FGAI4H-P-046-A04

Helsinki, 20-22 September 2022

Source:	dentalXrai
Title:	Att.4 – Presentation - The Value of Benchmarking for Dental AI
Purpose:	Discussion

Contact:	Joachim Krois dentalXrai, Germany	E-mail: Joachim.krois@charite.de

Abstract:	This PPT contains a presentation on the Value of Benchmarking for Dental AI.

AI for Health

ITU-WHO Focus Group

The TG Dental - The Value of Benchmarking for Dental AI

2022/09/19 - 1st TG DENTAL SYMPOSIUM

Joachim Krois
dentalXrai, Germany

itu.int/go/fgai4h

Urgency contact
hamdanai4h@gmail.com

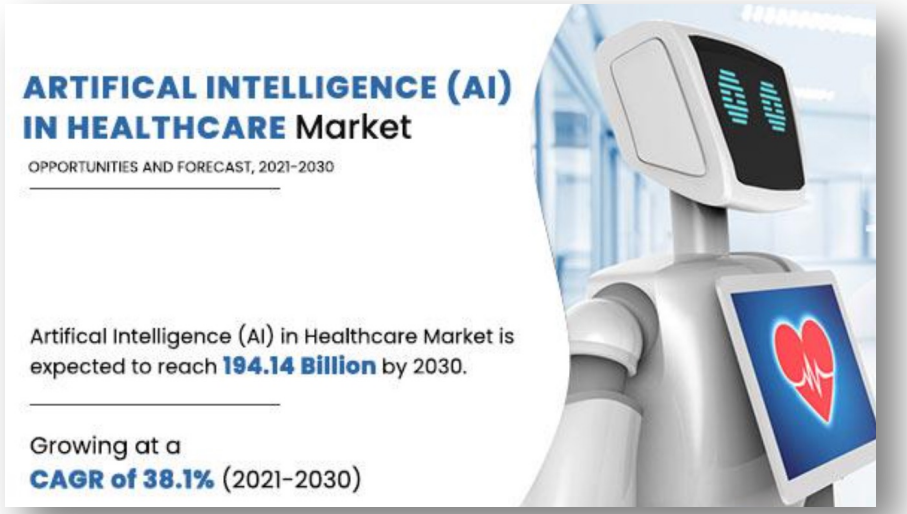
World Health Organization

ITU

Why we should be excited about Dental AI?

AI for Health
ITU-WHO Focus Group

https://online.stanford.edu/programs/artificial-intelligence-healthcare



https://www.alliedmarketresearch.com/artificial-intelligence-in-healthcare-market

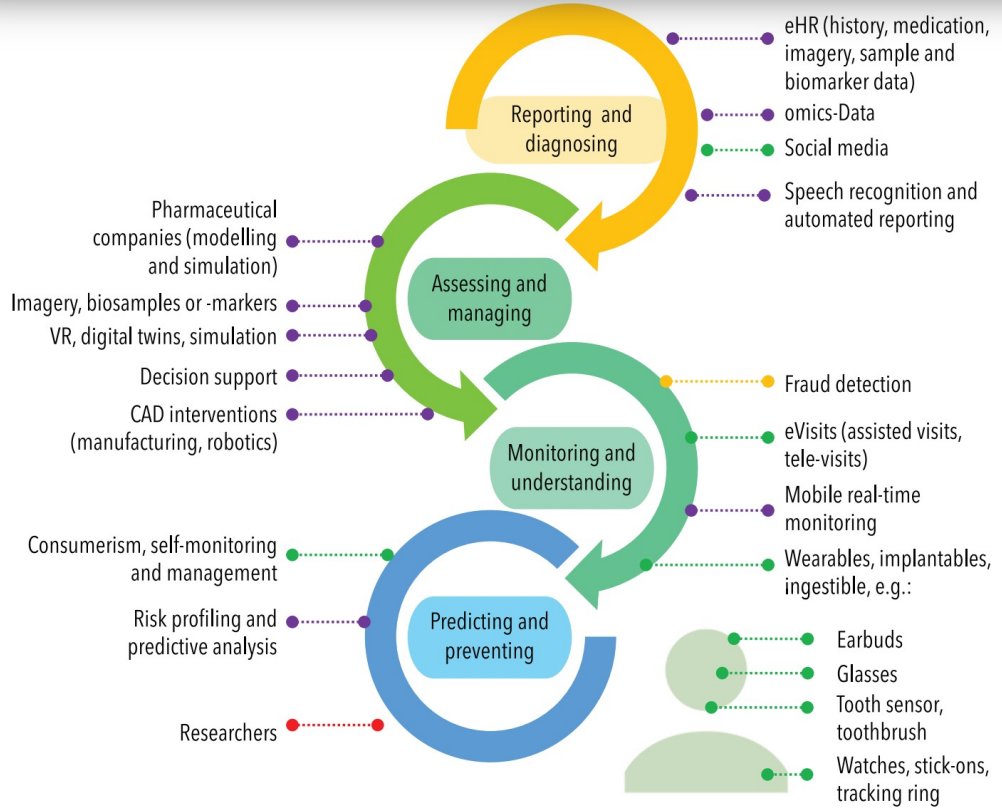
Critical Reviews in Oral Biology & Medicine

