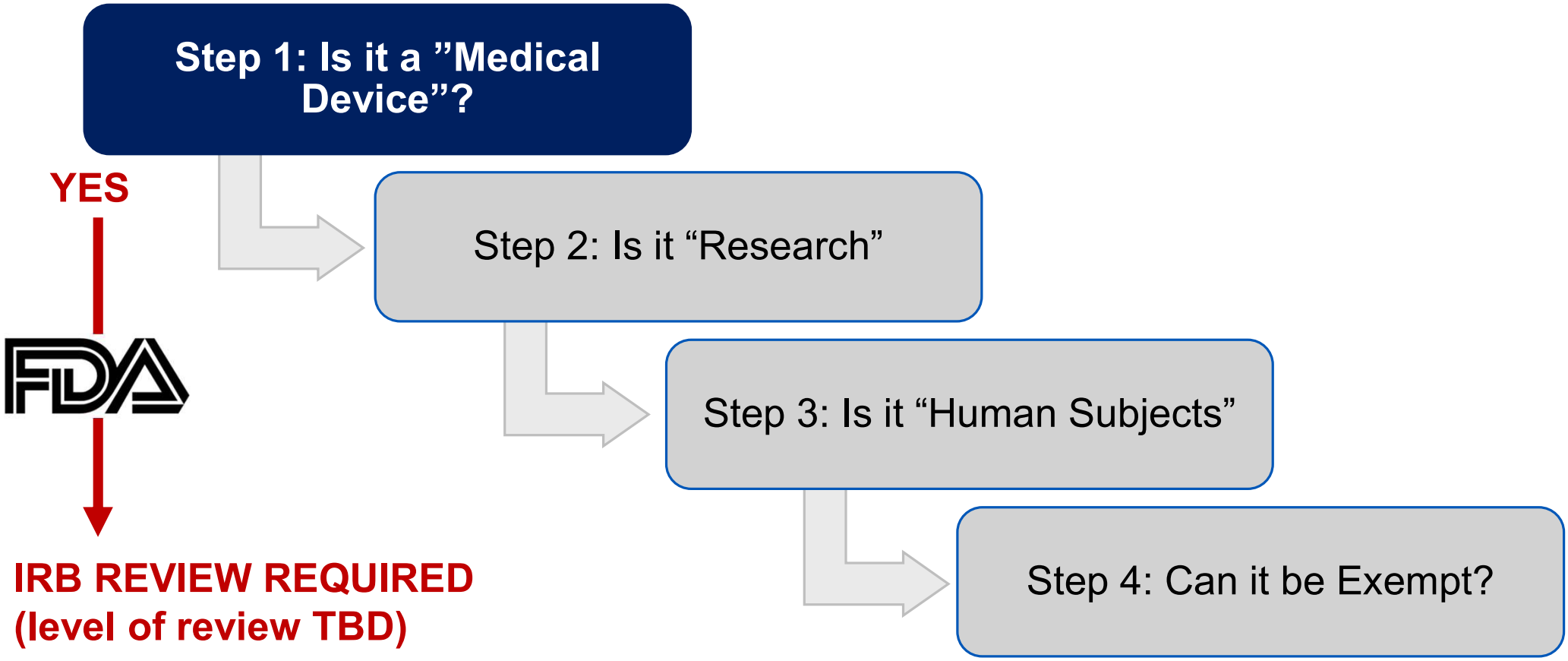
When is IRB Review Needed?
(21 CFR 56 & 45 CFR 46)

FDA (21 CFR 56)**:**
Clinical **Evaluations and Investigations** of devices
(Testing effectiveness of a model. *Including* Early Feasibility Studies of significant risk devices*)

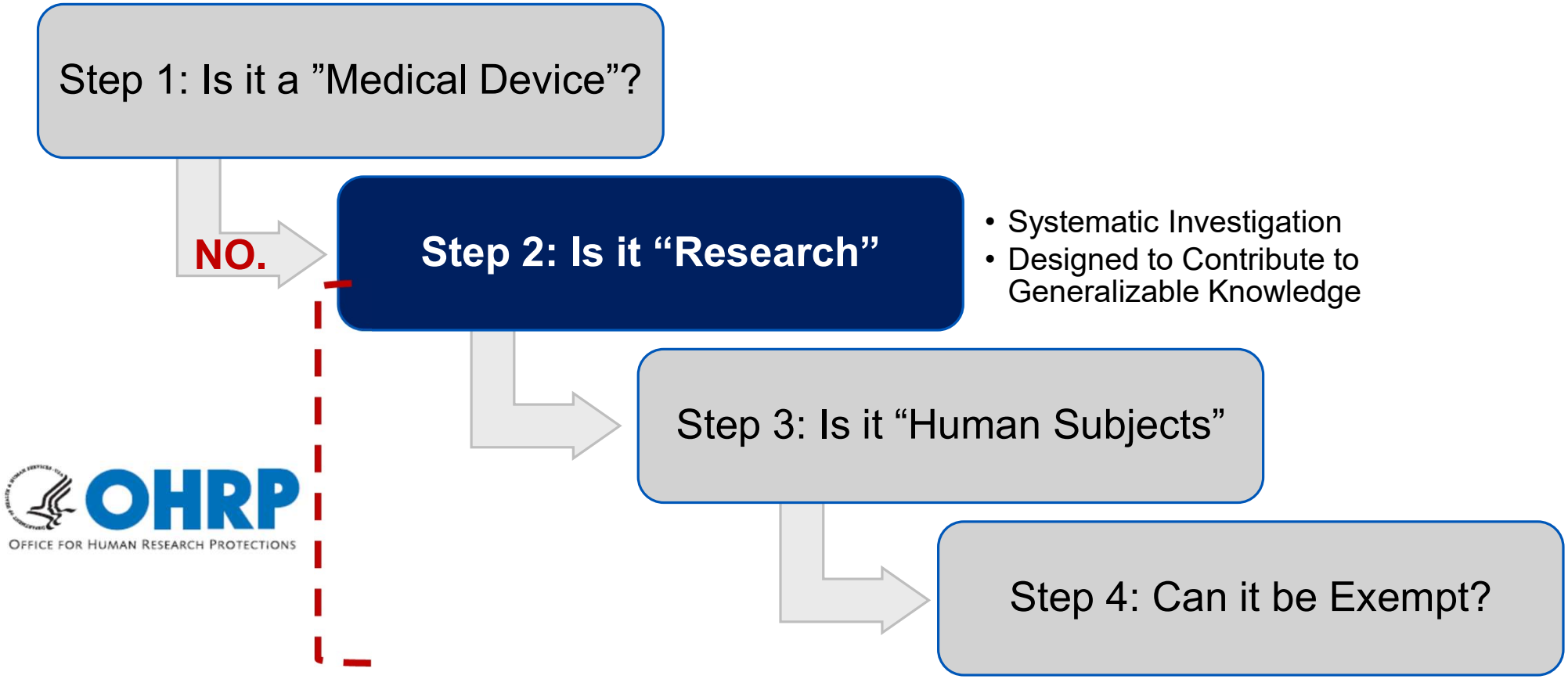
2. **Common Rule** (45 CFR 46)**:**
   Interaction/Intervention ***OR***
   Using / analyzing / generating identifiable information.

* Early Feasibility: A limited clinical investigation of a device early in development, typically before the device design has been finalized, for a specific indication. This information will further be used to determine necessary changes to ensure the safety and/or effectiveness of the model.

Determine What Regs Apply (4 steps)



Determine What Regs Apply (4 steps)



Definitions

AI in the Context of **H**uman **S**ubjects **R**esearch (AI HSR)

What is AI Human Subjects Research (AI HSR)?

AI HSR is:

"*Research*"

involving "*human subjects*",

conducted *to develop AI tools*.

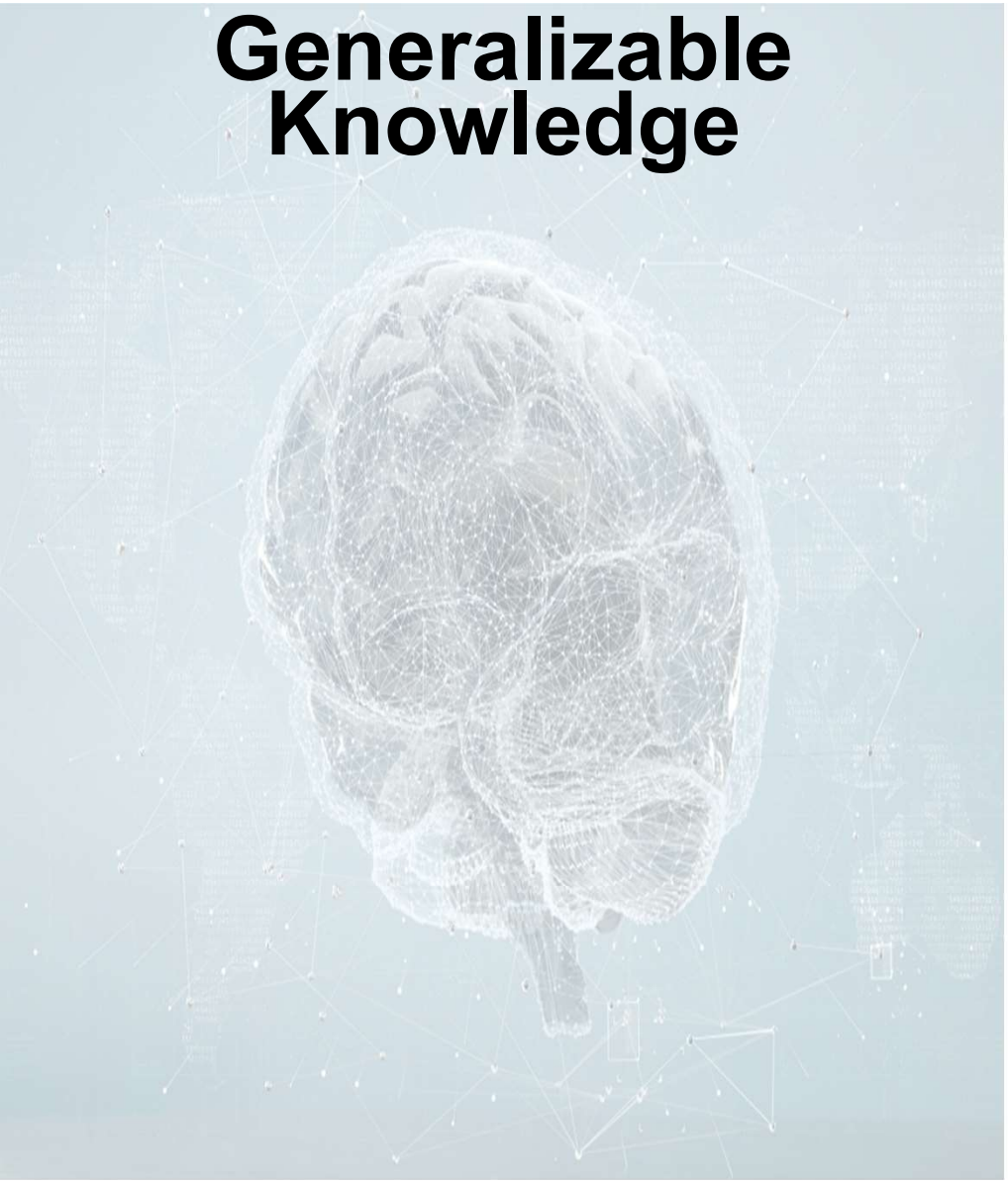
(Canca & Eto, 2020)

