# Artificial Intelligence in Medicine: The Need to Translate From Basic Science to Clinical Value

William Hersh, MD
Professor and Chair
Department of Medical Informatics & Clinical Epidemiology
School of Medicine
Oregon Health & Science University
Portland, OR, USA
http://www.ohsu.edu/informatics
Email: hersh@ohsu.edu
Web: www.billhersh.info
Blog: https://informaticsprofessor.blogspot.com/
Twitter: @williamhersh

References

Attia, Z.I., Friedman, P.A., Noseworthy, P.A., Lopez-Jimenez, F., Ladewig, D.J., Satam, G., Pellikka, P.A., Munger, T.M., Asirvatham, S.J., Scott, C.G., Carter, R.E., Kapa, S., 2019. Age and Sex Estimation Using Artificial Intelligence From Standard 12-Lead ECGs. Circ Arrhythm Electrophysiol 12, e007284. https://doi.org/10.1161/CIRCEP.119.007284

Bejnordi, B.E., Zuidhof, G., Balkenhol, M., Hermsen, M., Bult, P., van Ginneken, B., Karssemeijer, N., Litjens, G., van der Laak, J., 2017. Context-aware stacked convolutional neural networks for classification of breast carcinomas in whole-slide histopathology images. J Med Imaging (Bellingham) 4, 044504. https://doi.org/10.1117/1.JMI.4.4.044504

Coiera, E., Tong, H.L., 2021. Replication studies in the clinical decision support literature-frequency, fidelity, and impact. J Am Med Inform Assoc 28, 1815–1825. https://doi.org/10.1093/jamia/ocab049

DeGrave, A.J., Janizek, J.D., Lee, S.-I., 2021. AI for radiographic COVID-19 detection selects shortcuts over signal. Nat Mach Intell 3, 610–619. https://doi.org/10.1038/s42256-021-00338-7

Esteva, A., Chou, K., Yeung, S., Naik, N., Madani, A., Mottaghi, A., Liu, Y., Topol, E., Dean, J., Socher, R., 2021. Deep learning-enabled medical computer vision. npj Digital Medicine 4, 1–9. https://doi.org/10.1038/s41746-020-00376-2

Esteva, A., Kuprel, B., Novoa, R.A., Ko, J., Swetter, S.M., Blau, H.M., Thrun, S., 2017. Dermatologist-level classification of skin cancer with deep neural networks. Nature 542, 115–118. https://doi.org/10.1038/nature21056

Freeman, K., Geppert, J., Stinton, C., Todkill, D., Johnson, S., Clarke, A., Taylor-Phillips, S., 2021. Use of artificial intelligence for image analysis in breast cancer screening programmes: systematic review of test accuracy. BMJ 374, n1872. https://doi.org/10.1136/bmj.n1872

Gulshan, V., Peng, L., Coram, M., Stumpe, M.C., Wu, D., Narayanaswamy, A., Venugopalan, S., Widner, K., Madams, T., Cuadros, J., Kim, R., Raman, R., Nelson, P.C., Mega, J.L., Webster, D.R., 2016. Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. JAMA 316, 2402–2410. https://doi.org/10.1001/jama.2016.17216

Haenssle, H.A., Fink, C., Schneiderbauer, R., Toberer, F., Buhl, T., Blum, A., Kalloo, A., Hassen, A.B.H., Thomas, L., Enk, A., Uhlmann, L., Reader study level-I and level-II Groups, Alt, C., Arenbergerova, M., Bakos, R., Baltzer, A., Bertlich, I., Blum, Andreas, Bokor-Billmann, T., Bowling, J., Braghiroli, N., Braun, R., Buder-Bakhaya, K., Buhl, Timo, Cabo, H., Cabrijan, L., Cevic, N., Classen, A., Deltgen, D., Fink, Christine, Georgieva, I., Hakim-Meibodi, L.-E., Hanner, S., Hartmann, F., Hartmann, J., Haus, G., Hoxha, E., Karls, R., Koga, H., Kreusch, J., Lallas, A., Majenka, P., Marghoob, A., Massone, C., Mekokishvili, L., Mestel, D., Meyer, V., Neuberger, A., Nielsen, K., Oliviero, M., Pampena, R., Paoli, J., Pawlik, E., Rao, B., Rendon, A., Russo, T., Sadek, A., Samhaber, K., Schneiderbauer, Roland, Schweizer, A., Toberer, Ferdinand, Trennheuser, L., Vlahova, L., Wald, A., Winkler, J., Wölbing, P., Zalaudek, I., 2018. Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. Ann Oncol 29, 1836–1842. https://doi.org/10.1093/annonc/mdy166

Lakhani, P., Sundaram, B., 2017. Deep Learning at Chest Radiography: Automated Classification of Pulmonary Tuberculosis by Using Convolutional Neural Networks. Radiology 284, 574–582. https://doi.org/10.1148/radiol.2017162326

McDermott, M.B.A., Wang, S., Marinsek, N., Ranganath, R., Foschini, L., Ghassemi, M., 2021. Reproducibility in machine learning for health research: Still a ways to go. Sci Transl Med 13, eabb1655. https://doi.org/10.1126/scitranslmed.abb1655

