

Technology, Big Data and Transparency

Ida Sim, MD, PhD
January 24, 2020

University of California San Francisco
Co-Founder Open mHealth
Co-Founder Vivli

 @idasim

Digital Transparency for Real-World Evidence

Ida Sim, MD, PhD
January 24, 2020

University of California San Francisco
Co-Founder Open mHealth
Co-Founder Vivli

 @idasim

Supported by: NIBIB R24 EB025845, NSF 1640813, IEEE P1752

Core Objective: Trustworthiness of Evidence

Transparency: necessary but not sufficient

- Do I have all the details? Do I trust the people who ran this study? Are they hiding anything?

(Computational) Reproducibility

- Do I trust the execution of the study? Do I trust the computational results?

Replicability

- Do I trust the findings are true beyond this one study?



The
Economist

MAY 6TH - 12TH 2017

Crunch time in France

Ten years on: banking after the crisis

South Korea's unfinished revolution

Biology, but without the cells

The world's most valuable resource

Data and the new rules
of competition





WIKIPEDIA
The Free Encyclopedia

[Main page](#)
[Contents](#)
[Featured content](#)
[Current events](#)
[Random article](#)
[Donate to Wikipedia](#)
[Wikipedia store](#)

Interaction

[Help](#)
[About Wikipedia](#)
[Community portal](#)
[Recent changes](#)

Article [Talk](#)

[Read](#) [Edit](#) [View history](#)

Data-driven

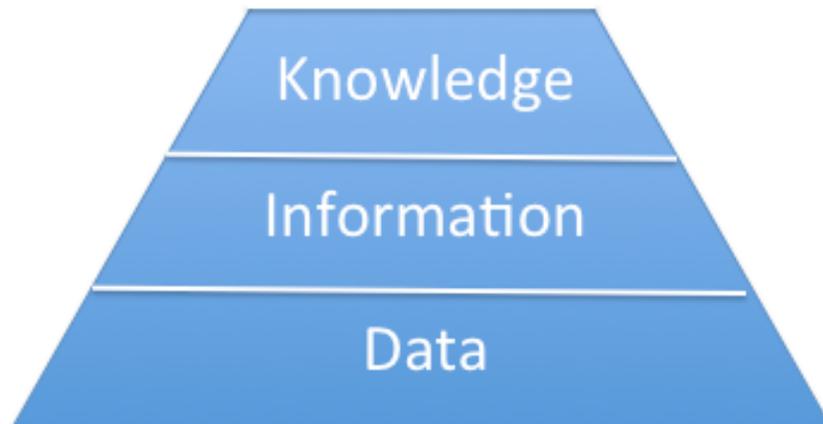
From Wikipedia, the free encyclopedia

The adjective **data-driven** means that progress in an activity is compelled by [data](#), rather than by [intuition](#) or by [personal experience](#).

Data-driven may refer to:

- [Data-driven programming](#), computer programming in which program statements describe data to be matched and the processing required
- [Data-driven journalism](#), a journalistic process based on analyzing and filtering large data sets
- [Data-driven testing](#), computer software testing done using a table of conditions directly as test inputs and verifiable outputs
- [Data-driven learning](#), a learning approach driven by research-like access to data
- [Data-driven science](#), an interdisciplinary field of scientific methods to extract knowledge from data
- [Data-driven control systems](#), systems of automatic control based on [system identification](#)
- [Data-driven security](#), a form of [model-driven security](#)
- [Data-driven marketing](#), a form of [digital marketing](#)

Data – Information - Knowledge



Data

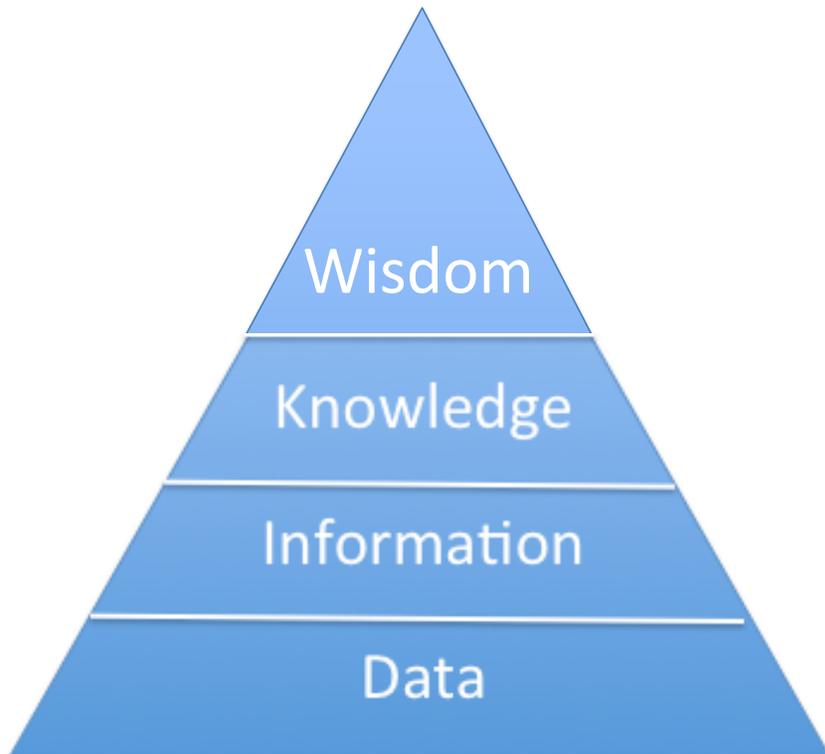
- raw observations, objective facts

Information

- data in meaningful context

Knowledge

- understanding about the world
 - explicit, codifiable (e.g., guideline)
 - tacit, not codifiable (e.g. expertise)
- process knowledge, i.e., put in a central line
- useful for explaining, predicting, and guiding future action



D-I-K Example

Data

- HgbA1C value 10.1%

Information

- occurred last Thursday
- 10.1% is above normal

Knowledge

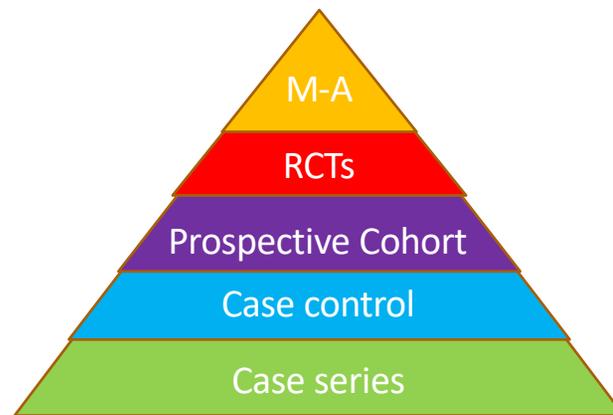
- high HgbA1C occurs in diabetes
- associated with higher risk for cardiovascular outcomes

Evidence?