*"A **systematic Investigation*** (including development, testing, and/or evaluation) designed to develop or contribute to ***generalizable information*****

*** Systematic Investigation:**
"A detailed or careful examination that has or involves a prospectively identified approach to the activity based on a system, method, or plan"

-University of Washington

Generalizable Knowledge



***What is "Generalizable knowledge"**:*

- *The information is expected to expand the knowledge base of a scientific discipline or other scholarly field of study and yield one or both of the following:*

- Results that are applicable to a larger population beyond the site of data collection or the specific subjects studied

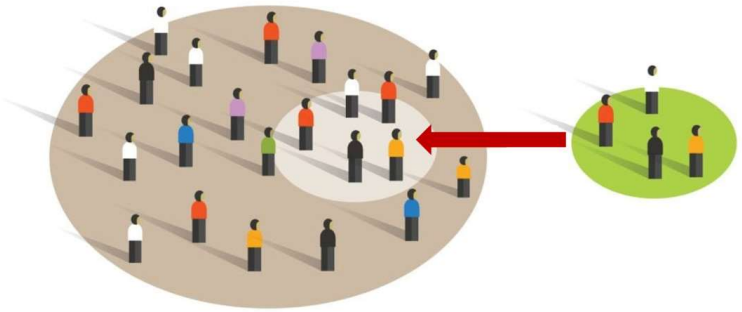
- Results that are intended to be used to develop, test, or support theories, principles, and statements of relationships, or to inform policy beyond the study.

  -University of Washington

Generalizable Knowledge and AI

***NOT* Generalizable AI:**

-*If the intended use of that algorithm is* ***limited to*** *its application to the original dataset.*



Generalizable AI:

-Intent *is to build a tool to be applied to a broader community* ***or*** *to* ***data not-yet-collected.***

-SACHRP (Oct 2022)

What is AI Human Subjects Research (AI HSR)?

AI HSR is:

"*Research*"

involving "***human subjects***",

conducted ***to develop AI tools***.

(Canca & Eto, 2020)

A Human Subject *is*

a ***living individual about whom*** *an investigator either…*

(i) **Obtains information or biospecimens** through ***intervention*** or ***interaction*** with the individual, *and stores,* uses, studies, or analyzes the information or biospecimens;

Or

(ii) **Obtains, stores, uses, studies, *analyzes*, or generates *identifiable private information*** or identifiable biospecimens.

What is AI Human Subjects Research (AI HSR)?

AI HSR is:

"*Research*"

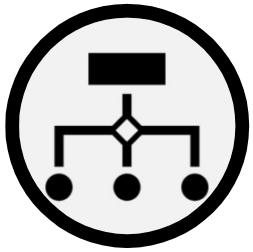
involving "***human subjects***",

conducted ***to develop AI tools***.

(Canca & Eto, 2020)

"To Develop AI Tools":

- The AI tool is under investigation
- Assessing AI tool performance, safety, or effectiveness
- AI tool needs validated
- Not currently legally marketed in US, ***or*** a legally marked device not being used as indicated



What About Quality Assurance or Quality Improvement Initiatives? (QA/QI)

(Projects NOT Subject to IRB Oversight)

NOTE: Still may require an official Determination at your institution

QUALITY ASSURANCE / QUALITY IMPROVEMENT (QA/QI) VS. RESEARCH

QA/QI Looks Like:

- ✓ Using models that are evidence based (SoC / non-investigational) (we know it works as intended, and is safe, and have scientific evidence to prove it)
- ✓ Limited to improving clinical workflows, health delivery, and quality (NOT improving health outcomes)
- ✓ Limited usefulness (to one's own clinic) (NOT for the field, your colleagues, or collaborators)
- ✓ Models developed by a licensed practitioner for their individual practice ONLY (not for hospital or colleague use) (FDA 2022)

Research Looks Like

- ✓ Comparing one model against another to assess performance or impact on health outcomes
- ✓ Determining efficacy of a model
- ✓ Developing, evaluating, validating a model
- ✓ Proving or answering a research question
- ✓ Randomizing or having control groups
- ✓ Models developed with the hopes of making it "generally available" (to the broader hospital or to other HCPs).
  - ✓ This triggers Sponsor-Investigator Requirements (FDA 2022)

WHEN PROJECTS MIGHT NOT QUALIFY AS QA/QI:

✓Has research components in it (see here for regulation)

✓Externally funded (NIH, industry, etc.)

✓Involves other sites

NOTE: *one should be careful not to call QI/QA projects "research", "investigation", or "a study" in their presentations or publications.*
Terminology matters!

3 CONDUCTING AN EFFECTIVE IRB REVIEW OF AI HSR - BELMONT REPORT (PART 1)

Respect For Persons (Transparency & Choice):

- **Autonomy:**
  - participation is voluntary;
  - informed consent;
  - **protection of privacy and confidentiality**;
  - right to withdraw without penalty; ***and***

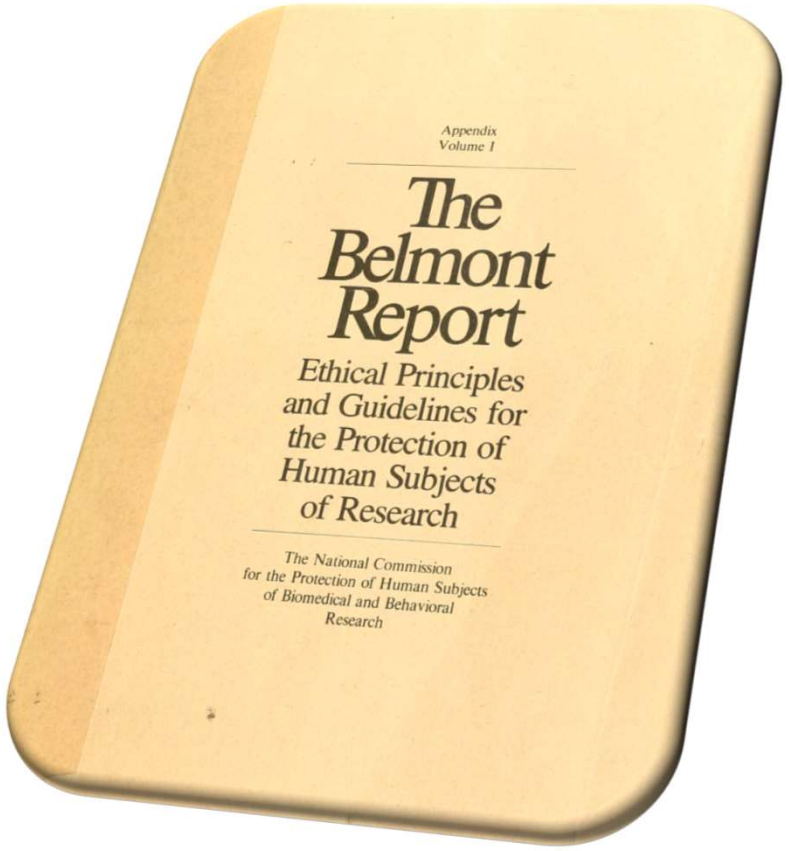
- Protect those with compromised autonomy

Justice (Equity)

- *No group bears the burden of testing (or being the test of) new technologies while other groups reap the rewards.*

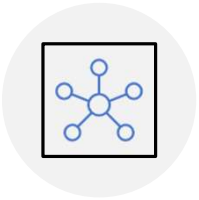
Beneficence (Don't hurt people)

- *Minimize harm, Maximize benefit.*
- *AI/ML projects demand the Responsible Conduct of Research*



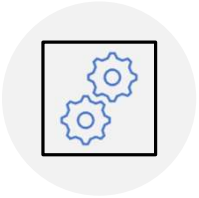
Principle of Respect for Persons

Is PHI or PII involved?
(Privacy & Confidentiality)



Will the proposed dataset(s) be combined?