Data Dentistry: How Data Are Changing Clinical Care and Research

F. Schwendicke[1] and J. Krois[1]

Journal of Dental Research
2022, Vol. 101(1) 21–29

DOI: 10.1177/00220345211020265
journals.sagepub.com/home/jdr

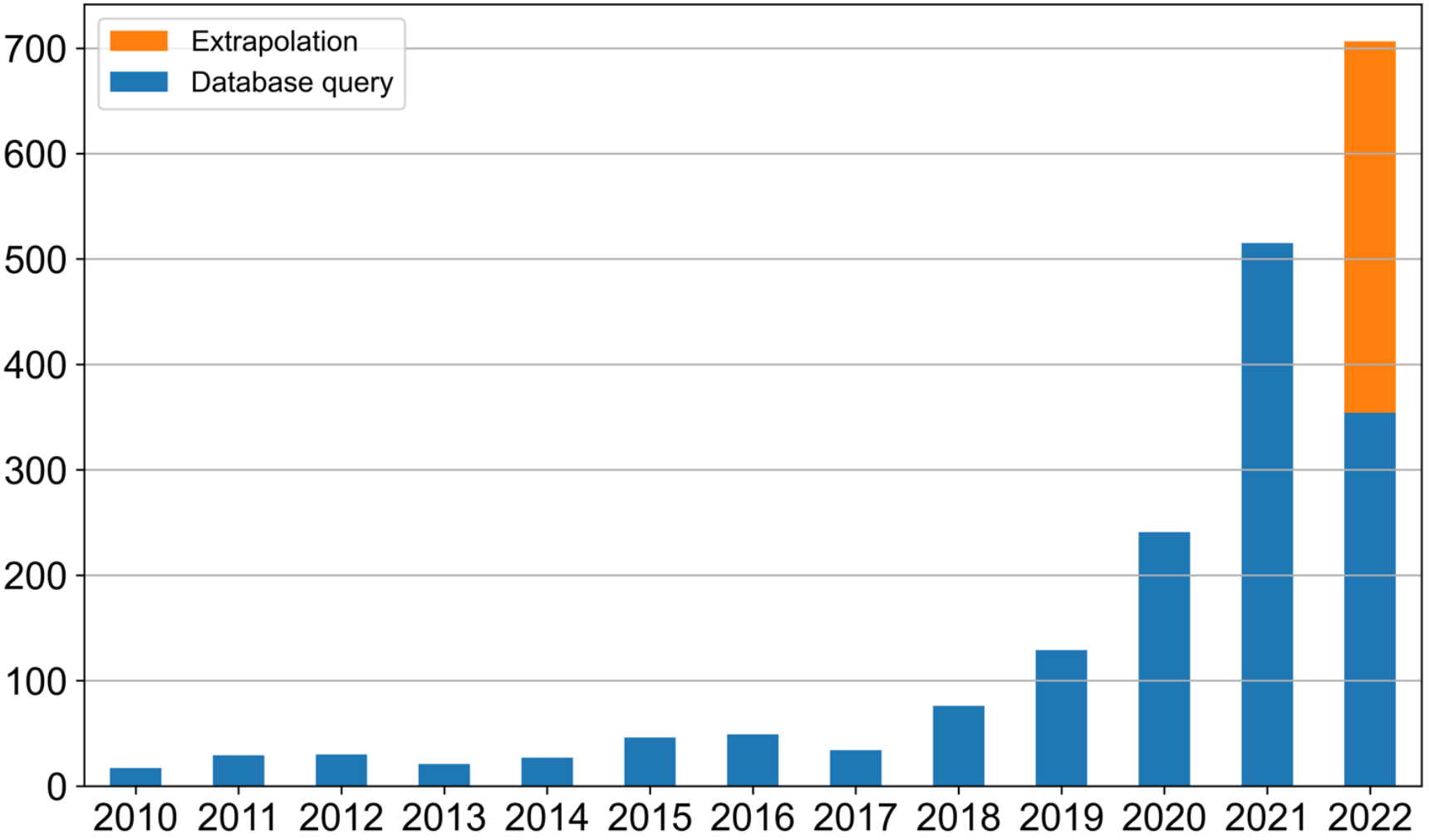