Evidence : basis for a claim of knowledge

Evidence : data + study design + analysis

“Best” Evidence : per-protocol data + RCT design + ITT/meta analysis



Real-world designs: pragmatic trials, N-of-1, observational studies



Real-World Evidence : data + study design + analysis



Big data: genomics, imaging;
Real-world data: EHR, claims, registries, mobile health (patient-reported outcomes, wearables, IoT)



data mining, predictive analytics, natural language processing (NLP)

FRAMEWORK FOR FDA'S

REAL-WORLD EVIDENCE PROGRAM

Real-World Evidence (RWE) is the clinical evidence about the usage and potential benefits or risks of a medical product derived from analysis of Real-World Data (RWD).

technical amendment in Section 901 of the FDA Reauthorization Act of 2017 (Public law 115-52)

How do we think about trustworthiness of
real-world evidence?

“A scholar’s positive contribution is measured by the sum of the original data that he contributes. Hypotheses come and go but data remain. Theories desert us, while data defend us. They are our true resources, our real estate, and our best pedigree.”

Santiago Ramón y Cajal, 1897

Real-world designs:
embedded in routine care or
patients' daily lives



Real-World Evidence : data + study design + analysis

Real-world
mobile health
data



data mining,
predictive analytics,
natural language
processing (NLP)



mHealth

The application of wearable and ambient sensors, mobile apps, social media, and location-tracking technology singly or in combination to obtain data pertinent to wellness and disease diagnosis, prevention, and management.

*Sim I. Mobile Devices and Health. N Engl J Med
2019;381:956-68.*

Frontiers in Medicine

Mobile Health

A VIDEO FROM

Mobile Devices and Health

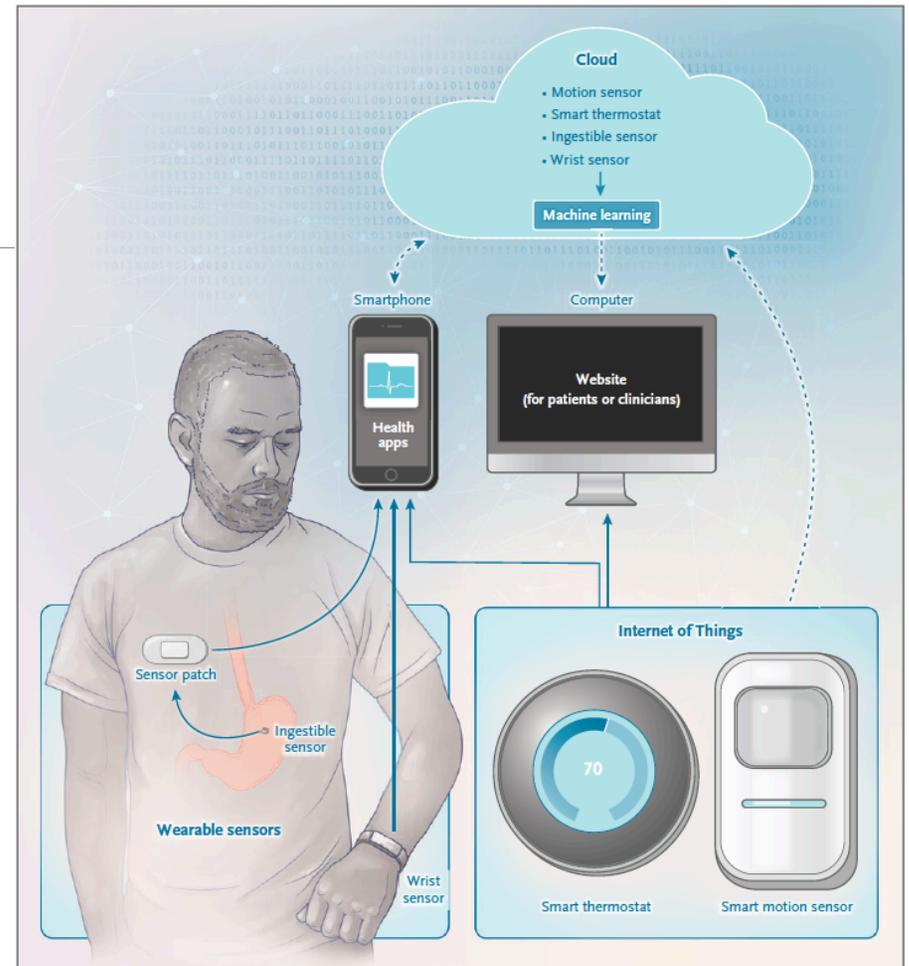
by Ida Sim

[10.1056/NEJMra1806949](https://doi.org/10.1056/NEJMra1806949)

<https://www.nejm.org/doi/10.1056/NEJMdo005587/full/?requestType=popUp&relatedArticle=10.1056%2FNEJMra1806949>

Digital Biomarkers

Digital biomarkers are physiological and behavioral measures collected by means of digital devices such as portables, wearables, implantables or digestibles that characterize, influence or predict health-related outcomes



Sim I. Mobile Devices and Health. *N Engl J Med* 2019;381:956-68.

Digital Biomarkers: Assumptions

Many new digital biomarkers will be defined

From new sensors and new patient-reported outcomes (PROs)

Continually changing hardware and software

Increasingly digital biomarkers that combine output from multiple sensors/sources

Data will flow far and wide to multiple actors for multiple purposes

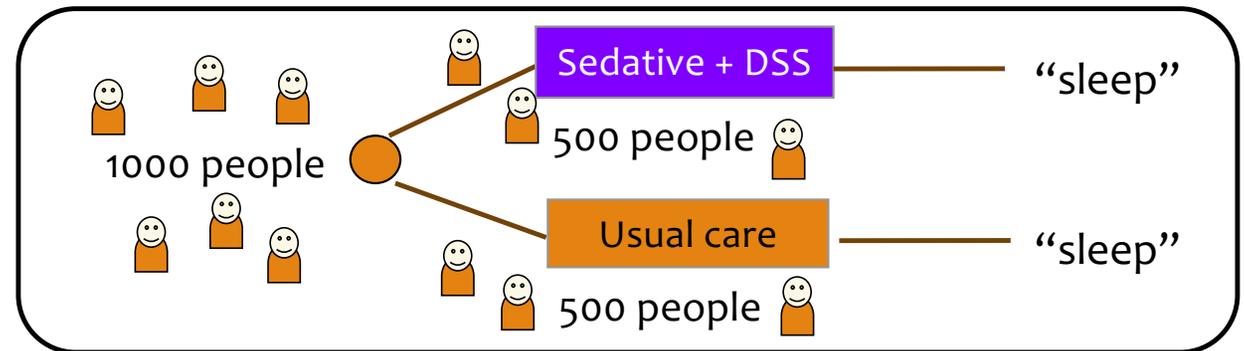
Example RWE Studies using Digital Biomarkers

RWE studies of

- a anti-hypertensive: measuring blood pressure remotely as a primary outcome
- an anti-hyperglycemic: measuring weight remotely as a secondary outcome
- a proton pump inhibitor: measuring depression through patient-reported outcomes (PROs) as a side effect
- a Parkinsons drug: measuring gait / mobility as a primary outcome
- real-world performance monitoring of a heart rate device: measuring heart rate variability

How do we think about trustworthiness of digital biomarkers?