* 3rd party ToS,
* use & storage, Previous
* Consent (secondary use),
* Contractual limitations from data source



* How did the technology come to the conclusion it did?
* Is the output interpretable to a lay person?

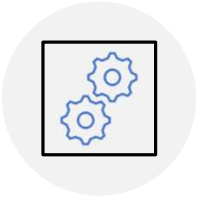
Is Study Team capable of answering participant questions about AI?

Principle of Respect for Persons

Is PHI or PII involved?

Will the proposed dataset(s) be combined?
(Privacy & Confidentiality)

* 3rd party ToS,
* use & storage, Previous
* Consent (secondary use),
* Contractual limitations from data source

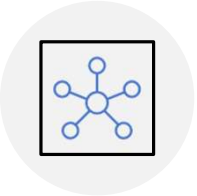


* How did the technology come to the conclusion it did?
* Is the output interpretable to a lay person?

Is Study Team capable of answering participant questions about AI?

Principle of Respect for Persons

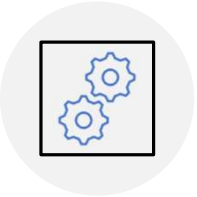
Is PHI or PII involved?



Will the proposed dataset(s) be combined?

* 3rd party Terms of Use,
* Consent for Future Use
* Long term storage/retention,
* Contractual limitations from data/model source

(Privacy & Confidentiality)

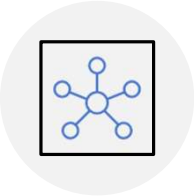


* How did the technology come to the conclusion it did?
* Is the output interpretable to a lay person?

Is Study Team capable of answering participant questions about AI?

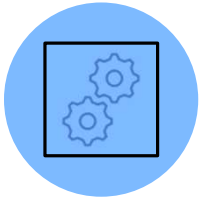
Principle of Respect for Persons

Is PHI or PII involved?



Will the proposed dataset(s) be combined?

* 3rd party ToS,
* use & storage,
* Previous consent (secondary use),
* Contractual limitations from data source



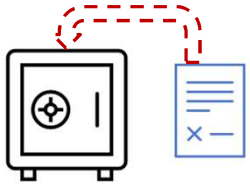
* How did the technology come
  to the conclusion it did?
* Is the output interpretable to
  a lay person?
(Informed Consent)

Is Study Team capable of answering participant questions about AI?
(Informed Consent)

Principle of Respect for Persons

- **Informed Consent:**
  - **Transparency:**
    - Informed about the investigational nature and role AI plays in the study
    - Informed if they have a choice/alternative (**AI Bill of Rights**).

- **Explainability/Human Interpretability:**
  - How the model functions/process;
  - Role of model's output in final decision-making are clearly explained;
  - Consent form is comprehensible to participants



- **Privacy & Confidentiality:**
  - **Data Disposition:**
    - What will happen to the data when the project ends? Will the model continue using the data for future training? Will model be shared? With whom?

Protocol Should Describe…

Principle of Justice

- **Representativeness:**
  - **Data Source:** source and characteristics of the data;
    - If external datasets will be combined (pooled);
  - **Diversity** (or lack of) in data
    - Justification for how data meets needs of study design
    - Procedures to ensure equitable selection (not just a race issue)
    - Target population of deployment match source data

- **Minimization of Disparities:**
  - Plan to mitigate algorithmic discriminatory decisions & unjust impacts
  - **Plan for pre-real-world deployment needs:**
    - External validation
    - Model re-calibration

- **Secondary Participants/Incidental Participant:**
  - Features of data used in final model
  - If collecting specific traits/individuals so that AI can learn how to single out the “noise” or “silence” that group out? (controls, non-cancer, offender vs non-offender)?

Additional info I need to assess this…

Principle of Justice

WHO …is directly (and indirectly) benefiting from this technology?

WHO's …data was used to train & validate the model

HOW …will these findings/technology benefit the *data-origin populations*?

HOW …will these findings/technology benefit the target *deployment populations*?

-Is benefit limited to specific population or setting? If so, why?

How to Mitigate Risk…

- *Belmont Report*
- ***ICH E6/GCP*** *(Declaration of Helsinki (WMA, 2008))*
- *45 CFR 46.111(A)(1)(i)*
- *Nuremberg Code (1947)*
- ***US (OSTP) AI Bill of Rights (2022) !NEW!***
- ***Executive Order 14110 (2023) !NEW!***

Principle of Beneficence

Evaluate study design:

Risks are minimized & Benefits are maximized

1. *Evaluate **quality of the science** in a research proposal based on **thorough knowledge of scientific literature**, etc.*
2. ***Qualifications of the Investigator:** PI's experience with AI/ML*
3. ***Resources** available to accomplish the study as planned*
4. ***Methods** used in study relative to available alternatives*
5. ***Characteristics** of the control group*
6. ***Statistical power** calculations*
7. ***Conflicts of Interest** (COI) are managed*

How to Mitigate Risk…

(2 Types of Risk)

Principle of Beneficence

TYPE 1 RISK: Privacy & Confidentiality

- **HIPAA Minimum Necessary:**
  - Don’t grab what you don’t need.
  - **Remember:** Deep Learning “needs” EVERYTHING.

- **External Disclosures:**
  - Does the training/validation data transfer with the model?
  - Is “derivative” data considered in external disclosures?
  - **Remember:** Inviting external collaborators into your firewall to access PHI is still a disclosure.

How to Mitigate Risk…

(2 Types of Risk)

Principle of Beneficence

TYPE 2 RISK: Direct Patient/Participant Risk

- Get ready for a high-maintenance relationship & long-term commitment!
  - Continuous monitoring
  - Post-Monitoring for true outcomes

- Future data usage, storage, and sharing for iterative changes/updates.
  - Who will do that?
  - Does the institution have the funds and FTE for required computational power, proper/safe upkeep?

4

CONDUCTING AN EFFECTIVE IRB REVIEW OF AI HSR (PART 2)

- THE .111 DETERMINATIONS

Navigating
& IRB Approval Criteria
via
“The .111 Determinations”

IRB Approval Criteria : “The .111 Determinations”

#1 & 2: Risks are minimized & reasonable in relation to benefits **(BENEFICENCE)**

#3: Subject selection is equitable **(JUSTICE)**

#4 & 5: Informed consent will be (a) sought and documented, or (b) waived as appropriate **(RESPECT FOR PERSONS)**

#6 & 7: Adequate provision are made for monitoring the data collected to (*protect privacy, maintain confidentiality and*) ensure the safety of subjects **(BENEFICENCE & JUSTICE)**

#8: Safeguards to protect rights and welfare of vulnerable subjects **(BENEFICENCE)**

Criteria 1 & 2 – Evaluating the Risk-to-Benefit Ratio

Belmont Report: Principle of Beneficence

RISKS- To Individual

- Privacy and confidentiality breach
- Harm from false positive or negative results
- Harm from future mis-application of the tool
- Dignitary harm from involvement w/o consent (learning post-hoc of data being used)

RISKS- To Group/Society

- Inappropriate or biased output
- Future misuse to stigmatize
- Inappropriate purpose

BENEFITS- To Individuals

- None

BENEFITS – To Society

- How can we know if there is "POTENTIAL" benefit without *evaluating* ***quality of the science*** *in a research proposal?*
  - ***Drugs studies have animal studies and other scientific evidence.***
  - ***What is available for AI/ML studies? Is it relevant?***

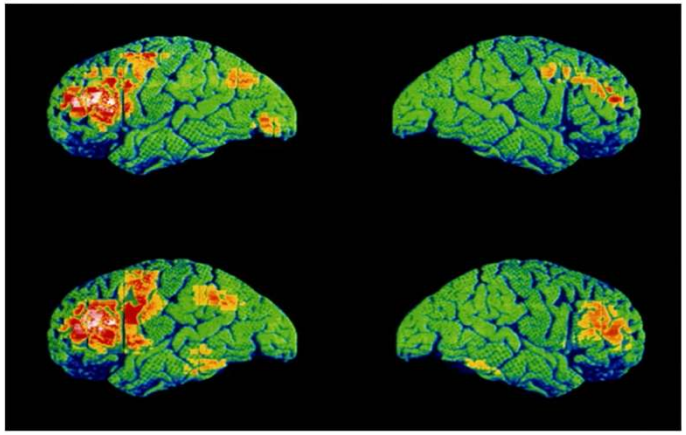
(Reflected in Executive Order 2023 and AI Bill of Rights)

NEWS | 11 January 2024

Medical AI falters when assessing patients it hasn't seen

Physicians rely on algorithms for personalized medicine – but an analysis of schizophrenia trials shows that the tools fail to adapt to new data sets.

By Miryam Naddaf



Scans showing brain activity during speech for a person with schizophrenia (bottom) and one without (top). Credit: Wellcome Centre Human Neuroimaging/Science Photo Library

RESEARCH ARTICLE

NEUROSCIENCE

Illusory generalizability of clinical prediction models

Adam M. Chekroud[1,2]*, Matt Hawrilenko[1], Hieronimus Loho[2], Julia Bondar[1], Ralitza Gueorguieva[3], Alkomiet Hasan[4], Joseph Kambeitz[5], Philip R. Corlett[2], Nikolaos Koutsouleris[6], Harlan M. Krumholz[7], John H. Krystal[2], Martin Paulus[8]

It is widely hoped that statistical models can improve decision-making related to medical treatments. Because of the cost and scarcity of medical outcomes data, this hope is typically based on investigators observing a model's success in one or two datasets or clinical contexts. We scrutinized this optimism by examining how well a machine learning model performed across several independent clinical trials of antipsychotic medication for schizophrenia. Models predicted patient outcomes with high accuracy within the trial in which the model was developed but performed no better than chance when applied out-of-sample. Pooling data across trials to predict outcomes in the trial left out did not improve predictions. These results suggest that models predicting treatment outcomes in schizophrenia are highly context-dependent and may have limited generalizability.