Schwendicke and Krois 2022

A surge in Dental AI research!

AI for Health
ITU-WHO Focus Group

Publications on "AI and Dentistry" since 2010



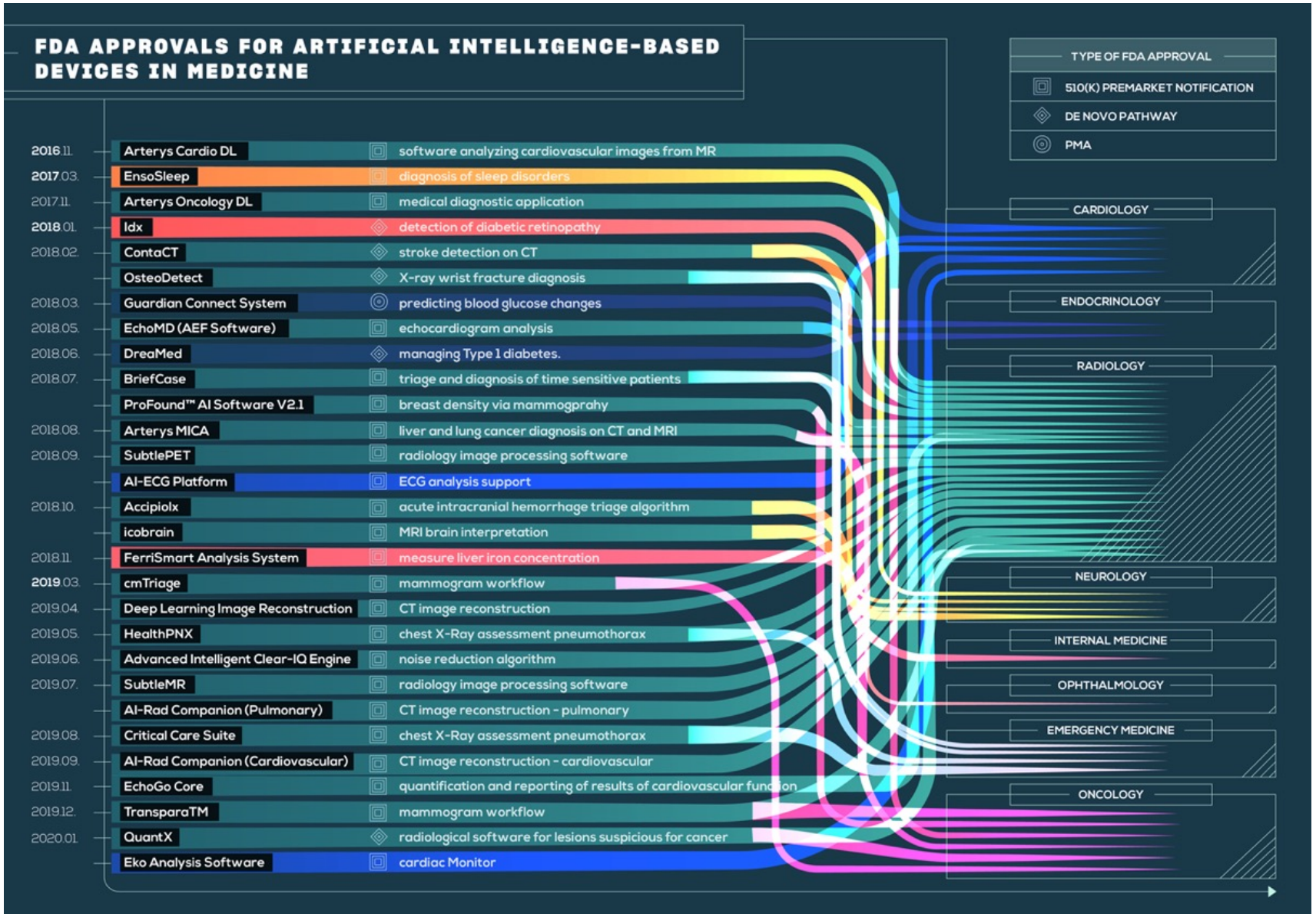
Overall, in 2019 PubMed lists 14,000 publications related to AI