Real-World Sedative Effectiveness



- A company is conducting a real-world study on the effectiveness of their approved sedative
 - real-world pragmatic design
 - using real world digital biomarkers collected using a mobile sensor in people’s nightly lives in the comfort of their own bed
 - supporting the clinician with a decision support system in prescribing the sedative when insomnia is detected
- What do we need to know to trust this study?

Categories of Digital Biomarker Metadata

January 24, 2020: I. Sim



Datapoint (may be a single observation or a computed biomarker)

Description of the datapoint itself



Source

From what sensor(s) did the datapoint come?



Acquisition / Processing

How was the datapoint acquired? processed?



Attribution

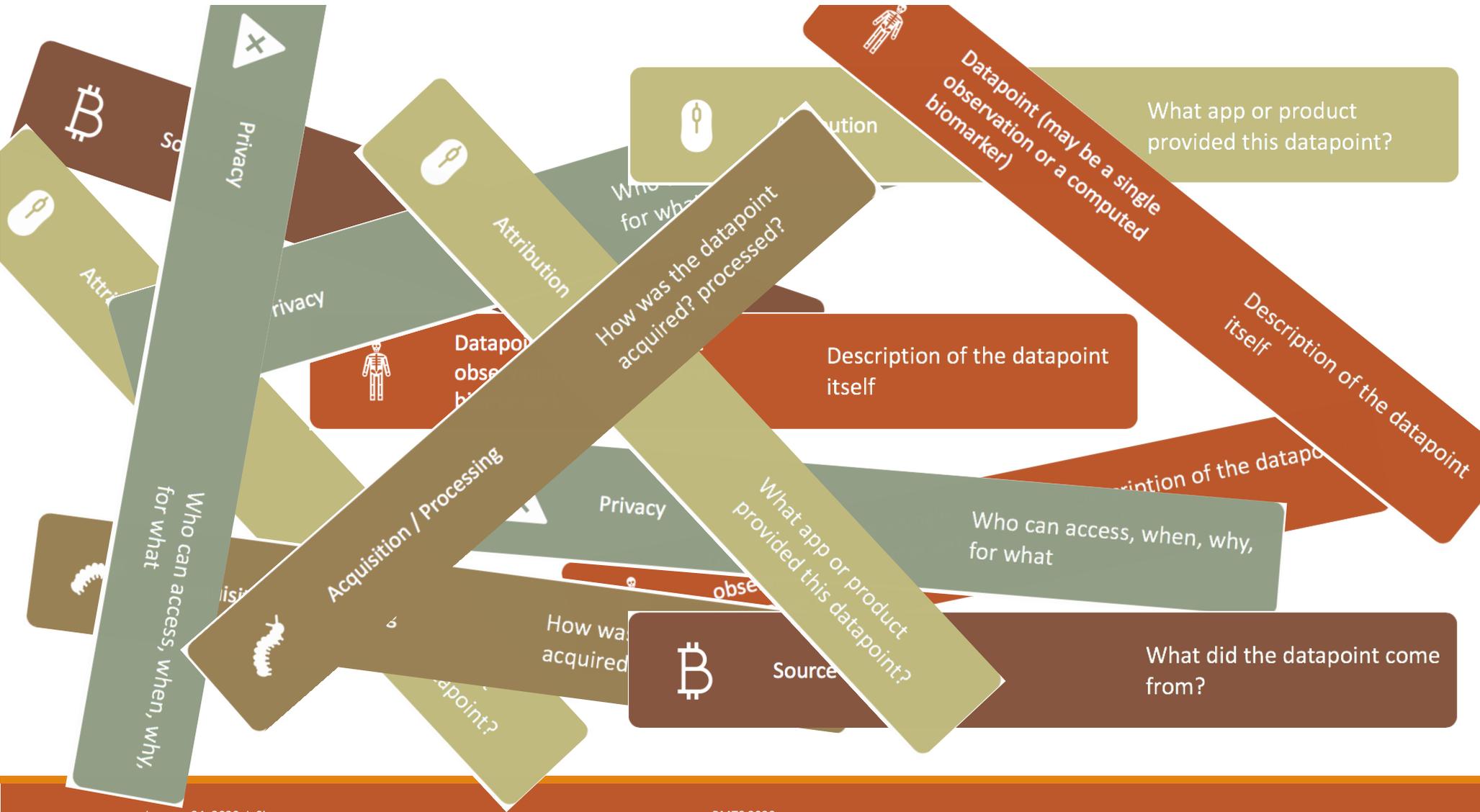
What app or product provided this datapoint?



Privacy

Who can access, when, why, for what

BMTS 2020



Categories of Metadata: Datapoint



Datapoint (may be a single observation or a computed biomarker)

Description of the datapoint itself



Source

From what sensor(s) did the datapoint come?



Acquisition / Processing

How was the datapoint acquired? processed?



Attribution

What app or product provided this datapoint?



Privacy

Who can access, when, why, for what

“Sleep” Digital Biomarker

QUESTIONS

What is the measure?

How is the measure defined?

Validity

Error

Natural variability

Uncertainty / confidence

Bias

Identity

Context

METADATA QUESTIONS FOR TRANSPARENCY

Sleep duration? Sleep quality? Sleep refreshment?

Time in bed? asleep? with/without microawakenings?

Comparison to a gold standard

Difference from “real” value

Variability within and among individuals

Probability that the person is asleep

Systematic errors, e.g., from skin color

Whose sleep duration is this?

In own bed? After jetlag?