One fundamental problem in medicine is that despite similar treatments some patients get better whereas others show no improvement. One goal of precision medicine is to use machine learning to find models that will help predict who will respond to what type of treatment (*1*). For precision medicine to affect clinical practice and improve outcomes, the models that we develop must robustly predict outcomes for unseen, future patients (*2*–*5*).

However, models are not usually tested on new patients in a different context because data—especially data from controlled designs—are scarce and expensive (*6*). Instead, researchers typically split a study's participants into two or more random groups, build a model using the data from one of the groups, and test its predictions on the other group (e.g., k-fold cross-validation) (*3*, *4*). When we use this kind of approximation based on one data set or clinical sample, we have a fundamentally limited insight into the true potential for a model to improve outcomes in the future. Validating clinical prediction models in different clinical samples is an essential step in the model development process. It generally results in predictive performance measures that are lower but allows for a more realistic assessment of the potential for statistical models to improve clinical practice (*7*–*9*).

Open data opens possibilities

As efforts toward mandatory randomized controlled trial (RCT) data deposition, archival data sharing, and open science continue to advance, opportunities arise to more rigorously examine how well treatment prediction models will fare in different contexts. The Yale Open Data Access (YODA) Project is one such effort, which now includes a data archive of over 246 clinical trials from all medical fields.

The YODA project included several RCTs evaluating the comparative efficacy of antipsychotic medications for treating schizophrenia. Predicting treatment outcomes in schizophrenia could be especially advantageous because the clinical response to pharmacological interventions is heterogeneous and depends on many environmental factors such as individual and family-related stress, drug abuse, homelessness, and social isolation. Depending on the clinical outcome definition, up to 20 to 30% of first-episode individuals (*10*) and more than 50% with a relapse do not respond sufficiently to antipsychotic medications (*11*).

We examined the generalizability of clinical prediction models across multiple clinical trials using the case study of antipsychotic treatments for schizophrenia. Critically, this study directly evaluated the performance of a model on its initial training sample as well as how the same model performed on truly independent clinical trial samples. This allowed us to assess two key risks: First, models may "overfit" the data by fitting the random noise of one particular dataset rather than a true signal likely to generalize across samples, leading to good predictions in the training data that do not generalize to the testing data. The second key risk is poor model transportability. Models may lack external validity due to patients, providers, or implementation characteristics varying across trials (*12*).

Data sources

We used treatment data from five international, multisite RCTs (NCT00518323, NCT00334126, NCT00085748, NCT00078039, and NCT00083668) obtained through the YODA Project (https://yoda.yale.edu/). These trials were selected because of their comparability and consistency. All patients had a current DSM-IV diagnosis of schizophrenia at the start of the trial; all trials randomized patients to an antipsychotic medication or placebo; all trials used the same scale to measure treatment outcomes (the Positive and Negative Syndrome Scale, PANSS); all trials included a 4-week timepoint to measure outcomes; and all trials collected similar data about the patients at baseline. Combined, the trials also provide a heterogeneous patient

[1]Spring Health, New York City, NY 10010, USA. [2]Department of Psychiatry, Yale University School of Medicine, New Haven, CT 06520, USA. [3]Department of Biostatistics, Yale University, New Haven, CT 06520, USA. [4]Department of Psychiatry, Psychotherapy and Psychosomatics, University Augsburg, 86159 Augsburg, Germany. [5]Department of Psychiatry and Psychotherapy, University of Cologne, Faculty of Medicine and University Hospital of Cologne, Cologne, Germany. [6]Department of Psychiatry and Psychotherapy, Ludwig-Maximilians-University, Munich, Germany. [7]Center for Outcomes Research and Evaluation, Yale New Haven Hospital, New Haven, CT 06520, USA. [8]Laureate Institute for Brain Research, Tulsa, OK 74136, USA.
*Corresponding author. Email: adam.chekroud@yale.edu

Table 1. Treatment outcomes across trials.

Outcome definition	Adults first episode (*n* = 321)	Adults - Chronic #1 (*n* = 430)	Adults - Chronic #2 (*n* = 481)	Older adults (*n* = 99)	Teens (*n* = 182)	Total (*n* = 1513)
25% Reduction PANSS	264 (82.2%)	208 (48.4%)	266 (55.3%)	32 (32.3%)	47 (25.8%)	816 (54.0%)
50% Reduction PANSS	119 (37.1%)	85 (19.8%)	82 (17.0%)	7 (7.1%)	12 (6.6%)	306 (20.3%)
RSWG remission criteria	152 (47.4%)	129 (30.0%)	153 (31.8%)	24 (24.2%)	58 (31.9%)	517 (34.2%)
Percentage change in PANSS total score (SD)	-44.1 (23.1)	-26.9 (28.2)	-28.4 (25.3)	-18.0 (21.8)	-13.7 (21.5)	-28.8 (26.7)
Baseline total PANSS (SD)	103.0 (14.3)	92.4 (13.0)	92.9 (10.9)	91.1 (8.8)	90.0 (13.1)	94.4 (13.2)

Editor's summary

A central promise of artificial intelligence (AI) in healthcare is that large datasets can be mined to predict and identify the best course of care for future patients. Unfortunately, we do not know how these models would perform on new patients because they are rarely tested prospectively on truly independent patient samples. Chekroud *et al*. showed that machine learning models routinely achieve perfect performance in one dataset even when that dataset is a large international multi-site clinical trial (see the Perspective by Petzschner). However, when that exact model was tested in truly independent clinical trials, performance fell to chance levels. Even when building what should be a more robust model by aggregating across a group of similar multisite trials, subsequent predictive performance remained poor. —Peter Stern

https://www.science.org/doi/10.1126/science.adg8538

Criteria 3 and 3(b): *Equitable Selection & Vulnerability*

Belmont Report:
Principle of Justice
+
Executive Order 14110 (2023)

(keep AI algorithms from exacerbating discrimination)

Protocol Describes Plan For…:

- **Equity:**
  - **Equitable selection:** those impacted by the findings should be included
- **Stigmatization:**
  - Consider minority groups/communities that will be impacted by findings.
- **Diversity:**
  - Ensure large and diverse datasets reflect the target deployment population.
- **Vulnerability:**
  - Avoid unnecessary **inclusion/exclusion** of certain groups (age, race, ethnicity, disability, gender, etc.) due to inconvenience or unavailability.

Criteria # 4 & 5:
Informed Consent

Belmont Report:

Respect for Persons

Protocol Confirms…:

HOW they are authorized for "secondary use"

- Was consent obtained in the past for future use in this manner?

Compliance with Any State Laws and Within Limitations

- Do you need to consider international or state laws re: the use of that data/images?
  - Cause of Death/National Death Index may have limitations
  - Extra protections for HIV, psych/mental health data, pregnancy data, or incriminating data, etc.

Strong Justification that Meets Waiver Criteria (if Requesting)

- Is a HIPAA and/or Consent waiver needed and appropriate?
- **Consent Required For:**
  - Survey/Interaction.
  - Taking/linking data from other restricted sources.
  - Testing and Validation as Primary Data Collection.
  - Application to patient clinical care or decision-making.
  - SACHRP: consent required if data collection is part of the research (primary data collection).

Criteria # 4 & 5:
Informed Consent

Belmont Report:
Respect for Persons

Can we obtain informed consent if we, ourselves, are not informed?

To be "informed": *having or showing a lot of knowledge about a particular subject or situation*

Protocol Describes Plan For…:

✓ **Explainability / Human Interpretability**
- How the output is presented as understandable to the operator/reader
- If the output will "drive" or "inform" clinical decision-making

✓ **Transparency: (see explainability above)**
- **AI Bill of Rights (2022)**
  - Is AI involved in a decision made about my healthcare?
- **AI Executive Order (2023)**
  - Requirement to share safety test results and other critical information with US Govt
  - Govt recv reports and act to remedy unsafe practices involving AI

Criteria #6: *Data Monitoring*

Belmont Report:

Respect for Persons
&
Principle of Justice

The research plan makes adequate provision for monitoring the data collected to ensure the safety of subjects.

What does this require for AI/ML?

Model iteration, data shift, and version changes

Post-deployment monitoring to identify possible harms

Scientifically established AI/ML-specific methodology for mitigating bias spelled out (and according to best practice)

What kind of problems could be anticipated?

Are they thoroughly described?

How are they handled?

Criteria 7:
Privacy & Confidentiality

Belmont Report:

Principle of Beneficence
&
Principle of Justice

Protocol Describes Plan For…:

Privacy:

Control over the extent, timing, and circumstances of sharing oneself (physically, behaviorally, or intellectually) with others; (OHRP, 1993)

- ✓ Adherence to HIPAA (Security Act, HITECH, Privacy Act)
- ✓ What PII/PHI will be used and by WHOM
  - ✓ If it involves Limited Datasets, acknowledge PHI
- ✓ If HIPAA does not apply, HOW is "private" identifiable data determined?
- ✓ Additional protections if involving small populations (increased risk of re-identifiability)
- ✓ Confirming compliance with authorization and ToU when using Public Datasets, Big Data, & linking through common identifiers (See Google/University of Chicago Case)
- ✓ Extra protections for Sensitive Data (Substance Use, Mental Health, Police Records, HIV, etc.)