(Source: PubMed, accessed: August 2022)

AI-based medical applications are real!

AI for Health
ITU-WHO Focus Group

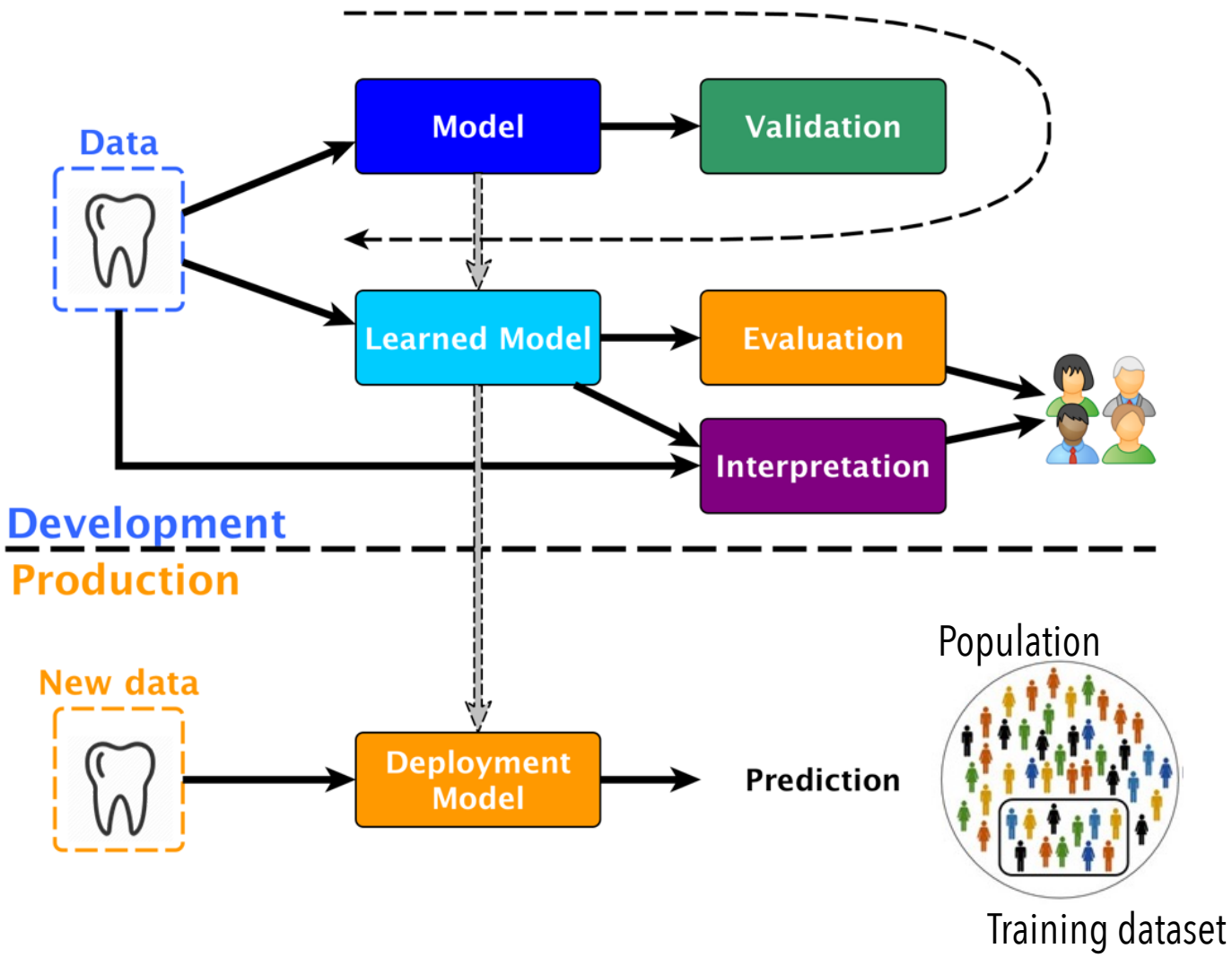
By March 2020, **64 AI/ML based medical devices** and algorithms were approved by FDA (U.S. Food and Drug Administration)