Poplin, R., Varadarajan, A.V., Blumer, K., Liu, Y., McConnell, M.V., Corrado, G.S., Peng, L., Webster, D.R., 2018. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. Nat Biomed Eng 2, 158–164. https://doi.org/10.1038/s41551-018-0195-0

Rajkomar, A., Oren, E., Chen, K., Dai, A.M., Hajaj, N., Hardt, M., Liu, P.J., Liu, X., Marcus, J., Sun, M., Sundberg, P., Yee, H., Zhang, K., Zhang, Y., Flores, G., Duggan, G.E., Irvine, J., Le, Q., Litsch, K., Mossin, A., Tansuwan, J., Wang, D., Wexler, J., Wilson, J., Ludwig, D., Volchenboum, S.L., Chou, K., Pearson, M., Madabushi, S., Shah, N.H., Butte, A.J., Howell, M.D., Cui, C., Corrado, G.S., Dean, J., 2018. Scalable and accurate deep learning with electronic health records. npj Digital Medicine 1, 1–10. https://doi.org/10.1038/s41746-018-0029-1

Rajpurkar, P., Irvin, J., Ball, R.L., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D., Bagul, A., Langlotz, C.P., Patel, B.N., Yeom, K.W., Shpanskaya, K., Blankenberg, F.G., Seekins, J., Amrhein, T.J., Mong, D.A., Halabi, S.S., Zucker, E.J., Ng, A.Y., Lungren, M.P., 2018. Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists. PLoS Med 15, e1002686. https://doi.org/10.1371/journal.pmed.1002686

Ravizza, S., Huschto, T., Adamov, A., Böhm, L., Büsser, A., Flöther, F.F., Hinzmann, R., König, H., McAhren, S.M., Robertson, D.H., Schleyer, T., Schneidinger, B., Petrich, W., 2019. Predicting the early risk of chronic kidney disease in patients with diabetes using real-world data. Nat Med 25, 57–59. https://doi.org/10.1038/s41591-018-0239-8

Roberts, M., Driggs, D., Thorpe, M., Gilbey, J., Yeung, M., Ursprung, S., Aviles-Rivero, A.I., Etmann, C., McCague, C., Beer, L., Weir-McCall, J.R., Teng, Z., Gkrania-Klotsas, E., Rudd, J.H.F., Sala, E., Schönlieb, C.-B., 2021. Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. Nat Mach Intell 3, 199–217. https://doi.org/10.1038/s42256-021-00307-0

Rodriguez, V.A., Bhave, S., Chen, R., Pang, C., Hripcsak, G., Sengupta, S., Elhadad, N., Green, R., Adelman, J., Metitiri, K.S., Elias, P., Groves, H., Mohan, S., Natarajan, K., Perotte, A., 2021. Development and validation of prediction models for mechanical ventilation, renal replacement

therapy, and readmission in COVID-19 patients. J Am Med Inform Assoc. https://doi.org/10.1093/jamia/ocab029

Ting, D.S.W., Cheung, C.Y.-L., Lim, G., Tan, G.S.W., Quang, N.D., Gan, A., Hamzah, H., Garcia-Franco, R., San Yeo, I.Y., Lee, S.Y., Wong, E.Y.M., Sabanayagam, C., Baskaran, M., Ibrahim, F., Tan, N.C., Finkelstein, E.A., Lamoureux, E.L., Wong, I.Y., Bressler, N.M., Sivaprasad, S., Varma, R., Jonas, J.B., He, M.G., Cheng, C.-Y., Cheung, G.C.M., Aung, T., Hsu, W., Lee, M.L., Wong, T.Y., 2017. Development and Validation of a Deep Learning System for Diabetic Retinopathy and Related Eye Diseases Using Retinal Images From Multiethnic Populations With Diabetes. JAMA 318, 2211–2223. https://doi.org/10.1001/jama.2017.18152

Tschandl, P., Rosendahl, C., Akay, B.N., Argenziano, G., Blum, A., Braun, R.P., Cabo, H., Gourhant, J.-Y., Kreusch, J., Lallas, A., Lapins, J., Marghoob, A., Menzies, S., Neuber, N.M., Paoli, J., Rabinovitz, H.S., Rinner, C., Scope, A., Soyer, H.P., Sinz, C., Thomas, L., Zalaudek, I., Kittler, H., 2019. Expert-Level Diagnosis of Nonpigmented Skin Cancer by Combined Convolutional Neural Networks. JAMA Dermatol 155, 58–65. https://doi.org/10.1001/jamadermatol.2018.4378