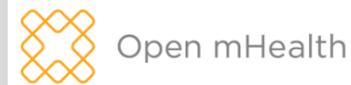
Unstandardized Sleep Datapoint Example

```
{  
  "sleep": 7.5  
  "date": 2019-04-05  
}
```

Open mHealth Sleep Datapoint Example

average sleep duration per day, in hours, from January 1, 2019 to April 5, 2019

```
{
  "sleep_duration": {
    "value": 7.5,
    "unit": "h"
  },
  "effective_time_frame": {
    "time_interval": {
      "start_date_time": "2019-01-01T07:25:00Z",
      "end_date_time": "2019-04-05 T07:25:00Z"
    }
  },
  "descriptive_statistic": "average"
  "descriptive_statistic_denominator": "d"
}
```



Digital Transparency Landscape



Open mHealth

data and metadata standards

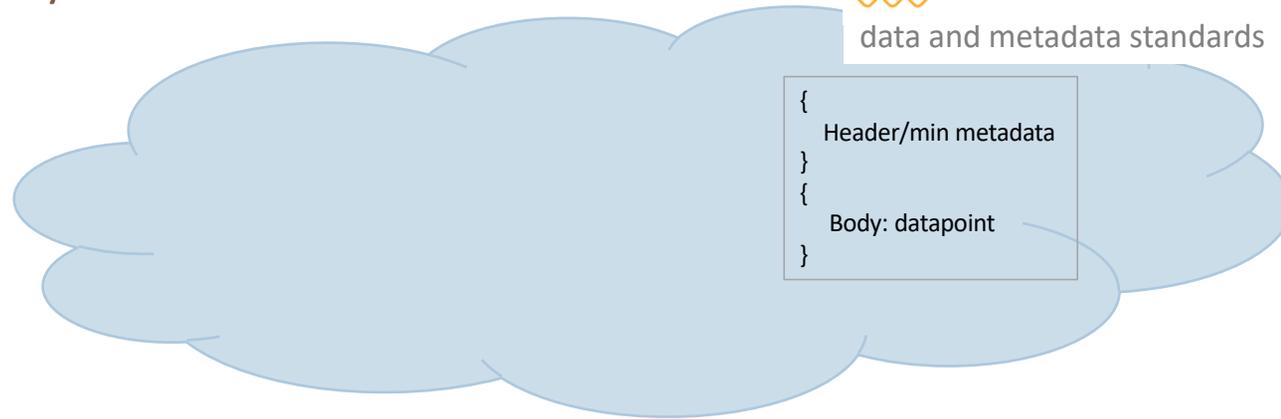
```
{  
  Header/min metadata  
}  
{  
  Body: datapoint  
}
```

Digital Transparency Landscape



Open mHealth

data and metadata standards



Study
Protocol

Data
Sharing
Policy

Categories of Metadata: Source



Datapoint (may be a single observation or a computed biomarker)

Description of the datapoint itself



Source

From what sensor(s) did the datapoint come?



Acquisition / Processing

How was the datapoint acquired? processed?



Attribution

What app or product provided this datapoint?



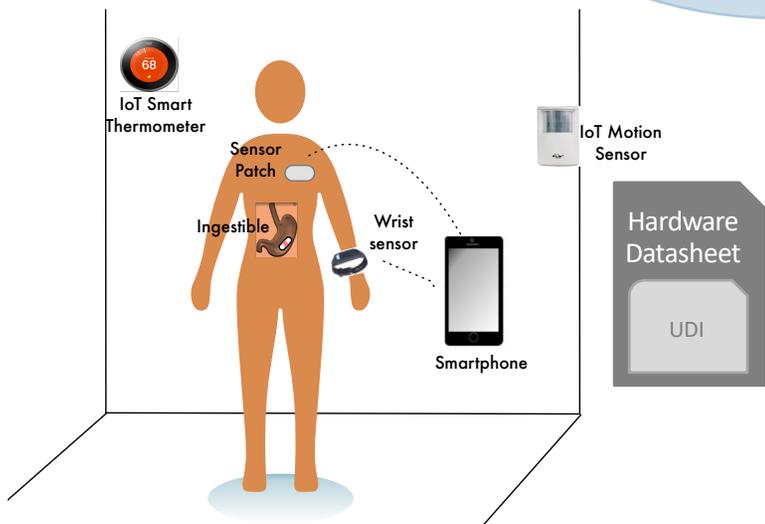
Privacy

Who can access, when, why, for what

Digital Transparency Landscape

 Open mHealth
data and metadata standards

```
{  
  Header/min metadata  
}  
{  
  Body: datapoint  
}
```



E.g., [HL7 FHIR Device Definition](#) and others

FDA UDI (Unique Device Identification) System

Categories of Metadata: Acquisition/ Processing



Datapoint (may be a single observation or a computed biomarker)

Description of the datapoint itself



Source

What did the datapoint come from?



Acquisition / Processing

How was the datapoint acquired? processed?



Attribution

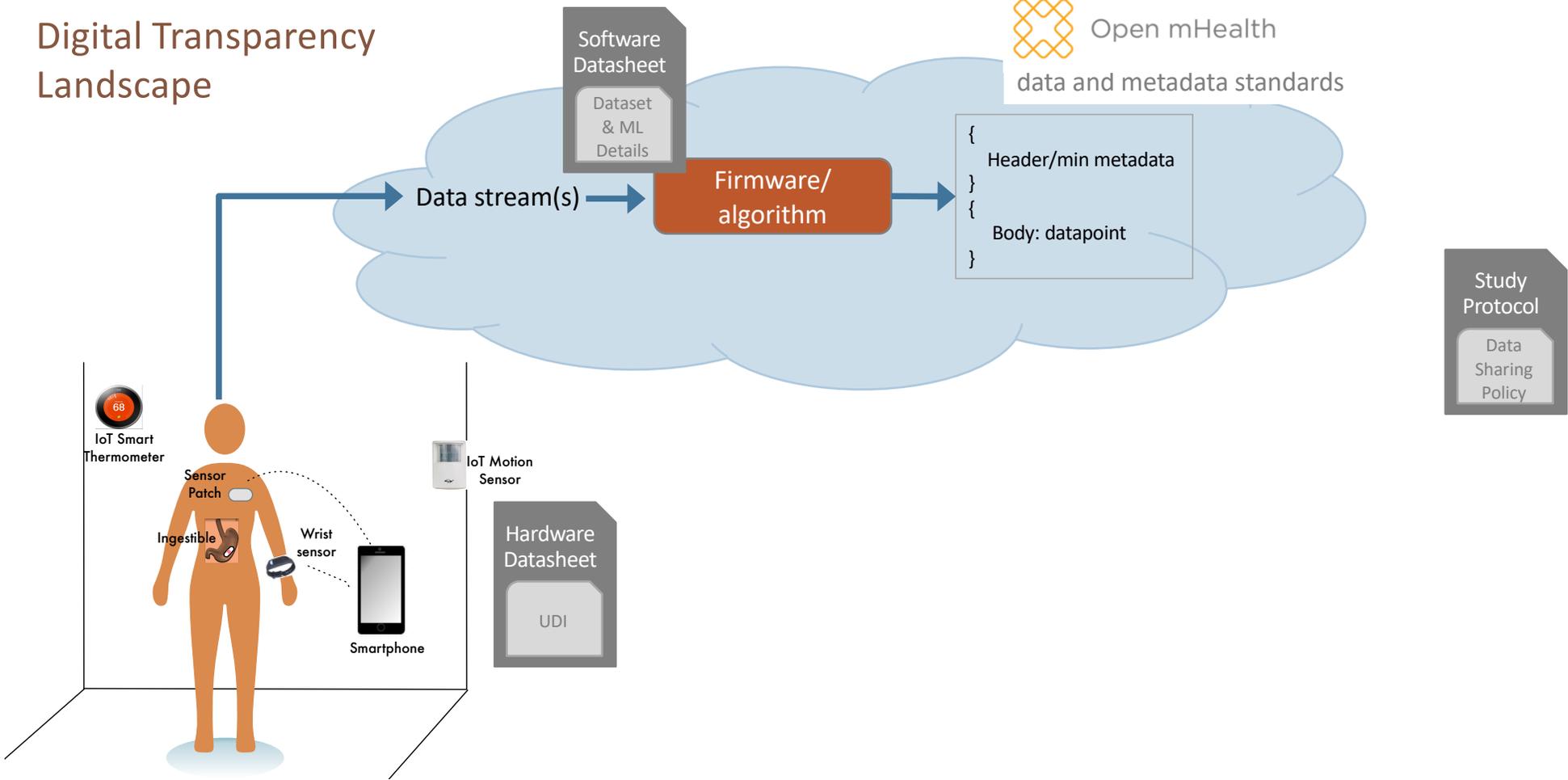
What app or product provided this datapoint?