Benjamens et al 2020

Topic group: Dental Diagnostics and Digital Dentistry (TG-Dental)

AI for Health
ITU-WHO Focus Group

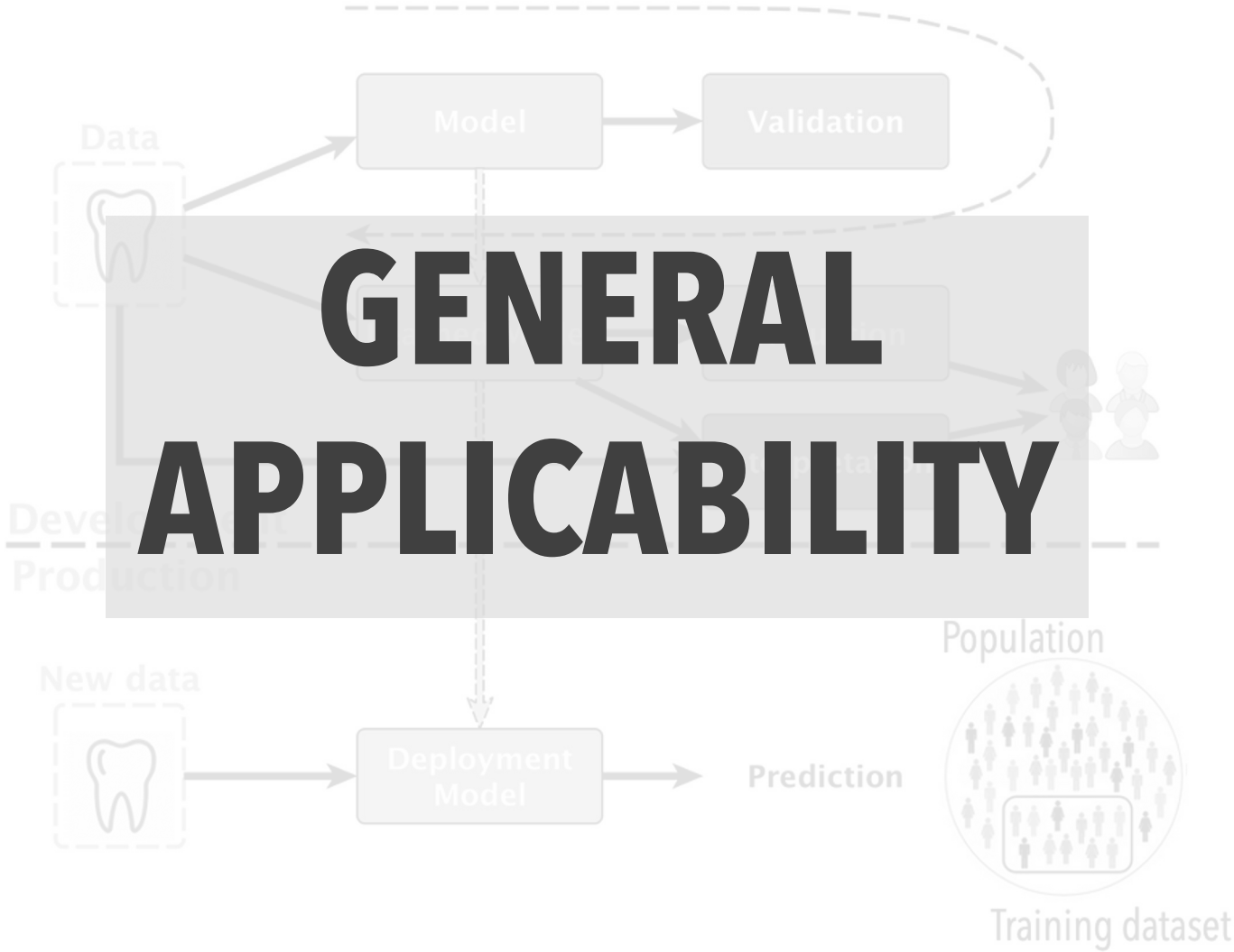


AI for Health
An ITU Focus Group
In collaboration with WHO

The International Telecommunication Union (ITU) and the World Health Organization (WHO) collaborate to establish an open group of experts to develop a generalizable benchmarking framework for health solutions based on artificial intelligence (AI), the ITU/WHO Focus Group on AI for Health (FG-AI4H).

Topic group: Dental Diagnostics and Digital Dentistry (TG-Dental)

AI for Health
ITU-WHO Focus Group



GENERAL APPLICABILITY