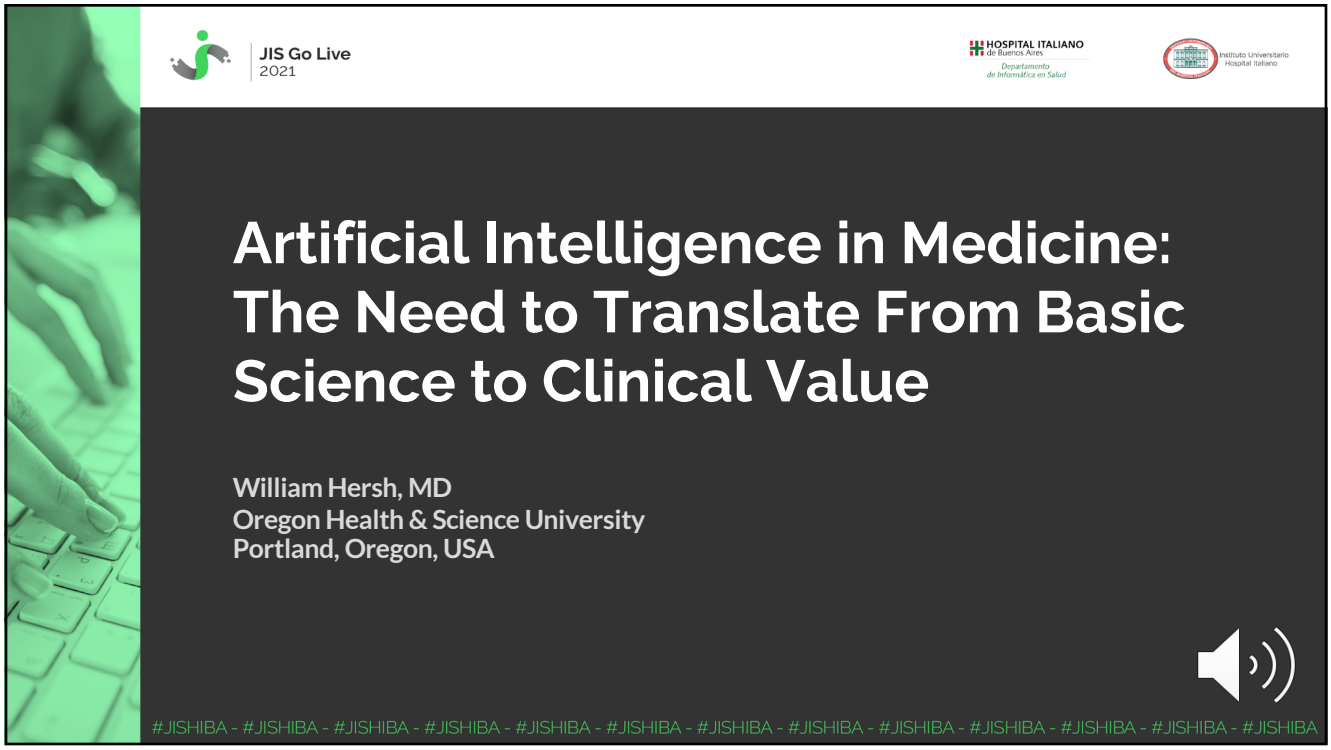
Veta, M., Heng, Y.J., Stathonikos, N., Bejnordi, B.E., Beca, F., Wollmann, T., Rohr, K., Shah, M.A., Wang, D., Rousson, M., Hedlund, M., Tellez, D., Ciompi, F., Zerhouni, E., Lanyi, D., Viana, M., Kovalev, V., Liauchuk, V., Phoulady, H.A., Qaiser, T., Graham, S., Rajpoot, N., Sjöblom, E., Molin, J., Paeng, K., Hwang, S., Park, S., Jia, Z., Chang, E.I.-C., Xu, Y., Beck, A.H., van Diest, P.J., Pluim, J.P.W., 2019. Predicting breast tumor proliferation from whole-slide images: The TUPAC16 challenge. Med Image Anal 54, 111–121. https://doi.org/10.1016/j.media.2019.02.012

Wang, L., Sha, L., Lakin, J.R., Bynum, J., Bates, D.W., Hong, P., Zhou, L., 2019. Development and Validation of a Deep Learning Algorithm for Mortality Prediction in Selecting Patients With Dementia for Earlier Palliative Care Interventions. JAMA Netw Open 2, e196972. https://doi.org/10.1001/jamanetworkopen.2019.6972