Privacy

Who can access, when, why, for what

Digital Transparency Landscape



Software Datasheet

Algorithm method and version

- Input data streams
- Training data
 - description of the data that the model was trained on (e.g., [Datasheets for Datasets](#), [Dataset Nutrition Label](#))

Performance

- E.g., average PPG HR accuracy for person with a specific skin tone

Describing validation and verification

- by the manufacturer and/or independent 3rd party
- if and where open data set is available for testing reproducibility

Dataset Nutrition Label

Dataset Facts

ProPublica's Dollars
for Docs Data

Metadata

Filename	201612v1-docdollars-product_payments
Format	csv
Url	https://projects.propublica.org/docdollars/
Domain	healthcare
Keywords	Physicians, drugs, medicine, pharmaceutical, transactions
Type	tabular
Rows	500
Columns	18
Missing	5.2%
License	cc
Released	JAN 2017
Range	
From	AUG 2013
To	DEC 2015

Description This is the data used in ProPublica's Dollars for Docs news application. It is primarily based on CMS's Open Payments data, but we have added a few features. ProPublica has standardized drug, device and manufacturer names, and made a flattened table (product_payments) that allows for easier aggregating payments associated with each drug/device. In [1], one payment record can be attributed to up to five different drugs or medical devices. This table flattens the payments out so that each drug/device related to each payment gets its own line.

<https://ahmedhosny.github.io/datanutrition/>

Dataset Nutrition Label

Provenance

Source

Name U.S. Centers for Medicare & Medicaid Services

Url <https://www.cms.gov/OpenPayments/>

Email openpayments@cms.hhs.gov

Author

Name Propublica

Url <https://www.propublica.org/datastore/>

Email data.store@propublica.org

Dataset Nutrition Label

Statistics

Ordinal

name	type	count	uniqueEntries	mostFrequent	leastFrequent	missing
id	number	500	488 including missing	missing value (13)	multiple detected	2.60%
applicable_manufact...	number	500	4	100000000232 (417)	multiple detected	0%
date_of_payment	date	500	213 including missing	missing value (27)	multiple detected	5.40%
general_transaction_id	number	500	467 including missing	missing value (34)	multiple detected	6.80%
program_year	number	500	2 including missing	2014 (495)	missing value (5)	1.00%

Nominal

name	type	count	uniqueEntries	mostFrequent	leastFrequent	missing
product_name	string	500	16 including missing	Xarelto (200)	Aciphex (1)	3.20%
original_product_name	string	500	15	Xarelto (212)	Aciphex (1)	0%
product_ndc	number	500	21 including missing	5045857810 (201)	multiple detected	5.00%
product_is_drug	boolean	500	2 including missing	t (492)	missing value (8)	1.60%
payment_has_many	boolean	500	3 including missing	f (267)	missing value (29)	5.80%
teaching_hospital_id	number	500	2 including missing	0 (464)	missing value (36)	7.20%
physician_profile_id	number	500	230 including missing	missing value (32)	multiple detected	6.40%
recipient_state	string	500	40	CA (56)	multiple detected	0%
applicable_manufact...	string	500	5 including missing	Janssen Pharmaceut...	multiple detected	7.00%
teaching_hospital_ccn	number	500	2 including missing	0 (481)	missing value (19)	3.80%
product_slug	string	500	15 including missing	drug-xarelto (196)	drug-aciphex (1)	8.20%

Continuous

name	type	count	min	median	max	mean	standardDevi...	missing	zeros
total_amount...	number	500	0.14	14.00	5000	134.21	501.99	9.40%	0%

Discrete

name	type	count	min	median	max	mean	standardDevi...	missing	zeros
number_of_p...	number	500	1	1.00	1	1.00	0.00	4.80%	0%

Dataset Nutrition Label

Variables

id	A unique ID number for this payment & product combination. This is assigned by ProPublica for internal use
Applicable_manufacturer_or_applicable_gpo_making_payment_id	ID of the applicable manufacturer or submitting applicable GPO making the payment or other transfer of value
Date_of_payment	If a singular payment, then this is the actual date the payment was issued; if a series of payments or an aggregated set of payments, this is the date of the first payment to the covered recipient in this program year
General_transaction_id	System-assigned identifier to the general transaction at the time of submission
Program_year	The calendar year for which the payment is reported in Open Payments
Product_name	Derived from the 'name_of_associated_covered_drug_or_biologicalX' field (for drugs) or 'name_of_associated_covered_device_or_medical_supplyX' field (for medical devices). Where possible, multiple versions of the same product are converted to the same product_name (i.e. records for 'Zorvolex 65mg' and 'Zorvolex 35mg' will be converted to 'Zorvolex'). The original value is contained in original_product_name
Original_product_name	A copy of the original name_of_associated_covered_drug_or_biologicalX' field (for drugs) or 'name_of_associated_covered_device_or_medical_supplyX' field (for medical devices)
Product_ndc	If the product is a drug, this a copy of the original 'ndc_of_associated_covered_drug_or_biologicalX' field
Product_is_drug	'1' if the product is a drug (contained in a 'name_of_associated_covered_drug_or_biologicalX' field). '0' if the product is a medical device (contained in a 'name_of_associated_covered_device_or_medical_supplyX' field)
Payment_has_many	'1' if the original payment record included data on more than one drug or device, i.e. 'name_of_associated_covered_drug_or_biological1' and 'name_of_associated_covered_drug_or_biological2', 'name_of_associated_covered_device_or_medical_supply1' and 'name_of_associated_covered_device_or_medical_supply2', etc.
Teaching_hospital_id	Open Payments system-generated unique identifier of the teaching hospital receiving the payment or other transfer of value
Physician_profile_id	ID of the physician receiving the payment or other transfer of value
Recipient_state	The state or territory abbreviation of the primary business address of the physician or teaching hospital or non-covered recipient entity receiving the payment or other transfer of value if the primary business address is in the United States