AI for Health
An ITU Focus Group
In collaboration with WHO

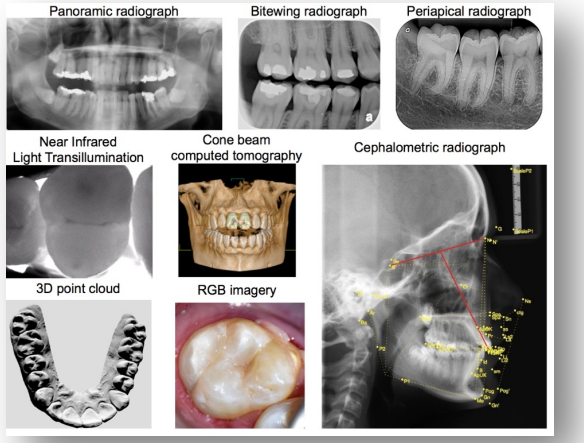
The International Telecommunication Union (ITU) and the World Health Organization (WHO) collaborate to establish an open group of experts to develop a generalizable benchmarking framework for health solutions based on artificial intelligence (AI), the ITU/WHO Focus Group on AI for Health (FG-AI4H).

What is AI benchmarking?

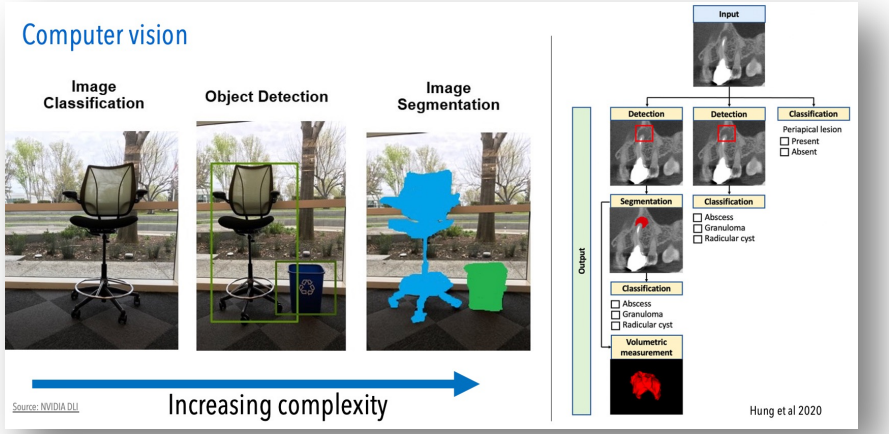
AI for Health
ITU-WHO Focus Group

"Evaluate performance based on the success of a model's utility within some practical applications."

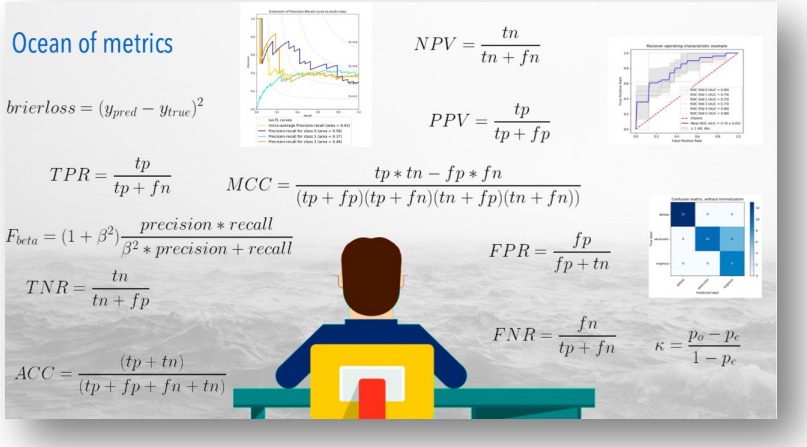
(Raji et al. 2021)