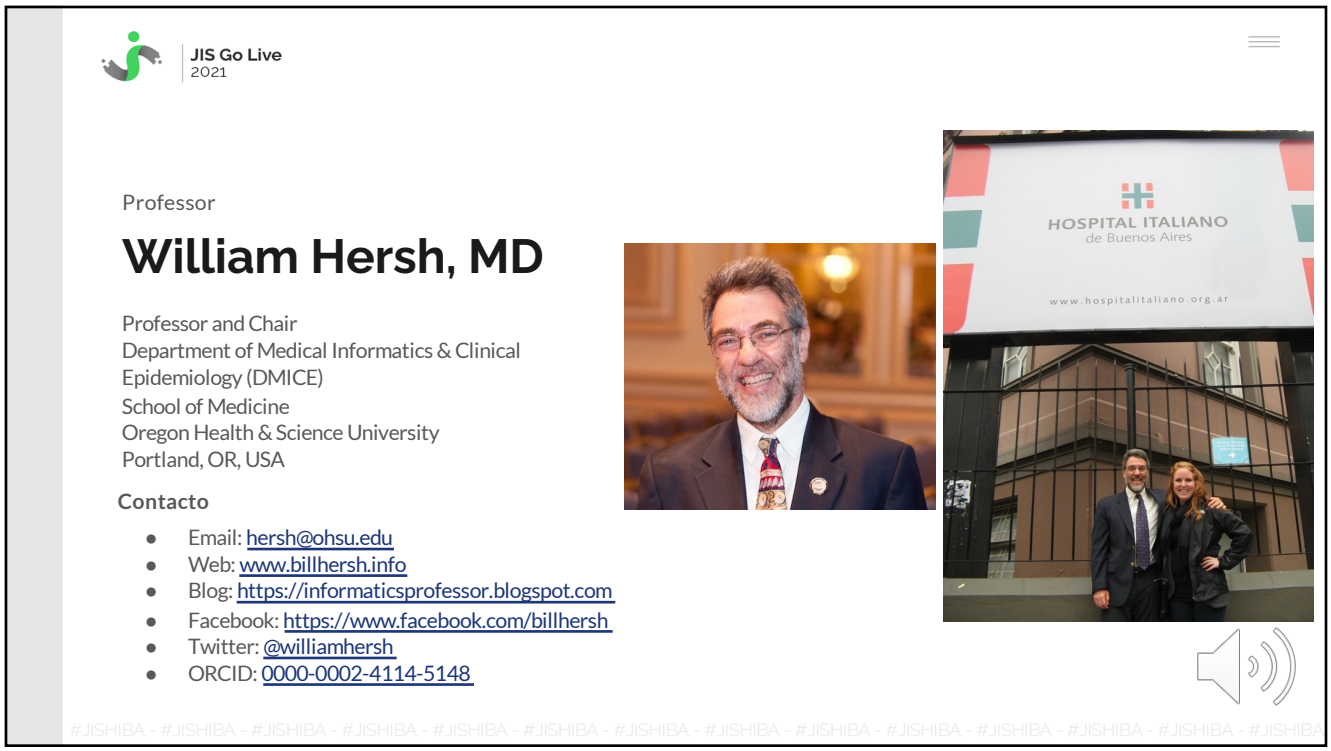
Wu, E., Wu, K., Daneshjou, R., Ouyang, D., Ho, D.E., Zou, J., 2021. How medical AI devices are evaluated: limitations and recommendations from an analysis of FDA approvals. Nat Med 27, 582–584. https://doi.org/10.1038/s41591-021-01312-x

Wynants, L., Van Calster, B., Collins, G.S., Riley, R.D., Heinze, G., Schuit, E., Bonten, M.M.J., Dahly, D.L., Damen, J.A.A., Debray, T.P.A., de Jong, V.M.T., De Vos, M., Dhiman, P., Haller, M.C., Harhay, M.O., Henckaerts, L., Heus, P., Kammer, M., Kreuzberger, N., Lohmann, A., Luijken, K., Ma, J., Martin, G.P., McLernon, D.J., Andaur Navarro, C.L., Reitsma, J.B., Sergeant, J.C., Shi, C., Skoetz, N., Smits, L.J.M., Snell, K.I.E., Sperrin, M., Spijker, R., Steyerberg, E.W., Takada, T., Tzoulaki, I., van Kuijk, S.M.J., van Bussel, B., van der Horst, I.C.C., van Royen, F.S., Verbakel, J.Y., Wallisch, C., Wilkinson, J., Wolff, R., Hooft, L., Moons, K.G.M., van Smeden, M., 2020. Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. BMJ 369, m1328. https://doi.org/10.1136/bmj.m1328

Zhou, Q., Chen, Z.-H., Cao, Y.-H., Peng, S., 2021. Clinical impact and quality of randomized controlled trials involving interventions evaluating artificial intelligence prediction tools: a systematic review. NPJ Digit Med 4, 154. https://doi.org/10.1038/s41746-021-00524-2


JIS Go Live
2021
HOSPITAL ITALIANO
de Buenos Aires
Departamento
de Informática en Salud
Instituto Universitario
Hospital Italiano
Artificial Intelligence in Medicine: The Need to Translate From Basic Science to Clinical Value
William Hersh, MD
Oregon Health & Science University
Portland, Oregon, USA
#JISHIBA - #JISHIBA - #JISHIBA - #JISHIBA - #JISHIBA - #JISHIBA - #JISHIBA - #JISHIBA - #JISHIBA - #JISHIBA - #JISHIBA - #JISHIBA

1


JIS Go Live
2021
Professor
William Hersh, MD
Professor and Chair
Department of Medical Informatics & Clinical Epidemiology (DMICE)
School of Medicine
Oregon Health & Science University
Portland, OR, USA
Contacto
Email: hersh@ohsu.edu
Web: www.billhersh.info
Blog: https://informaticsprofessor.blogspot.com
Facebook: https://www.facebook.com/billhersh
Twitter: @williamhersh
ORCID: 0000-0002-4114-5148
HOSPITAL ITALIANO
de Buenos Aires
www.hospitalitaliano.org.ar
#JISHIBA - #JISHIBA - #JISHIBA - #JISHIBA - #JISHIBA - #JISHIBA - #JISHIBA - #JISHIBA - #JISHIBA - #JISHIBA - #JISHIBA - #JISHIBA