Criteria 7:
Privacy & Confidentiality

Belmont Report:

Principle of Beneficence
&
Principle of Justice

Protocol Describes Plan For…:

Confidentiality:
Treatment of information that an individual has disclosed in a relationship of trust and with the expectation that it will NOT be divulged to others in ways that are inconsistent with the understanding of the original disclosure without permission (ORHP, 1993)

- ✓ How confidentiality of datasets are maintained
- ✓ Mitigation if confidentiality is breached
- ✓ How re-identifiability is minimized
- ✓ Adequate de-identification method for biometric identifiers
  - ✓ **Example:** Video, Audio, Gait, Retina scans
- ✓ Consent process, as required (state-based laws), for use of biometric data
- ✓ How external or internal sets will be pooled/combined, and confidentiality maintained
- ✓ BAAs for third party vendors; congruence with authorization.

OTHER REGULATORY CONSIDERATIONS

FDA (21 CFR)

- What is “FDA-regulated”
- Clinical Decision Support System Exceptions
- Making Device Risk Determinations
- SaMD Action Plan
- GMLP
- Software Validation / QSM (21 CFR §820.30)
- Performance Assessment of Quantitative Imaging
- Using the IRB as an FDA-surrogate
- 21 CFR Part 11, SBOM, and PATCH Act

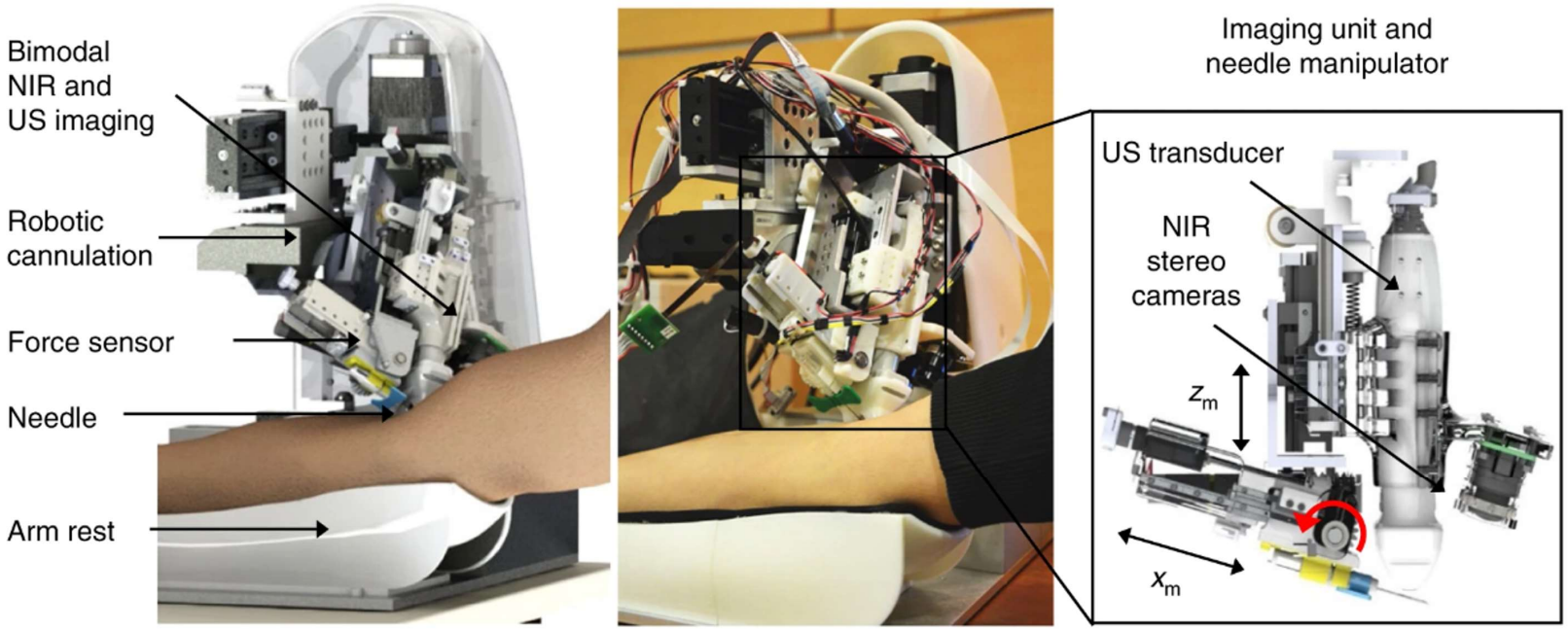
Others

- HIPAA Privacy Rule
- HIPAA Security Rule
- HITECH Act
- 42 CFR Part 2
- **FTC** Breach Notification Act, FTC Act, FCRA, & ECOA; Model as Service Privacy Laws
- State (patchwork) Laws (25 states to-date!)
- Cause Of Death (NDI & State) Limitations
- NIST RMF (Ntl’ Inst for Standards & Tech)
- AI Bill of Rights (Blueprint)(2022)
- AI Executive Order 14110 (2023)
- Pending: *Algorithmic Accountability Act, etc.*

5 FDA CONSIDERATIONS - SOFTWARE AS A MEDICAL DEVICE (DEFINITIONS + VALIDATION/TESTING AI/ML SYSTEMS+ RISK DETERMINATIONS) (PART 1)

"Software as a Medical Device"

Software intended to be used for one or more medical purposes that perform these purposes without being part of a hardware medical device.



Source: Chen, A.I., Balter, M.L., Maguire, T.J. et al. Deep learning robotic guidance for autonomous vascular access. Nat Mach Intell 2, 104–115 (2020)

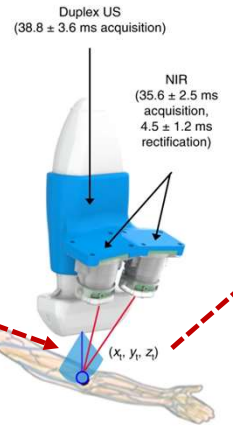
"Medical purpose"

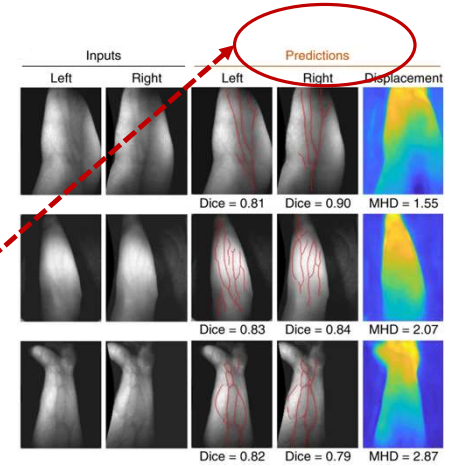
Examples:

- Diagnosis, prevention, monitoring, treatment or alleviation of disease
  - analysis of clinical samples that help with disease diagnosis.
  - helps monitor sleep apnea using the microphone of a smart device to detect breathing patterns.
  - Use of data from individuals for predicting risk score for developing stroke or heart disease for creating prevention or interventional strategies.
- Disease management
  - Provides info by taking pictures (ex: for monitoring or supplementing other info) for disease monitoring.

Examples:

- *Breast Cancer Prediction Score*
- *Sepsis Prediction*
- *Stroke Prediction*
- *Suicidality Prediction*
- *Schizophrenia treatment success prediction*
- *Treatment Effectiveness Prediction*





https://www.imdrf.org/sites/default/files/docs/imdrf/final/technical/imdrf-tech-131209-samd-key-definitions-140901.pdf

Validating & Testing an AI/ML SaMD

Clinical Evaluations ≠ Clinical Investigations

But…

***Both** Clinical Evaluations & Investigations Require IRB Review*

CLINICAL *INVESTIGATION* VS CLINICAL *EVALUATION*

Clinical Investigation

- Not always necessary (e.g., if device qualifies for a 510(k))
- Clinical Trial (interventions)
- **Research Question:** what works and doesn't work in treating humans
- Establish safety, device performance, benefits, effectiveness
- Standards:
  - ISO 14155 Standard
  - QSM (Design control, etc.)
- Final step of R&D process