A ***dataset*** *or sets of datasets (at least test data, sometimes also training data)*



A ***task****, which is a particular specification of a problem, as represented in the dataset.*



One or more ***metric(s)*** *to summarize system performance over some set or sets of tasks as a single number or score.*

Benchmarking – The AI researcher's daily bread

AI for Health
ITU-WHO Focus Group

Research Reports: Clinical

Benchmarking Deep Learning Models for Tooth Structure Segmentation

Journal of Dental Research
1–7

DOI: 10.1177/00220345221100169
journals.sagepub.com/home/jdr

L. Schneider[1,2], L. Arsiwala-Scheppach[1,2], J. Krois[1,2], H. Meyer-Lueckel[3], K.K. Bressem[4,5], S.M. Niehues[4], and F. Schwendicke[1,2]

Tufts Dental Database: A Multimodal Panoramic X-Ray Dataset for Benchmarking Diagnostic Systems

Karen Panetta, *Fellow, IEEE*, Rahul Rajendran, *Student Member, IEEE*, Aruna Ramesh, Shishir Paramathma Rao, *Student Member, IEEE*, and Sos Agaian, *Fellow, IEEE*

Model	CheXpert AUC	#Params (M)
DenseNet121	0.859 (0.846, 0.871)	6.968
DenseNet169	0.860 (0.848, 0.873)	12.508
DenseNet201	0.864 (0.850, 0.876)	18.120
EfficientNetB0	0.859 (0.846, 0.871)	4.025
EfficientNetB1	0.858 (0.844, 0.872)	6.531
EfficientNetB2	0.866 (0.853, 0.880)	7.721
EfficientNetB3	0.853 (0.837, 0.867)	10.718
InceptionV3	0.862 (0.848, 0.876)	27.161
InceptionV4	0.861 (0.846, 0.873)	42.680
MNASNet	0.858 (0.845, 0.871)	5.290
MobileNetV2	0.854 (0.839, 0.869)	2.242
MobileNetV3	0.859 (0.847, 0.872)	4.220
ResNet101	0.863 (0.848, 0.876)	44.549
ResNet18	0.862 (0.847, 0.875)	11.690
ResNet34	0.863 (0.849, 0.875)	21.798
ResNet50	0.859 (0.843, 0.871)	25.557

Table 1: CheXpert AUC (with 95% Confidence Intervals) and Number of Parameters for 16 ImageNet-Pretrained Models.

Ke et al. 2021

Why is AI benchmarking for general applicability so hard?

AI for Health
ITU-WHO Focus Group

The data problem

- **Limited Scope** – Ultimately closed systems
- **Subjectivity** – Embedded biases hinder neutral or universal datasets
- **Ground truth establishment** – Complexity and difficulty of the clinical task yields noise and fuzziness in the labels/annotations

The task problem

input space ←→ output space

The metric problem

Why is AI benchmarking for general applicability so hard?

AI for Health
ITU-WHO Focus Group

The data problem

- **Limited Scope** – Ultimately closed systems
- **Subjectivity** – Embedded biases hinder neutral or universal datasets
- **Ground truth establishment** – Complexity and difficulty of the clinical task yields noise and fuzziness in the labels/annotations

The task problem

input space ←→ output space

- **Construct validity** – How well the benchmark dataset and associated metrics of evaluation represents a task?
- **Samples of convenience** – Tasks and collections of tasks arbitrarily built out of what is easily available to the team developing these benchmarks
- **(Clinical) usefulness** – Are defined tasks relevant for the discipline?

The metric problem

Why is AI benchmarking for general applicability so hard?