2

JIS Go Live 2021

# Outline

- Definitions
- Results
- Challenges



3

JIS Go Live 2021

# Artificial intelligence and machine learning

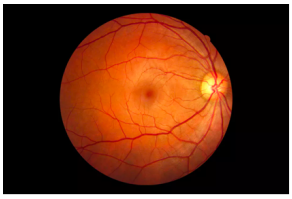
- Artificial intelligence (AI) is ability of computers to perform tasks normally associated with human intelligence
- Modern AI has been greatly advanced by machine learning (ML), which is ability of computer programs to learn without being explicitly programmed
- ML has been advanced by new algorithms, enhanced computer processing abilities, and availability of large and diverse quantities of data
    - Most important has been deep learning (DL), involving use of multi-layered neural networks
- Don't need to understand math or theory to understand use or evaluate clinical benefit

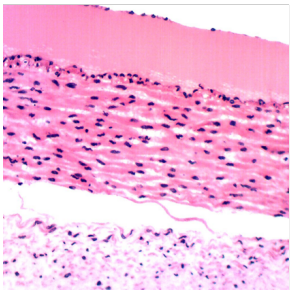


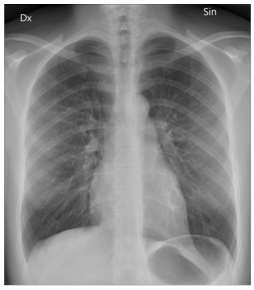
4

JIS Go Live 2021

# Most success has come from image classification








Early studies

- Diabetic retinopathy (DR) (Gulshan, 2016; Ting, 2017)
- Histology of cancer (Benjordi, 2017) and metastases (Veta, 2019)
- Tuberculosis (Lakhani, 2017) and pneumonia (Rajpurkar, 2018)
- Skin cancer (Esteva, 2017; Haenssle, 2018; Tschandi, 2019)

State of the art (Esteva, 2021)

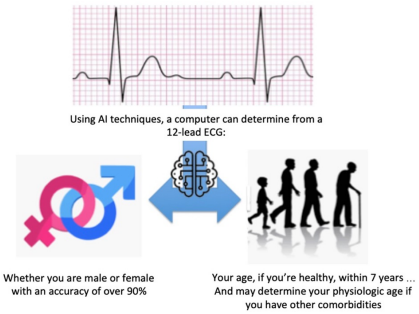


5

JIS Go Live 2021

# Other successes in clinical prediction

- Length of stay, mortality, readmission, and diagnosis at two large medical centers (Rajkomar, 2018)
- Age and sex determination from retinal images (Poplin, 2018) or ECG (Attia, 2019)
- Early risk of chronic kidney disease in patients with diabetes (Ravizza, 2019)
- Dementia from EHR data up to two years before clinical diagnosis (Wang, 2019)
- Prediction models for mechanical ventilation, renal replacement therapy, and readmission in COVID-19 (Rodriguez, 2021)






6

JIS Go Live 2021

# What is the evidence for the benefit of AI?

- Best evidence for interventions (treatment or prevention) comes from randomized controlled trials (RCTs)
  - Ideally RCTs that are well-conducted, generalizable, and well-reported
- There are other clinical questions that can be answered by AI
  - Diagnosis – can AI methods improve ability to diagnose disease?
  - Harm – can AI identify harms from environment, medical care, etc.?
  - Prognosis – can AI inform the prognosis of health and disease?
- Ultimately, however, AI interventions must be demonstrated experimentally to benefit patients, clinicians, and populations
  - Some instances when RCTs are infeasible so observational studies may be justified



7

JIS Go Live 2021

# Systematic review of clinical interventions

- Zhou, Q., Chen, Z.-H., Cao, Y.-H., Peng, S., 2021. Clinical impact and quality of randomized controlled trials involving interventions evaluating artificial intelligence prediction tools: a systematic review. *NPJ Digit Med* 4, 154. https://doi.org/10.1038/s41746-021-00524-2
- Review of all randomized controlled trials (RCTs) using
  - Traditional statistical (TS) – mostly regression
  - Machine learning (ML) – all but deep learning
  - Deep learning (DL) – neural networks
- Found use for
  - Assistive treatment decisions
  - Assistive diagnosis
  - Risk stratification



8

JIS Go Live 2021

Identified 65 RCTs with following characteristics

- 61.5% positive results
- Variety of disease categories – cancer, other chronic disease, acute disease, and primary care
- Types of algorithms – TS > ML > DL
- Predictive tool function – assistive treatment decisions > assistive diagnosis > risk stratification

Some concerns of bias in studies

- One-third no sample size estimation
- Three-fourths no masking (open-label)
- Majority did not reference CONSORT, use intent-to-treat analysis, or provide study protocol

- Caveat: number of positive studies does not necessarily indicate general superiority of methods