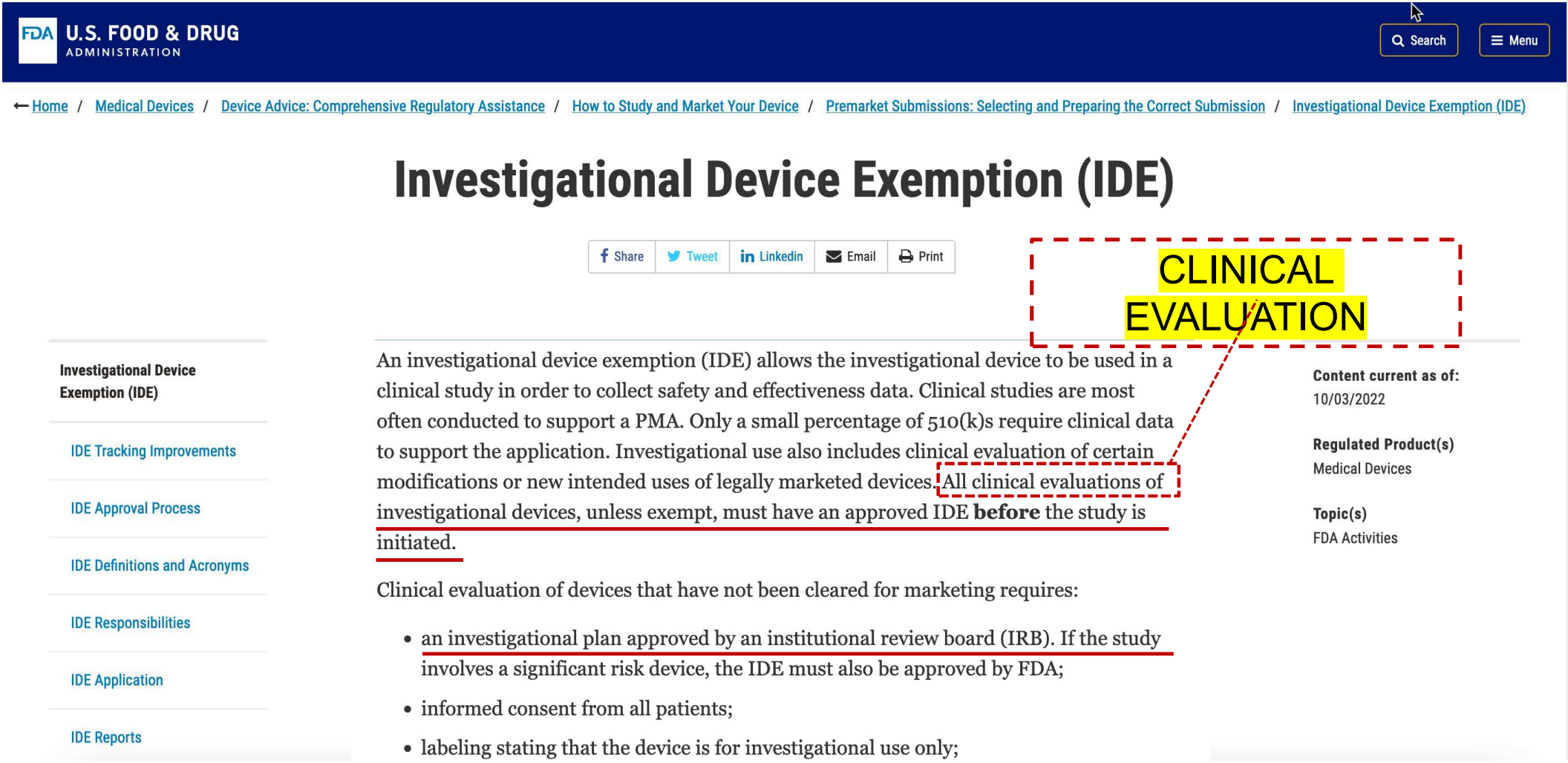
Clinical Evaluation

- **ALWAYS** necessary
- Product development (lit review, analysis of available data)
- Non-interventional assessment of existing data
- **Research Question:** Can the medical device achieve its intended purpose
- Establish safety, benefits outweigh risk, if any predicate devices
- Continuously monitored and updated over time (post-market surveillance)

https://www.fda.gov/media/100714/download

https://www.raps.org/news-and-articles/news-articles/2022/3/clinical-evaluation-of-software

FDA's Approach to Investigational Devices



Unpacking "*Clinical Evaluation*" of SaMD

Software as a Medical Device (SAMD): Clinical Evaluation

Guidance for Industry and Food and Drug Administration Staff

Document issued on December 8, 2017.

7.0 SaMD Clinical Evaluation

Clinical evaluation is a systematic and planned process to **continuously generate, collect, analyze, and assess** the **clinical data** pertaining to a SaMD in order to **generate clinical evidence verifying the clinical association and the performance metrics** of a SaMD when used as intended by the manufacturer. The quality and breadth of the clinical evaluation is determined by the role of the SaMD for the target clinical condition and assures that the **output of the SaMD** is clinically valid and **can be used reliably and predictably**.

https://www.fda.gov/media/100714/downloa

7.0 SaMD Clinical Evaluation

Clinical evaluation is a systematic and planned process to continuously generate, collect, analyze, and assess the ***clinical data*** *pertaining to a SaMD in order to generate* ***clinical evidence*** *verifying the clinical association and the performance metrics of a SaMD when used as intended by the manufacturer. The quality and breadth of the clinical evaluation is determined by the role of the SaMD for the target clinical condition and assures that the output of the SaMD is clinically valid and can be used reliably and predictably.*

- *safety and/or* ***performance information*** *that is*
- *generated from the use of the "device" (e.g., the AI system)*

(using EMR to validate and test the AI system)

- *the technical documentation of a medical device…*
- *along with other design verification and validation documentation,*
- *device description, labelling,*
- *risk analysis and*
- *manufacturing information…*

cGMP § 820.30 as per § 812

Software as a Medical Device (SaMD):

STEP 1

*Intended **PURPOSE***
(NOT current phase in IRB application)

WHAT IS BEING EVALUATED?

SaMD Definition Statement

- Intended Medical Purpose of a SaMD
  - Treat or Diagnose
  - Drive Clinical Management
  - Inform Clinical Management
- Targeted Healthcare Situation or Condition of a SaMD
  - Critical
  - Serious
  - Non-Serious

SaMD Categories

	Treat or Diagnose	Drive Clinical Mgmt	Inform Clinical Mgmt
Critical	IV	III	II
Serious	III	II	I
Non-Serious	II	I	I

SaMD N12 Risk Categorization Framework

SaMD N41 Clinical Evaluation

Clinical Association
between a SaMD output and a Clinical Condition

Literature searches, Original Clinical Research, Professional Society Guidelines, Secondary Data Analysis, Clinical Trials

Product Performance
Verify & Validate

Analytical / Technical Validation
Accuracy, Reliability, Precision…

Clinical Validation
Sensitivity, Specificity, Odds Ratio…

STEP 2

HOW IS IT BEING EVALUATED?

*Requirements, Design, Develop, **Verify & Validate**, Deploy, Maintain, Retire*
SaMD Realization and Use Processes

*Planning, **Risk Management**, Documentation, Configuration, Measurement, Outsourcing*
SaMD Lifecycle Support Processes

Personnel, Infrastructure, Work Environment
Leadership and Organizational Support

SaMD N23 Quality Management System

STEP 3

THROUGH WHAT STANDARDS & PROCESSES?

§ 820.30, ISO standards 14155, 42001, etc.

Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) Action Plan

January 2021

Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD)

1. Quality Systems and Good Machine Learning Practices (GMLP):

The FDA expects every medical device manufacturer to have an established quality system that is geared towards developing, delivering, and maintaining high-quality products throughout the lifecycle that conforms to the appropriate standards and regulations.[19] Similarly, for AI/ML-based SaMD, we expect that SaMD developers embrace the excellence principles of culture of quality and organizational excellence.[20]

As is the case for all SaMD, devices that rely on AI/ML are expected to demonstrate analytical and clinical validation, as described in the SaMD: Clinical Evaluation guidance (Figure 3).[21] The specific types of data necessary to assure safety and effectiveness during the premarket review, including study design, will depend on the function of the AI/ML, the risk it poses to users, and its intended use.

This is an AI/ML version of the standard QSM/cGMP (from 2017 guidance)

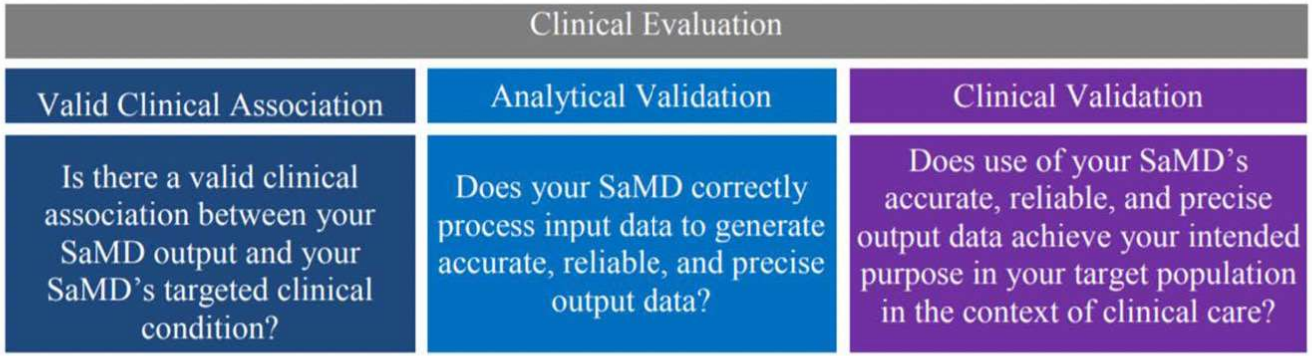


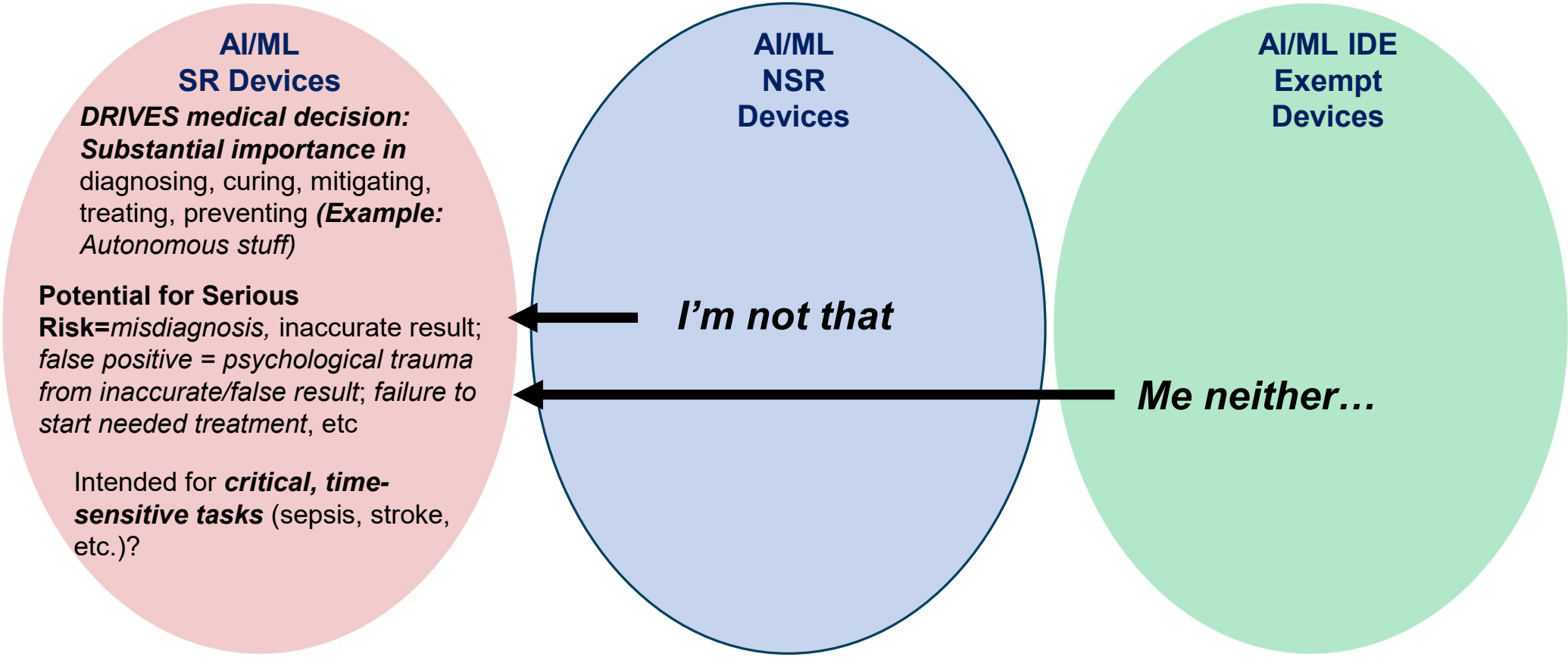
Figure 3: IMDRF description of Clinical Evaluation components

