Trustworthiness of Person-generated Health Data

JULY 2019
Luca Foschini, PhD
Co-founder & Chief Data Scientist
luca@evidation.com | @calimagna
Examples of AI applications across the human lifespan. Credits: Debbie Maizels / Springer Nature

### Table 3 | Selected reports of machine- and deep-learning algorithms to predict clinical outcomes and related parameters

<table>
<thead>
<tr>
<th>Prediction</th>
<th>$n$</th>
<th>AUC</th>
<th>Publication (Reference number)</th>
</tr>
</thead>
<tbody>
<tr>
<td>In-hospital mortality, unplanned readmission, prolonged LOS, final discharge diagnosis</td>
<td>216,221</td>
<td>0.93<em>0.75</em>0.85*</td>
<td>Rajkomar et al.$^{96}$</td>
</tr>
<tr>
<td>All-cause 3-12 month mortality</td>
<td>221,284</td>
<td>0.93</td>
<td>Avati et al.$^{91}$</td>
</tr>
<tr>
<td>Readmission</td>
<td>1,068</td>
<td>0.78</td>
<td>Shameer et al.$^{106}$</td>
</tr>
<tr>
<td>Sepsis</td>
<td>230,936</td>
<td>0.67</td>
<td>Horng et al.$^{102}$</td>
</tr>
<tr>
<td>Septic shock</td>
<td>16,234</td>
<td>0.83</td>
<td>Henry et al.$^{103}$</td>
</tr>
<tr>
<td>Severe sepsis</td>
<td>203,000</td>
<td>0.85*</td>
<td>Culliton et al.$^{104}$</td>
</tr>
<tr>
<td><em>Clostridium difficile</em> infection</td>
<td>256,732</td>
<td>0.82**</td>
<td>Oh et al.$^{93}$</td>
</tr>
<tr>
<td>Developing diseases</td>
<td>704,587</td>
<td>range</td>
<td>Miotto et al.$^{97}$</td>
</tr>
<tr>
<td>Diagnosis</td>
<td>18,590</td>
<td>0.96</td>
<td>Yang et al.$^{95}$</td>
</tr>
<tr>
<td>Dementia</td>
<td>76,367</td>
<td>0.91</td>
<td>Cleret de Langavant et al.$^{92}$</td>
</tr>
<tr>
<td>Alzheimer’s Disease (+ amyloid imaging)</td>
<td>273</td>
<td>0.91</td>
<td>Mathotaarachchi et al.$^{98}$</td>
</tr>
<tr>
<td>Mortality after cancer chemotherapy</td>
<td>26,946</td>
<td>0.94</td>
<td>Elfiky et al.$^{95}$</td>
</tr>
<tr>
<td>Disease onset for 133 conditions</td>
<td>298,000</td>
<td>range</td>
<td>Razavian et al.$^{105}$</td>
</tr>
<tr>
<td>Suicide</td>
<td>5,543</td>
<td>0.84</td>
<td>Walsh et al.$^{86}$</td>
</tr>
<tr>
<td>Delirium</td>
<td>18,223</td>
<td>0.68</td>
<td>Wong et al.$^{100}$</td>
</tr>
</tbody>
</table>

LOS, length of stay; $n$, number of patients (training+validation datasets). For AUC values: *, in-hospital mortality; +, unplanned readmission; #, prolonged LOS; **, all patients; @, structured+unstructured data; ++, for University of Michigan site.

To trust an MI system requires to trust all steps of its workflow

Fig. 1. The nine stages of the machine learning workflow. Some stages are data-oriented (e.g., collection, cleaning, and labeling) and others are model-oriented (e.g., model requirements, feature engineering, training, evaluation, deployment, and monitoring). There are many feedback loops in the workflow. The larger feedback arrows denote that model evaluation and monitoring may loop back to any of the previous stages. The smaller feedback arrow illustrates that model training may loop back to feature engineering (e.g., in representation learning).

Source: Software Engineering for Machine Learning: A Case Study Amershi et al., International Conference on Software Engineering (ICSE) 2019
To trust an MI system requires to trust all steps of its workflow.

Fig. 1. The nine stages of the machine learning workflow. Some stages are data-oriented (e.g., collection, cleaning, and labeling) and others are model-oriented (e.g., model requirements, feature engineering, training, evaluation, deployment, and monitoring). There are many feedback loops in the workflow. The larger feedback arrows denote that model evaluation and monitoring may loop back to any of the previous stages. The smaller feedback arrow illustrates that model training may loop back to feature engineering (e.g., in representation learning).