Categories of Metadata: Attribution

January 24, 2020: I. Sim



Datapoint (may be a single observation or a computed biomarker)

Description of the datapoint itself



Source

What did the datapoint come from?



Acquisition / Processing

How was the datapoint acquired? processed?



Attribution

What app or product provided this datapoint?

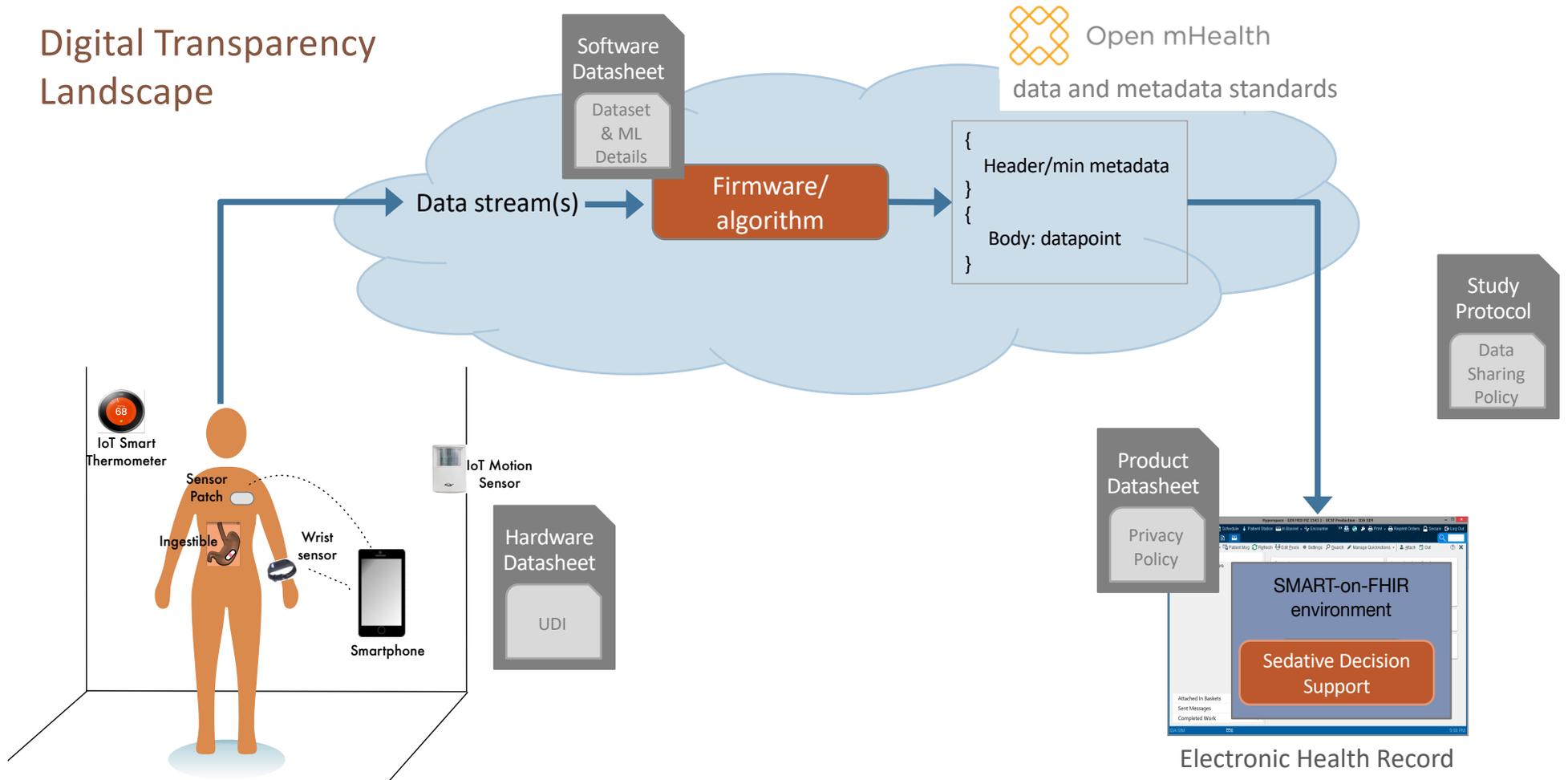


Privacy

Who can access, when, why, for what

BMTS 2020

Digital Transparency Landscape



Product Datasheet: ADviCE Package Insert

ADviCE: UCSF-Stanford Center of Excellence in Regulatory Science and Innovation (CERSI)

Facilitating adoption of digital health software tools by healthcare systems

INVOKANA®
(canagliflozin) tablets, for oral use

Revised: 07/2017
076265-170714

HIGHLIGHTS OF PRESCRIBING INFORMATION

These highlights do not include all the information needed to use INVOKANA® safely and effectively. See full prescribing information for INVOKANA.

INVOKANA (canagliflozin) tablets, for oral use
Initial U.S. Approval: 2013

WARNING: LOWER LIMB AMPUTATION
See full prescribing information for complete boxed warning.