https://www.fda.gov/media/122535/download

Device Determinations & Assessing Device Risk

Is the AI SaMD SR? NSR? Or IDE Exempt?

§812.2(c)



Need IDE from FDA

Need NSR Determination from IRB or FDA

PI Justifies w Evidence / IRB Confirms

Is the AI SaMD SR? NSR? Or IDE Exempt?

§812.2(c)

AI/ML NSR Devices

…I am not used without confirmation by another FDA approved product.

I must be either an SR or NSR device….

…but which one?

Need SR/NSR Determination from IRB or FDA

"Another FDA-approved diagnostic or medically established procedure":

Is there one?? What is it?

Example:

Software ***function*** must enable HCPs to ***independently review*** the ***basis for the output*** so that they do not rely on the output (recommendations), but rather ***on their own judgment***, to ***make clinical decisions*** for individual patients.

AI/ML IDE Exempt Devices

- ✓ Non-invasive
- ✓ Does not require invasive procedure
- ✓ Does not introduce energy (laser, radiation, etc.) ***and***
- ✓ Not used as a diagnostic **without confirmation by another** FDA-approved diagnostic product or medically established procedure.

PI Justifies w Evidence / IRB Confirms

https://www.fda.gov/media/109618/download

Is the AI SaMD SR? NSR? Or IDE Exempt?

§812.2(c)

AI/ML NSR Device?

If not SR, and not IDE Exempt.

Need NSR Determination from IRB or FDA

No FDA approved Alternative Used

FDA approved Alternative Used

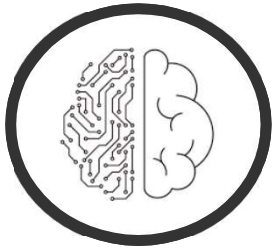
AI/ML IDE Exempt Devices

- ✓ Non-invasive
- ✓ Does not require invasive procedure
- ✓ Does not introduce energy (laser, radiation, etc.) ***and***
- ✓ Not used as a diagnostic **without confirmation by another** FDA-approved diagnostic product or medically established procedure.

PI Justifies w Evidence / IRB Confirms

6 FDA CONSIDERATIONS - SAMD: WHICH REGS APPLY? (PART 2)

What Regs Apply to My AI Medical Device?

Device Type	Applicable FDA Regulation
 **IDE-Exempt studies** *(Not requiring an IDE)*	21 CFR §50, 56, 809.10(c)(2), 820.30 & Part 11 Must meet 21st Century Cures Act Criteria **(2022)**. **NOTE:** Not eligible for Common Rule "Exempt 4" (45 CFR 46.104)
 **Non-Significant Risk (NSR)** *(If granted, is considered as having an IDE)*	21 CFR §50, 56, 820.30, + abbreviated 812 & Part 11 **NOTE:** Not eligible for Common Rule "Exempt" Cat. 4 (45 CFR 46.104); Possibly eligible for "Expedited" 2 or 9 (Requires Full Board review for determination)
 **Significant Risk** *(Studies requiring an IDE)* *e.g., AI-driven Brain Computer Interface (BCI)*	21 CFR §50, 56, 812, 820, & Part 11 *(and more)* (Full Board review)

But Aren't ALL Clinical Decision Support (CDS) Tools "EXEMPT" Under the Cures Act?

Not All CDS Tools Are Created Equal

https://www.fda.gov/media/109618/download

Your Clinical Decision Support Software: Is It a Device?

FDA

The FDA issued a guidance, Clinical Decision Support Software, to describe the FDA's regulatory approach to Clinical Decision Support (CDS) software functions. This graphic gives a general and summary overview of the guidance and is for illustrative purposes only. Consult the guidance for the complete discussion and examples. Other software functions that are not listed may also be device software functions. *

Your software function must meet all four criteria to be Non-Device CDS.

Summary interpretation of CDS criteria

1. Your software function does **NOT** acquire, process, or analyze medical images, signals, or patterns.
2. Your software function displays, analyzes, or prints medical information normally communicated between health care professionals (HCPs).
3. Your software function provides recommendations (information/options) to a HCP rather than provide a specific output or directive.
4. Your software function provides the basis of the recommendations so that the HCP does not rely primarily on any recommendations to make a decision.

Your software function may be non-device CDS.

Non-Device Examples

Non-Device examples display, analyze, or print the following examples of medical information, which must also not be images, signals, or patterns:

- Information whose relevance to a clinical decision is well understood
- A single discrete test result that is clinically meaningful
- Report from imaging study

AND

Non-Device examples provide:

- Lists of preventive, diagnostic, or treatment options
- Clinical guidelines matched to patient-specific medical info
- Relevant reference information about a disease or condition

AND

Non-Device examples provide:

- Plain language descriptions of the software purpose, medical input, underlying algorithm
- Relevant patient-specific information and other knowns/unknowns for consideration

Information is well understood; no images, signals, patterns, or continuous monitoring

Device Examples

Device examples acquire, process, or analyze:

- Signal acquisition systems
- In vitro diagnostics
- Magnetic resonance imaging (MRI)
- Next Generation Sequencing (NGS)
- Continuous Glucose Monitoring (CGM)
- Computer aided detection/diagnosis (CADe/CADx)

OR

Device examples display, analyze or print:

- Continuous signals/patterns
- Medical images
- Waveforms (ECG)
- More continuous sampling (aka – a signal or pattern)

OR

Device examples provide:

- Risk scores for disease or condition
- Probability of disease or condition
- Time-critical outputs

OR

Device examples:

- Basis of recommendations is not provided

Your software function is a device.

Your Clinical Decision Support Software: Is It a Device?

FDA

The FDA issued a guidance, Clinical Decision Support Software, to describe the FDA's regulatory approach to Clinical Decision Support (CDS) software functions. This graphic gives a general and summary overview of the guidance and is for illustrative purposes only. Consult the guidance for the complete discussion and examples. Other software functions that are not listed may also be device software functions. *

Your software function must meet all four criteria to be Non-Device CDS.

Summary interpretation of CDS criteria

1. Your software function does **NOT** acquire, process, or analyze medical images, signals, or patterns.
2. Your software function displays, analyzes, or prints medical information normally communicated between health care professionals (HCPs).
3. Your software function provides recommendations (information/options) to a HCP rather than provide a specific output or directive.
4. Your software function provides the basis of the recommendations so that the HCP does not rely primarily on any recommendations to make a decision.

Your software function may be non-device CDS.

Non-Device Examples

Non-Device examples display, analyze, or print the following examples of medical information, which must also not be images, signals, or patterns:

- Information whose relevance to a clinical decision is well understood
- A single discrete test result that is clinically meaningful
- Report from imaging study

AND

Non-Device examples provide:

- Lists of preventive, diagnostic, or treatment options
- Clinical guidelines matched to patient-specific medical info
- Relevant reference information about a disease or condition

AND

Non-Device examples provide:

- Plain language descriptions of the software purpose, medical input, underlying algorithm
- Relevant patient-specific information and other knowns/unknowns for consideration

Options; not specific output, risk score, or probability.

Device Examples

Device examples acquire, process, or analyze:

- Signal acquisition systems
- In vitro diagnostics
- Magnetic resonance imaging (MRI)
- Next Generation Sequencing (NGS)
- Continuous Glucose Monitoring (CGM)
- Computer aided detection/diagnosis (CADe/CADx)

OR

Device examples display, analyze or print:

- Continuous signals/patterns
- Medical images
- Waveforms (ECG)
- More continuous sampling (aka – a signal or pattern)

OR

Device examples provide:

- Risk scores for disease or condition
- Probability of disease or condition
- Time-critical outputs

e.g., Sepsis, Stroke, etc.

OR

Device examples:

- Basis of recommendations is not provided

Your software function is a device.

Your Clinical Decision Support Software: Is It a Device?

FDA

The FDA issued a guidance, Clinical Decision Support Software, to describe the FDA's regulatory approach to Clinical Decision Support (CDS) software functions. This graphic gives a general and summary overview of the guidance and is for illustrative purposes only. Consult the guidance for the complete discussion and examples. Other software functions that are not listed may also be device software functions. *

Your software function must meet all four criteria to be Non-Device CDS.