Source: Software Engineering for Machine Learning: A Case Study Amershi et al., International Conference on Software Engineering (ICSE) 2019
Define "bad data": a Real-World Evidence (RWE) perspective

- Data Reliability
- Data Relevancy
- Fit-for-Use
- RWD
- Accrual
- Quality Control
- Verification
- Validation
- Conformance
- Completeness
- Plausibility

Credits: Duke-Margolis Center for Health Policy, reproduced with permission. Adapted from:
• Framework for FDA's Real World Evidence Program. Food and Drug Administration, 2018
Define “bad data”: a Real-World Evidence (RWE) perspective

Fit-for-Use

RWD

Data Reliability

Accrual

Quality Control

Verification

Validation

Conformance

Completeness

Plausibility

Data Relevancy

Completeness

Plausibility

Credits: Duke-Margolis Center for Health Policy, reproduced with permission. Adapted from:


• Framework for FDA’s Real World Evidence Program. Food and Drug Administration, 2018

Evidation Study Platform ingests and analyze high-frequency Person-generated Health Data (PGHD) from a variety of sensors, services, and applications.

Examples of PGHD:
- Biometric data
- Symptoms
- Self-reported medical history
- Activity tracking
- Medication adherence/effects
- Sleep tracking
Conformance: Does the data fit our expected schema?

- Ingesting sleep data from Fitbit
- Two different data formats depending on device and data availability

<table>
<thead>
<tr>
<th>timestamp</th>
<th>level_deep_seconds</th>
<th>level_light_seconds</th>
<th>level_rem_seconds</th>
<th>level_wake_seconds</th>
</tr>
</thead>
<tbody>
<tr>
<td>2017-04-01T23:58:30.000</td>
<td>NaN</td>
<td>NaN</td>
<td>NaN</td>
<td>1080</td>
</tr>
<tr>
<td>2017-04-02T00:16:30.000</td>
<td>NaN</td>
<td>NaN</td>
<td>NaN</td>
<td>NaN</td>
</tr>
</tbody>
</table>

```json
"data": [
  {
    "dateTime": "2017-04-01T23:58:30.000",
    "level": "deep",
    "seconds": <value>
  },
  {
    "dateTime": "2017-04-02T00:16:30.000",
    "level": "rem",
    "seconds": <value>
  },
  ...
],
"shortData": [
  {
    "dateTime": "2017-04-02T05:58:30.000",
    "level": "wake",
    "seconds": <value>
  }
]
```

```json
"data": [
  {
    "dateTime": "2017-04-01T23:58:30.000",
    "level": "sleep",
    "seconds": <value>
  },
  {
    "dateTime": "2017-04-02T12:13:00.000",
    "level": "REM",
    "seconds": <value>
  },
  {
    "dateTime": "2017-04-02T12:14:00.000",
    "level": "wake",
    "seconds": <value>
  }
]
```
Completeness: How much data do we observe, and do our observations meet our expectations?

- Data availability relies on both the systems that capture/store and the patients who generate it
  - Participant Missing: stopped wearing the device; does not sync
  - Device Missing: battery died and stopped recording
  - System Missing: server/API errors
Completeness: How much data do we observe, and do our observations meet our expectations?

- Data availability relies on both the systems that capture/store and the patients who generate it
  - Participant Missing: stopped wearing the device; does not sync
  - Device Missing: battery died and stopped recording
  - System Missing: server/API errors

- Multi-sensor systems allow for data completeness checks that take into account multiple data streams
  - Allows you to differentiate true missing from null values
Plausibility: Can we trust the values we observe in the data?
Plausibility: Can we trust the values we observe in the data?
Need: Data documentation frameworks for medical MI applications

**Motivation**

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

Labeled Faces in the Wild was created to provide images that can be used to study face recognition in the unconstrained setting where image characteristics (such as pose, illumination, resolution, focus), subject demographic makeup (such as age, gender, race) or appearance (such as hairstyle, makeup, clothing) cannot be controlled. The dataset was created for the specific task of pair matching; given a pair of images each containing a face, determine whether or not the images are of the same person.

Who created this dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

The initial version of the dataset was created by Gary B. Huang, Manu Ramesh, Tamara Berg, and Eric Learned-Miller, most of whom were researchers at the University of Massachusetts Amherst at the time of the dataset’s release in 2007.

Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grant and the grant name and number.

The construction of the LFW database was supported by a United States National Science Foundation CAREER Award.

**Collection Process**

How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

The names for each person in the dataset were determined by an operator by looking at the caption associated with the person’s photograph.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software procedures, API)? How were these mechanisms or procedures validated?

The raw images for this dataset were obtained from the Faces in the Wild database collected by Tamara Berg at Berkeley.

**Uses**

Has the dataset been used for any tasks already? If so, please provide a description.

Papers using this dataset and the specified evaluation protocol are listed in [http://vis-www.cs.umass.edu/lfw/results.html](http://vis-www.cs.umass.edu/lfw/results.html)

Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.

Papers using this dataset and the specified training/evaluation protocols are listed under “Methods” section of [http://vis-www.cs.umass.edu/lfw/results.html](http://vis-www.cs.umass.edu/lfw/results.html).