- In patients with type 2 diabetes who have established cardiovascular disease (CVD) or at risk for CVD, INVOKANA has been associated with lower limb amputations, most frequently of the toe and midfoot; some also involved the leg (5.1)
- Before initiating, consider factors that may increase the risk of amputation. Monitor patients receiving INVOKANA for infections or ulcers of the lower limbs, and discontinue if these occur. (5.1)

----- **RECENT MAJOR CHANGES** -----
Boxed Warning 07/2017
Warnings and Precautions (5.1) 07/2017

----- **INDICATIONS AND USAGE** -----
INVOKANA is a sodium-glucose co-transporter 2 (SGLT2) inhibitor indicated as an adjunct to diet and exercise to improve glycemic control in adults with type 2 diabetes mellitus (1)

Limitation of Use:
• Not for treatment of type 1 diabetes mellitus or diabetic ketoacidosis (1)

----- **DOSAGE AND ADMINISTRATION** -----
• The recommended starting dose is 100 mg once daily, taken before the first meal of the day (2.1)
• Dose can be increased to 300 mg once daily in patients tolerating INVOKANA 100 mg once daily who have an eGFR of 60 mL/min/1.73 m² or greater and require additional glycaemic control (2.1)
• Assess renal function before initiation and periodically thereafter (2.2)

INVOKANA® (canagliflozin) tablets

- **Hypotension:** Before initiating INVOKANA, assess volume status and correct hypovolemia in patients with renal impairment, the elderly, in patients with low systolic blood pressure, or if on diuretics, ACEi, or ARB. Monitor for signs and symptoms during therapy (5.3)
 - **Ketoacidosis:** Assess patients who present with signs and symptoms of metabolic acidosis for ketoacidosis, regardless of blood glucose level. If suspected, discontinue INVOKANA, evaluate and treat promptly. Before initiating INVOKANA, consider risk factors for ketoacidosis. Patients on INVOKANA may require monitoring and temporary discontinuation of therapy in clinical situations known to predispose to ketoacidosis (5.3)
 - **Acute kidney injury and impairment in renal function:** Consider temporarily discontinuing in settings of reduced oral intake or fluid losses. If acute kidney injury occurs, discontinue and promptly treat. Monitor renal function during therapy (5.4)
 - **Hyperkalemia:** Monitor potassium levels in patients with impaired renal function and in patients predisposed to hyperkalemia (2.2, 5.5, 6.1, 8.6)
 - **Urosepsis and pyelonephritis:** Evaluate patients for signs and symptoms of urinary tract infections and treat promptly, if indicated (5.6)
 - **Hypoglycemia:** Consider a lower dose of insulin or the insulin secretagogue to reduce the risk of hypoglycemia when used in combination with INVOKANA (5.7)
 - **Genital mycotic infections:** Monitor and treat if indicated (5.8)
 - **Hypersensitivity reactions:** Discontinue INVOKANA and monitor until signs and symptoms resolve (5.9)
 - **Bone fracture:** Consider factors that contribute to fracture risk before initiating INVOKANA (5.10)
 - **Increased LDL-C:** Monitor LDL-C and treat if appropriate (5.11)
- **ADVERSE REACTIONS** -----
• Most common adverse reactions associated with INVOKANA (5% or greater incidence): female genital mycotic infections, urinary tract infection, and increased urination (6.1)
- To report SUSPECTED ADVERSE REACTIONS, contact Janssen Pharmaceuticals, Inc. at 1-800-526-7736 or FDA at 1-800-FDA-1088 or www.fda.gov/medwatch.**
- **DRUG INTERACTIONS** -----
• **UGT inducers (e.g., rifampin):** Canagliflozin exposure is reduced. Consider

- Software version
- Target users
- Desired implementation approach
- Data on effectiveness or safety (if any)
- Privacy Policy: Where data go. Who owns data. Is the tool actually used anywhere?



Categories of Metadata: Privacy



Datapoint (may be a single observation or a computed biomarker)

Description of the datapoint itself



Source

What did the datapoint come from?



Acquisition / Processing

How was the datapoint acquired? processed?



Attribution

What app or product provided this datapoint?