Summary interpretation of CDS criteria

1. Your software function does **NOT** acquire, process, or analyze medical images, signals, or patterns.
2. Your software function displays, analyzes, or prints medical information normally communicated between health care professionals (HCPs).
3. Your software function provides recommendations (information/options) to a HCP rather than provide a specific output or directive.
4. Your software function provides the basis of the recommendations so that the HCP does not rely primarily on any recommendations to make a decision.

Your software function may be non-device CDS.

Non-Device Examples

Non-Device examples display, analyze, or print the following examples of medical information, which must also not be images, signals, or patterns:

- Information whose relevance to a clinical decision is well understood
- A single discrete test result that is clinically meaningful
- Report from imaging study

AND

Non-Device examples provide:

- Lists of preventive, diagnostic, or treatment options
- Clinical guidelines matched to patient-specific medical info
- Relevant reference information about a disease or condition

AND

Non-Device examples provide:

- Plain language descriptions of the software purpose, medical input, underlying algorithm
- Relevant patient-specific information and other knowns/unknowns for consideration

Device Examples

Device examples acquire, process, or analyze:

- Signal acquisition systems
- In vitro diagnostics
- Magnetic resonance imaging (MRI)
- Next Generation Sequencing (NGS)
- Continuous Glucose Monitoring (CGM)
- Computer aided detection/diagnosis (CADe/CADx)

OR

Device examples display, analyze or print:

- Continuous signals/patterns
- Medical images
- Waveforms (ECG)
- More continuous sampling (aka – a signal or pattern)

OR

Device examples provide:

- Risk scores for disease or condition
- Probability of disease or condition
- Time-critical outputs

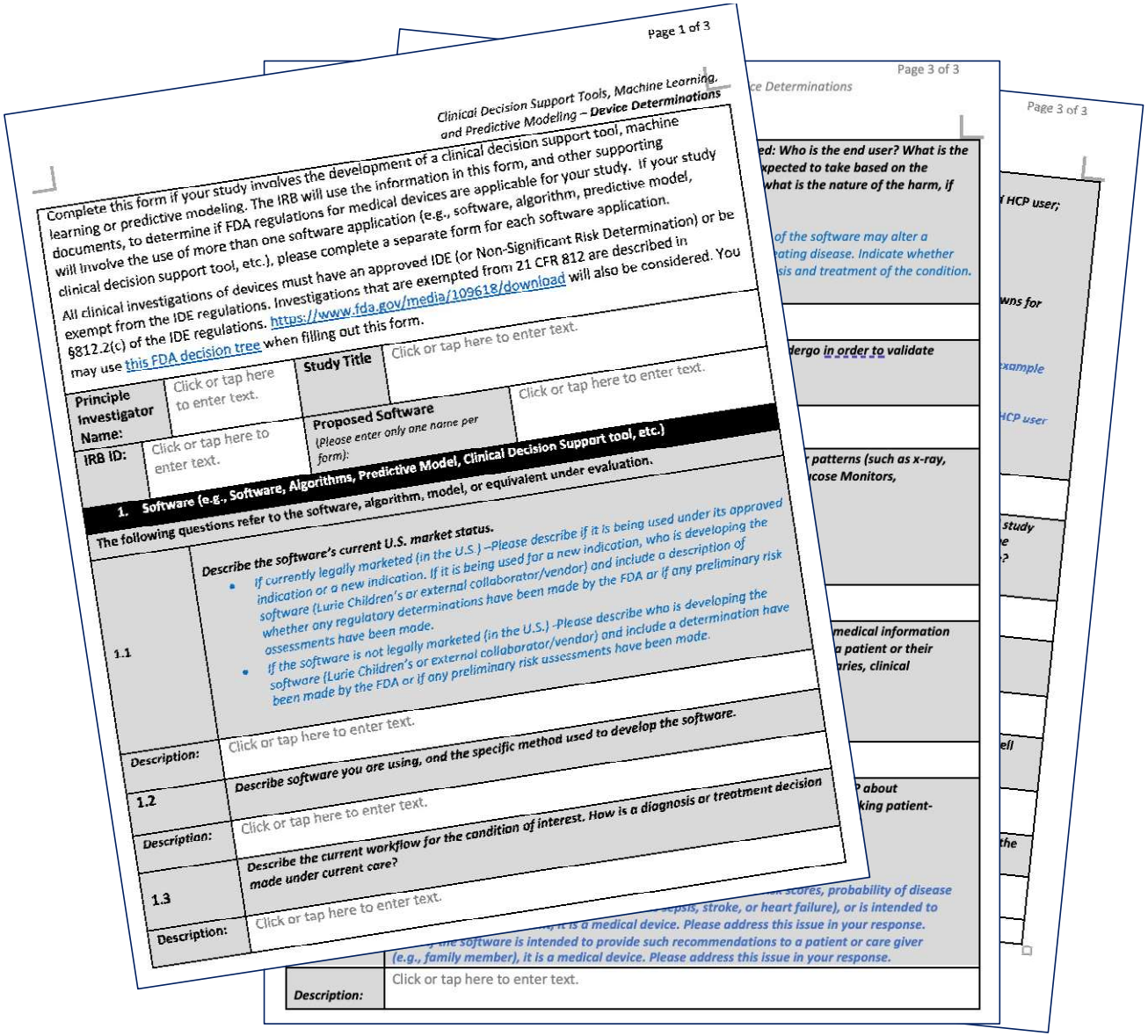
OR

Device examples:

- Basis of recommendations is not provided

Your software function is a device.

Basis of Recommendation MUST be provided



Is My Clinical Decision Support Tool a "Medical Device"?

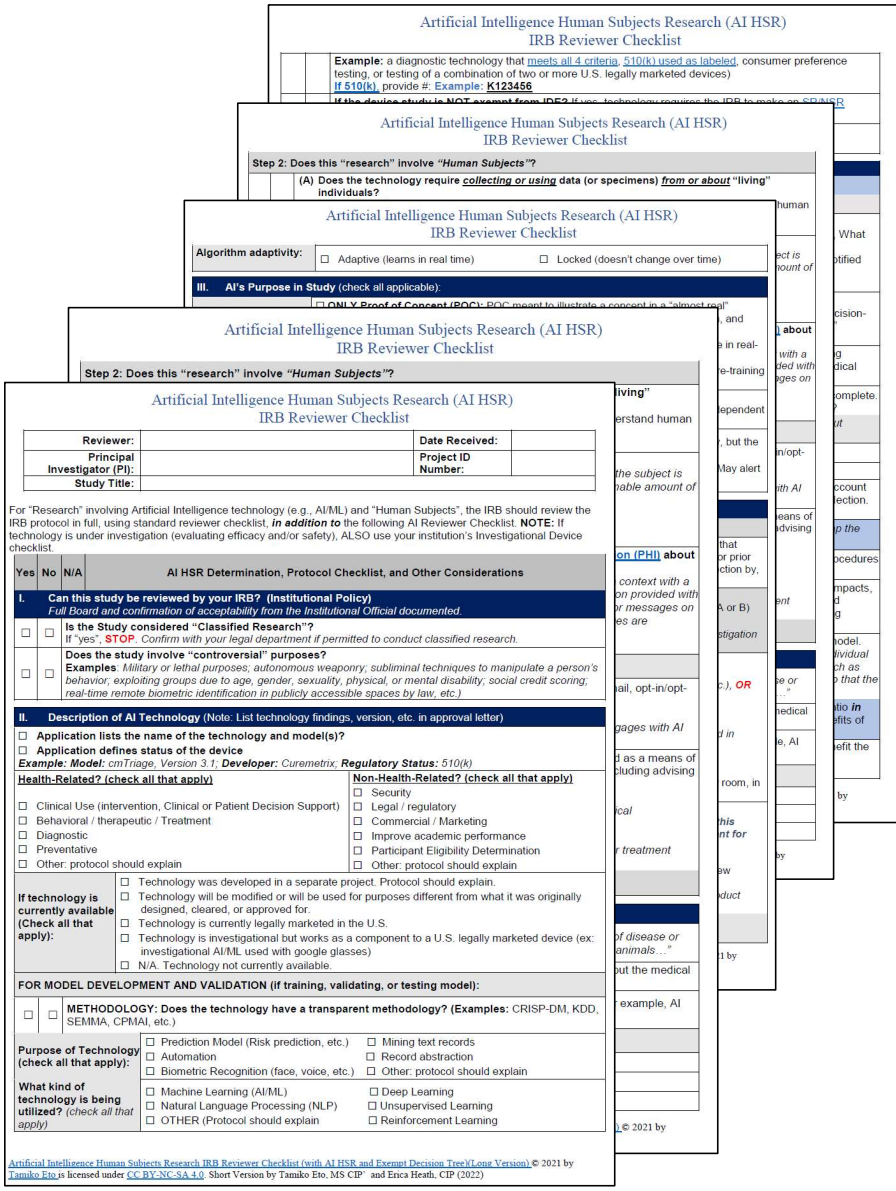
There's a CHECKLIST for that!
(email me)

IRB & HRPP CHECKLISTS...

Visit here to access the most recent/updated:

- ✓ AI HSR IRB Reviewer Checklist
- ✓ AI HSR Exempt Determination Decision Tree
- ✓ AI HSR Human Subjects Research Decision Tree

Learn how to use the AI HSR Checklist here (must be a **PRIM&R** member): https://www.pathlms.com/primr/courses/43595/documents/64223



RESPONSIBLE AI STARTS WITH US!

MAYO CLINIC
HEALTH SYSTEM

THANK YOU! LET'S CONNECT!

Tamiko Eto

Director:

Research Operations, HRPP & IRB

Mayo Clinic

Eto.Tamiko@Mayo.edu