**Table 1.** General characteristics of the 65 randomized controlled trials.

| Variables | Levels | Total (*n* = 65) |
|---|---|---|
| Results (%) | Negative | 25 (38.5) |
| | Positive | 40 (61.5) |
| Duration of study (*n*[a] = 59, months, median [IQR]) | | 12 [6, 24] |
| Sample size (median [IQR]) | | 435 [192, 999] |
| Sample size estimation (%) | Larger or equal than expected | 37 (56.9) |
| | Less than expected | 7 (10.8) |
| | Not performed | 21 (32.3) |
| Publication year (%) | 2010–2015 | 21 (32.3) |
| | 2016–2020 | 44 (67.7) |
| Study design (%) | RCT superiority (individualized) | 48 (73.8) |
| | RCT superiority with crossover (individualized) | 1 (1.5) |
| | RCT non-inferiority (individualized) | 2 (3.1) |
| | Clustered RCT superiority (clustered) | 7 (10.8) |
| | Stepped-wedge design (clustered) | 7 (10.8) |
| Allocation ratio (%) | 1:1 parallel | 55 (84.6) |
| | Others | 10 (15.4) |
| Masking (%) | Open-label | 49 (75.4) |
| | Single-blinded | 12 (18.5) |
| | Double-blinded | 4 (6.2) |
| Centers (%) | Single | 33 (50.8) |
| | Multi | 32 (49.2) |
| Disease category (%) | Cancer | 11 (16.9) |
| | Chronic disease not included cancer | 18 (27.7) |
| | Acute disease | 19 (29.2) |
| | Primary care | 9 (13.8) |
| | Others | 8 (12.3) |
| Types of algorithms (%) | Traditional statistical model | 37 (56.9) |
| | Machine learning | 17 (26.2) |
| | Deep learning | 11 (16.9) |
| Prediction tools function (%) | Assistive treatment decision | 35 (53.8) |
| | Assistive diagnosis | 16 (24.6) |
| | Risk stratification | 12 (18.5) |
| | Others | 2 (3.1) |
| Referenced CONSORT (%) | No | 47 (72.3) |
| | Yes | 18 (27.7) |
| Intent-to-treat analysis (%) | No | 39 (60.0) |
| | Yes | 26 (40.0) |
| Study protocol available | No | 49 (75.4) |
| | Yes | 16 (24.6) |
| Model development (%) | No | 7 (10.8) |
| | Yes—independent publication | 49 (75.4) |
| | Yes—published in the same article with RCT | 9 (13.8) |
| Internal validation (%) | No | 23 (35.4) |
| | Yes | 42 (64.6) |
| External validation (%) | No | 25 (38.5) |
| | Yes | 40 (61.5) |
| AUC in model development (*n*[a] = 21, median [IQR]) | | 0.81 [0.75, 0.91] |
| AUC in internal validation (*n*[a] = 18, median [IQR]) | | 0.83 [0.7, 0.91] |
| AUC in external validation (*n*[a] = 26, median [IQR]) | | 0.74 [0.75, 0.77] |

*IQR* interquartile range, *AUC* area under the receiver operating characteristic curve.
[a]Available numbers used for description



9

JIS Go Live 2021

Characteristics by tool type varied

- Model input – clinical quantitative data for TS/ML, images for DL, little use of natural language
- Disease category – varied for TS, chronic disease for ML, cancer for DL
- Tool function – risk stratification and treatment for TS, treatment for ML, diagnosis for DL
- Results – mixed for TS, more positive for ML/DL

| Variables | Levels | TS (*n* = 37) | ML (*n* = 17) | DL (*n* = 11) | *P* value |
|---|---|---|---|---|---|
| Duration of study (*n* = 59, months, median [IQR]) | | 17 [8, 32] | 7 [4, 19] | 6 [4, 9] | 0.005 |
| Sample size (median [IQR]) | | 435 [194, 999] | 258 [90, 537] | 700 [548, 994] | 0.122 |
| Clinical settings (%) | Outpatients | 19 (51.4) | 6 (35.3) | 1 (9.1) | 0.015 |
| | Inpatients | 17 (45.9) | 8 (47.1) | 10 (90.9) | |
| | Home | 1 (2.7) | 3 (17.6) | 0 (0.0) | |
| Publication year (%) | 2010–2015 | 14 (37.8) | 7 (41.2) | 0 (0.0) | 0.041 |
| | 2016–2020 | 23 (62.2) | 10 (58.8) | 11 (100.0) | |
| Model input (%) | Clinical quantitative data | 36 (97.3) | 16 (94.1) | 0 (0.0) | <0.001 |
| | Images or videos | 1 (2.7) | 0 (0.0) | 10 (90.9) | |
| | Natural language | 0 (0.0) | 1 (5.9) | 1 (9.1) | |
| Disease category (%) | Cancer | 2 (5.4) | 0 (0.0) | 9 (81.8) | <0.001 |
| | Chronic disease | 4 (10.8) | 13 (76.5) | 1 (9.1) | |
| | Acute disease | 16 (43.2) | 2 (11.8) | 1 (9.1) | |
| | Primary care | 9 (24.3) | 0 (0.0) | 0 (0.0) | |
| | Others | 6 (16.2) | 2 (11.8) | 0 (0.0) | |
| Prediction tools function (%) | Assistive diagnosis | 3 (8.1) | 2 (11.8) | 11 (100.0) | <0.001 |
| | Risk stratification | 11 (29.7) | 1 (5.9) | 0 (0.0) | |
| | Assistive treatment decision | 22 (59.5) | 13 (76.5) | 0 (0.0) | |
| | Others | 1 (2.7) | 1 (5.9) | 0 (0.0) | |
| Results (%) | Negative | 18 (48.6) | 5 (29.4) | 2 (18.2) | 0.136 |
| | Positive | 19 (51.4) | 12 (70.6) | 9 (81.8) | 0.044 (*P* for trend) |