Privacy

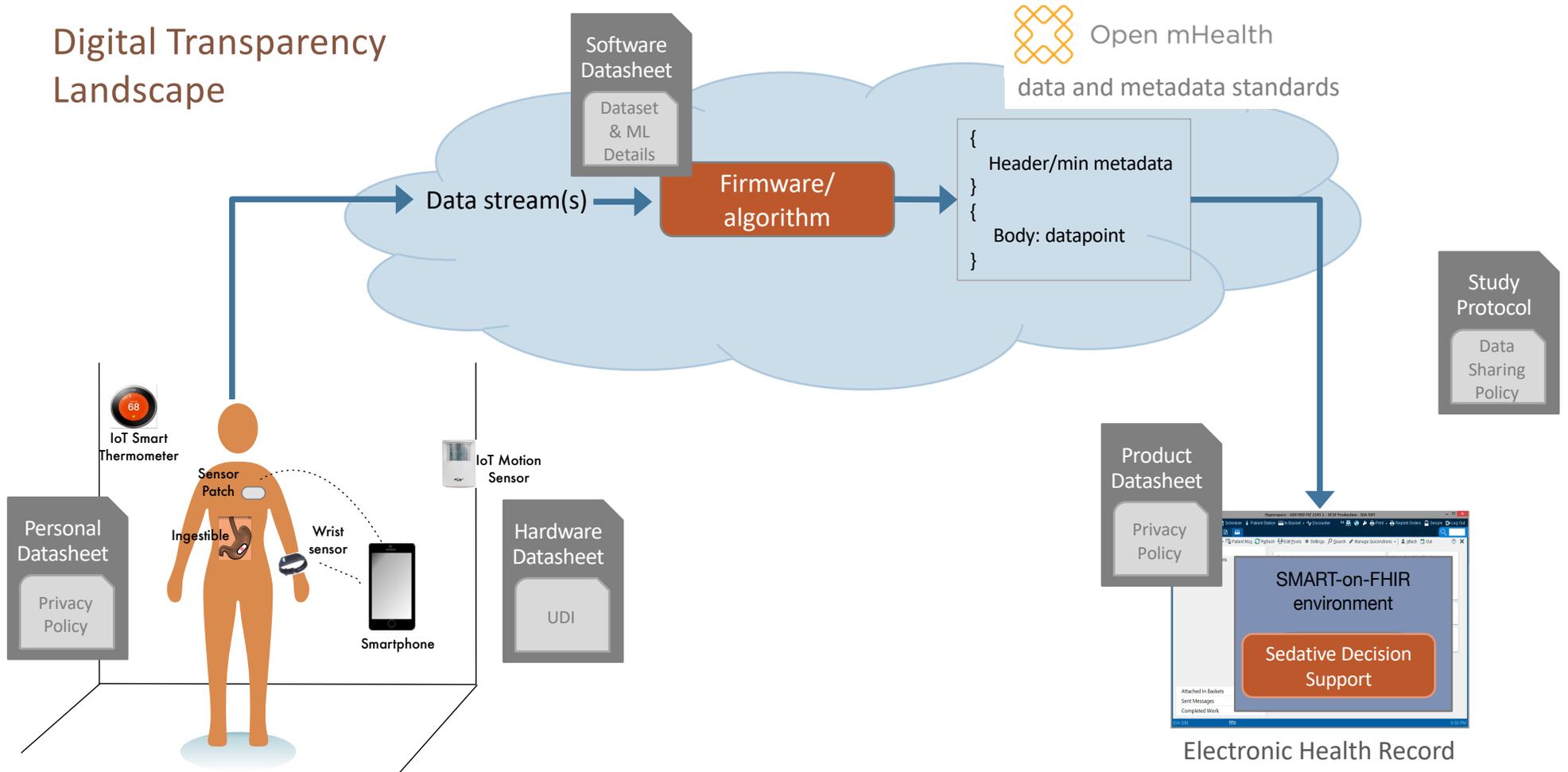
Who can access, when, why, for what

Privacy

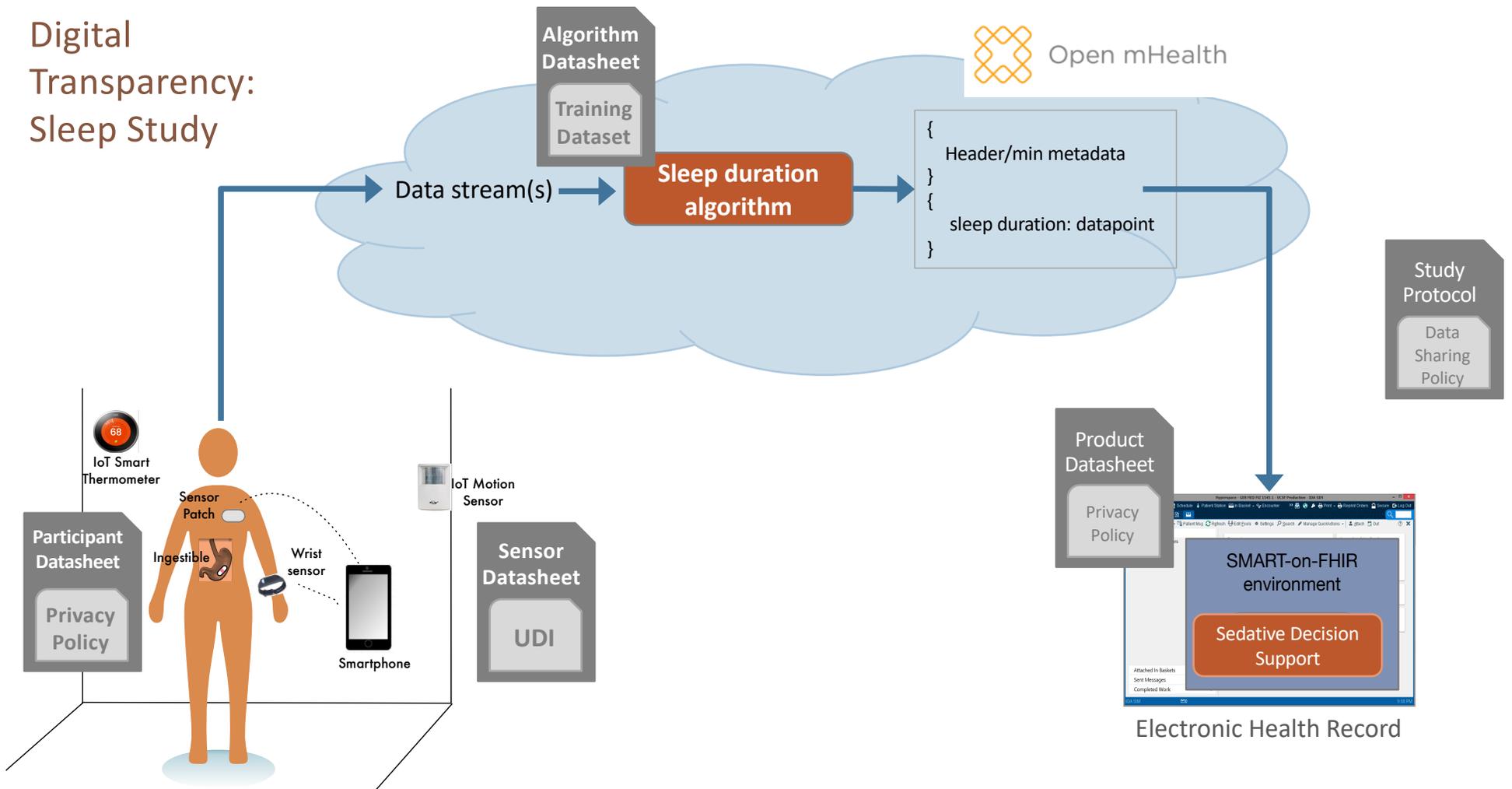
Who can access which digital biomarkers when and for what?

- Study's Data Sharing Statement
- Product Datasheet – privacy policy
- Personal Datasheet – privacy policy

Digital Transparency Landscape



Digital Transparency: Sleep Study



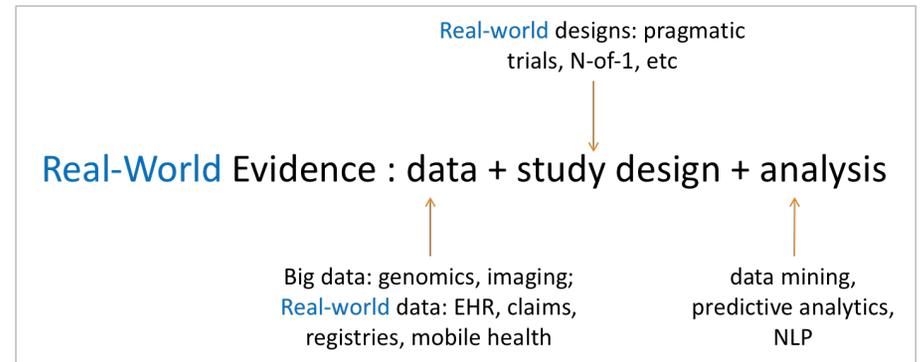
Open Questions

To support digital transparency for RWE

- Are these the most important metadata?
- How to coordinate the collection of metadata?
- How to establish pointers/references?
- Who needs to do this? Who *will* do this? Who are the stakeholders?
- How is the digital transparency ecosystem governed?
- How is the digital transparency ecosystem sustained?
- ...

Conclusion: Towards a Digital Transparency Ecosystem

Evidence : basis for a claim of knowledge



Metadata categories: datapoint, source, acquisition and processing, attribution, and privacy

Digital transparency with a suite of datasheets is necessary for trustworthy real-world digital biomarkers and real-world evidence

Transparency/trustworthiness is a property of the ecosystem. Many stakeholders will need to work together