What (other) tasks could the dataset be used for?

The LFW dataset can be used for the face identification problem. Some researchers have developed protocols to use the images in the LFW dataset for face identification.

**Preprocessing/cleaning/labeleding**

Was any preprocessing/cleaning/labeleding of the data done (e.g., digitalization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removing instances, processing of missing values)? If so, please provide a description. If not, you may skip the remainder of the questions in this section.

The following steps were taken to process the data:

1. Gathering raw images: First the raw images for this dataset were obtained from the Faces in the Wild dataset consisting of images and associated captions gathered from news articles found on the web.
2. Running the Viola-Jones face detector: The OpenCV vendor

**Distribution**

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.

Yes, the dataset is publicly available.

How will the dataset be distributed (e.g., on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?


When will the dataset be distributed?

The dataset was released in October, 2007.

**Maintenance**

Who will be supporting/hosting/maintaining the dataset?

The dataset is hosted at the University of Massachusetts.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)?

All questions and comments can be sent to Gary Huang: gh huang@cs.umass.edu.

Is there an erratum? If so, please provide a link or other access point.

All changes to the dataset will be announced through the LFW mailing list. Those who would like to sign up should send an email to lfw-subscribe@cs.umass.edu. Errata are listed under the “Errata” section of [http://vis-www.cs.umass.edu/lfw/index.html](http://vis-www.cs.umass.edu/lfw/index.html).

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub).

All changes to the dataset will be announced through the LFW mailing list.

**Other efforts (not an exhaustive list):**

- **Data statements for NLP:** Toward mitigating system bias and enabling better science. Bender et al. Transactions of the Association for Computational Linguistics (2018).

Need: Data documentation frameworks for medical MI applications

To trust an MI system requires to trust all steps of its workflow

Fig. 1. The nine stages of the machine learning workflow. Some stages are data-oriented (e.g., collection, cleaning, and labeling) and others are model-oriented (e.g., model requirements, feature engineering, training, evaluation, deployment, and monitoring). There are many feedback loops in the workflow. The larger feedback arrows denote that model evaluation and monitoring may loop back to any of the previous stages. The smaller feedback arrow illustrates that model training may loop back to feature engineering (e.g., in representation learning).

Source: *Software Engineering for Machine Learning: A Case Study* Amershi et al., International Conference on Software Engineering (ICSE) 2019
Trusting model building: Reproducibility

Evaluation Metrics:

A. Technical replicability
   1. Code available
   2. Public dataset

B. Statistical replicability
   1. Variance reported

C. Conceptual replicability
   1. Multiple datasets

Source: Reproducibility in Machine Learning for Health McDermott et al., Workshop in Reproducibility in Machine Learning (ICLR) 2019
https://arxiv.org/abs/1907.01463
Trusting model building: Adversarial examples

Original tracing
Prediction: AF
100% confidence

+ Smooth Perturbation

Combined tracing
Prediction: Normal
100% confidence

Source: Adversarial Examples for Electrocardiograms Han et al., Workshop on debugging Machine Learning systems (ICLR) 2019 https://arxiv.org/abs/1905.05163

See also: Adversarial attacks on medical machine learning Finlayson et al. Science (2019)

Source: Synthesizing Robust Adversarial Examples Athalye et al., (ICML) 2018
To trust an MI system requires to trust all steps of its workflow.

Fig. 1. The nine stages of the machine learning workflow. Some stages are data-oriented (e.g., collection, cleaning, and labeling) and others are model-oriented (e.g., model requirements, feature engineering, training, evaluation, deployment, and monitoring). There are many feedback loops in the workflow. The larger feedback arrows denote that model evaluation and monitoring may loop back to any of the previous stages. The smaller feedback arrow illustrates that model training may loop back to feature engineering (e.g., in representation learning).

Source: Software Engineering for Machine Learning: A Case Study Amershi et al., International Conference on Software Engineering (ICSE) 2019
Trusting model outputs: test-time monitoring

- **General idea**: An additional system (can be human) evaluates trustworthiness of model output.

- **High level mechanism**: Monitoring system determines if test data is "anomalous" as compared to the data the model has been trained on.

- **Active area of research in ML community**:

- Can be used to detect distribution shifts post-deployment.

Machine Intelligence systems are Complex Systems. System Engineering principles can be applied to ensure end-to-end reliability.

Figure 1: Only a small fraction of real-world ML systems is composed of the ML code, as shown by the small black box in the middle. The required surrounding infrastructure is vast and complex.

Source: *Hidden technical debt in machine learning systems*. Sculley et al., Advances in neural information processing systems (NeurIPS) 2015

Thank you

Luca Foschini, PhD
Co-founder & Chief Data Scientist
luca@evidation.com  |  @calimagna