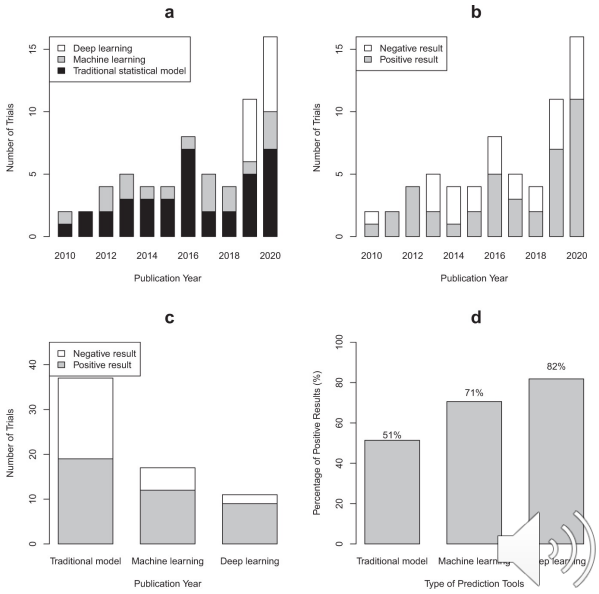


10

JIS Go Live 2021

By publication year
- Increasing per year
- Increasing DL per year

By tool type, more positive for DL > ML > TS





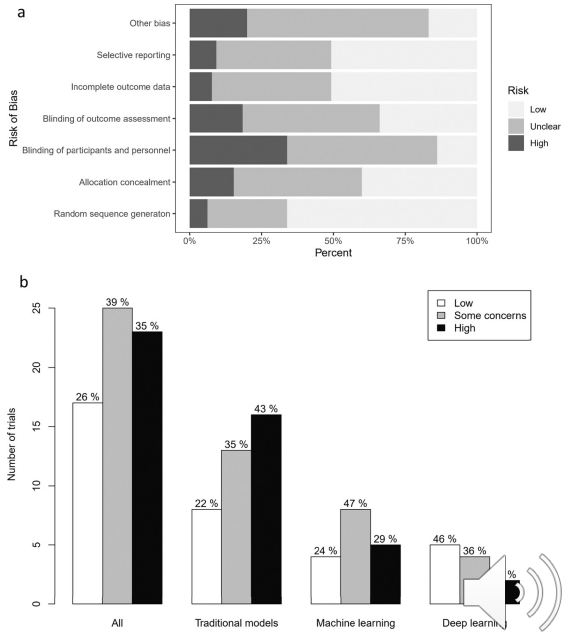

11

JIS Go Live 2021

Only 17 of 65 trials with low risk of bias

Risk of bias high or unclear for most studies – higher for TS > ML > DL

Suboptimal use of CONSORT, sample size pre-estimation, randomization, and intent-to-treat analysis






12

JIS Go Live 2021

Characteristics of DL trials

- Of 11 RCTs, 9 evaluate assisting endoscopy – all positive results
- 2 other RCTs have negative results

**Table 2. Procedures of predictive tool interventions in the eleven randomized controlled trials involving interventions evaluating deeplearning tools**