AI for Health
ITU-WHO Focus Group

The data problem

- **Limited Scope** – Ultimately closed systems
- **Subjectivity** – Embedded biases hinder neutral or universal datasets
- **Ground truth establishment** – Complexity and difficulty of the clinical task yields noise and fuzziness in the labels/annotations

The task problem

input space ←→ output space

- **Construct validity** – How well the benchmark dataset and associated metrics of evaluation represents a task?
- **Samples of convenience** – Tasks and collections of tasks arbitrarily built out of what is easily available to the team developing these benchmarks
- **(Clinical) usefulness** – Are defined tasks relevant for the discipline?

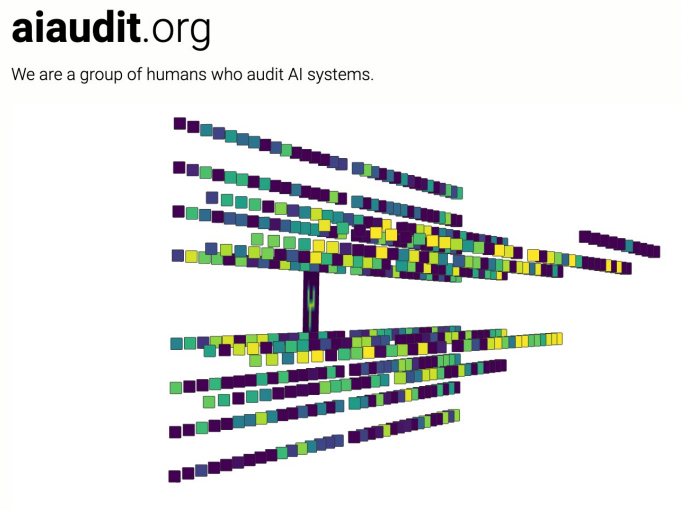
The metric problem

- **Inappropriate use**
  – Uncritically chasing algorithmic improvement as measured by the metric
  – Focusing on empirical and incremental work rather than hypothesis-based scientific inquiry
  – Aggregate performance evaluation tells little about generalization or impact and hinders meaningful interpretation
  – (Clinical) usefulness is not captured by the applied metrics

Alternative AI benchmarking techniques

AI for Health
ITU-WHO Focus Group

- Systematic development of test items (test suites, audits, adversarial testing)
- Output analysis and behavioural testing (error analysis, disaggregated analysis, counterfactual analysis)
- Ablation testing
- Analysing model properties (energy consumption, memory requirements, stability in the face of perturbations to training data)



AI for Health
ITU-WHO Focus Group

Evaluating AI models in health

The ITU/WHO Focus Group on artificial intelligence for health (FG-AI4H) works in partnership with the World Health Organization (WHO) to establish a standardized assessment framework for the evaluation of AI-based methods for health, diagnosis, triage or treatment decisions. The group was established by ITU-T Study Group 16 at its meeting in Ljubljana, Slovenia, 9-20 July 2018.

Tuesday, Sep 20, 1:20 – 2:40 (EEST)
Biomedicum, Lecture Hall 1 (Sali 1)

Open Code Initiative
AI for Health
ITU-WHO Focus Group
World Health Organization

Conclusion

AI for Health
ITU-WHO Focus Group

- The effective development of benchmarks is critical to progress of Dental AI research and to the value of Dental AI applications.
- Benchmarking helps us to understand how Dental AI systems work–and how they don't.
- Raise awareness for the shortcomings of benchmarking results related to datasets, task design and the applied metric(s).
- Alternative AI benchmarking techniques should be further explored and leveraged when feasible.

Conclusion

AI for Health
ITU-WHO Focus Group

- The effective development of benchmarks is critical to progress of Dental AI research and to the value of Dental AI applications.
- Benchmarking helps us to understand how Dental AI systems work–and how they don't.
- Raise awareness for the shortcomings of benchmarking results related to datasets, task design and the applied metric(s).
- Alternative AI benchmarking techniques should be further explored and leveraged when feasible.

However, universal and neutral (dental) datasets don't exist (yet), and current methods of benchmarking do not (yet) offer meaningful measures of general applicability.

ITUTech

AI for Health

ITU-WHO Focus Group

HTTPS://REMOTE.ITU.INT

itu.int/go/fgai4h

World Health Organization

ITU

Urgency contact
hamdanai4h@gmail.com