| Reference | Conditions | Sample size | Tools for intervention | Control | Algorithms | Tool function | Tool input | Tool output | How the output being used in clinical settings | Trial primary outcomes | Gold standard | Trial findings |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Chen 2019 | Upper gastrointestinal lesions | 437 | Routine EGD examination stratified by three types with the assistance of ENDOANGEL AI system | Routine EGD examination stratified by three types without AI | DCNN (VGG-16) | Assistive diagnosis | EGD images | A virtual stomach model monitoring blind spots; timing; scoring and grading | Experts referenced AI output to make EGD examination and monitor blind spots. | Mean blind spot rate | Experts | Positive |
| Lin 2019 | Childhood cataracts | 700 | CC-cruiser web diagnosis platform | Regular ophthalmic diagnosis | DCNN (ImageNet) | Assistive diagnosis | Ocular images from slit-lamp photography | Diagnosis outcome; comprehensive evaluation; treatment recommendation | AI made diagnosis independently, and its results would be comparted with experts and not impact clinical decision making. | Accuracy of diagnosis | Experts | Negative |
| Su 2019 | Colorectal cancer | 659 | Routine colonoscopies with the assistance of an AI automatic quality control system | Routine colonoscopies | DCNN (AlexNet, ZFNet, YOLO V2) | Assistive diagnosis | Colonoscopy images | Location of colorectal polyps; timing; reminding retest and clean | Endoscopists referenced AI output to make endoscopic examination and report of polyps and adenomas. | Adenoma detection rate | Pathology | Positive |
| Wang 2019 | Colorectal cancer | 1058 | Routine colonoscopies with the assistance of an automatic polyp detection system | Routine colonoscopies | Deep learning architecture | Assistive diagnosis | Colonoscopy images | Location of polyps; alarming | Endoscopists were required to check every polyp location detected by the system and report of polyps and adenomas. | Adenoma detection rate | Pathology | Positive |
| Wu 2019 | Upper gastrointestinal lesions | 303 | Routine EGD examination with the assistance of WISENSE AI system | Routine EGD examination | DCNN (VGG-16 and DenseNet) | Assistive diagnosis | EGD images | A virtual stomach model monitoring blind spots; timing; scoring and grading; extracting frames with the highest confidence | Experts referenced AI output to make EGD examination and monitor blind spots. | Mean blind spot rate | Experts | Positive |
| Gong 2020 | Colorectal cancer | 704 | ENDOANGEL-assisted routine colonoscopy | Routine colonoscopy | DCNN and perceptual hash algorithms (VGG-16) | Assistive diagnosis | Colonoscopy images | Timing; safe, alarm, and dangerous ranges of withdrawal speed for real-time monitoring; slipping warning | Operating endoscopists referenced AI output to make endoscopic examination and report of polyps and adenomas. | Adenoma detection rate | Pathology | Positive |
| Liu 2020 | Colorectal cancer | 1026 | Routine colonoscopy with CADe assistance | Routine colonoscopy | DCNN-3D | Assistive diagnosis | Colonoscopy images | The probability of polyps in each frame; lesions alarming | Endoscopists focused mainly on the main monitor during the examination process, and a voice alarm prompted them to view the system monitor to check the location of each polyp detected by the system. | Detection rate of polyps and adenomas | Pathology | Positive |
| Luo 2020 | Colorectal cancer | 157 | AI-assisted colonoscopy | Traditional colonoscopy | CNN (YOLO) | Assistive diagnosis | Colonoscopy images | Location of polyps | Endoscopists referenced AI output to make endoscopic examination and report of polyps. | Polyp detection rate | Not reported | Positive |
| Repici 2020 | Colorectal cancer | 685 | High-definition colonoscopies with the AI-based CADe system | Routine colonoscopy | CNN | Assistive diagnosis | Colonoscopy images | Location of polys | Endoscopists referenced AI output to make endoscopic examination and report of polyps and adenomas. | Adenoma detection rate | Pathology | Positive |
| Wang 2020 | Colorectal cancer | 962 | White light colonoscopy with assistance from the CADe system | White light colonoscopy with assistance from a sham system | Deep learning architecture | Assistive diagnosis | Colonoscopy images | Location of polyps; alarming | Endoscopists were required to check every polyp location detected by the system and report of polyps and adenomas. | Adenoma detection rate | Pathology | Positive |
| Blomberg 2021 | Out-of-hospital cardiac arrest (OHCA) | 5242 | Normal protocols with alert | Normal protocols without alert | Speech recognition using deep neural networks | Assistive diagnosis | Emergency calls | OHCA Alert | Dispatchers in the intervention group were alerted when the machine learning model identified out-of-hospital cardiac arrest. | The rate of dispatcher recognition of subsequently confirmed OHCA | Danish Cardiac [illegible] | egative |

Abbreviations: AI = Artificial intelligence; DL = Tools using deep learning algorithms; ML = Tools using machine learning algorithms; CNN = Convolutionalneural networks; DCNN = Deep convolutional neural networks; CADe = Computer-aided detection; EGD = Esephagogastroduodenoscopy; OHCA = Out-of-hos[illegible]cardiac a



13

JIS Go Live 2021

# Conclusions

- AI predictive tools show great promise in improving clinical decisions for diagnosis, treatment, and risk stratification but comprehensive evidence lacking
  - Number of clinical trials assessing clinical benefit is small
  - Majority of the clinical trials have indeterminate or high risk of bias
  - Most trials of deep learning focused on endoscopic procedures
- Concerns about review
  - Missing column in Table 2 of DL interventions
    - Does not include Yao et al. 2021 – published after review done?
  - Difficult to use data in Supp Table 4 of ML interventions
    - Includes Wijnberge et al. 2020 (62) but not in ML table – considered TS?
  - No data/table for TS interventions



14

JIS Go Live 2021

## Other concerns about AI algorithms

- CDS studies rarely (0.3%) replicated (Coiera, 2021)
- AI for radiographic COVID-19 detection selects non-signal features that may be idiosyncratic (DeGrave, 2021)
- In case of using AI for screening mammography, quality of studies is poor, mostly retrospective, and mostly inferior to radiologists (Freeman, 2021)
- Almost all published prediction models for COVID-19 prognosis poorly reported and at high risk of bias (Wynant, 2020)
- Studies of CXR used for diagnosis of COVID-19 had inadequate reporting of methodology (Roberts, 2021)
- Machine learning for health compared poorly to other areas regarding reproducibility metrics, such as dataset and code accessibility (McDermott, 2021)
- Of 130 medical AI devices approved by FDA, only 4 underwent prospective evaluation and number of evaluation sites and samples often not reported (Wu, 2021)



15

JIS Go Live 2021

## Summary

- Much promise, but clinically proven benefit still mostly lacking
- Need "translation" from "basic science" of AI to "clinical value"
  - Clinically-driven applications and trials
  - Attention to patient safety and benefit and to clinician workflow
  - Robust AI and RCT methods that are replicable and generalizable
- Most conduct implementation and trials from a health equity standpoint



16


JIS Go Live
2021
HOSPITAL ITALIANO
de Buenos Aires
Departamento
de Informática en Salud
Instituto Universitario
Hospital Italiano
Gracias!
#JISHIBA - #JISHIBA - #JISHIBA - #JISHIBA - #JISHIBA - #JISHIBA - #JISHIBA - #JISHIBA - #JISHIBA - #JISHIBA - #JISHIBA - #JISHIBA

17