Translational Artificial Intelligence (AI): 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
https://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

Balwani, M., Sardh, E., Ventura, P., Peiró, P.A., Rees, D.C., Stölzel, U., Bissell, D.M., Bonkovsky, H.L., Windyga, J., Anderson, K.E., Parker, C., Silver, S.M., Keel, S.B., Wang, J.-D., Stein, P.E., Harper, P., Vassiliou, D., Wang, B., Phillips, J., Ivanova, A., Langendonk, J.G., Kauppinen, R., Minder, E., Horie, Y., Penz, C., Chen, J., Liu, S., Ko, J.J., Sweetser, M.T., Garg, P., Vaishnaw, A., Kim, J.B., Simon, A.R., Gouya, L., ENVISION Investigators, 2020. Phase 3 Trial of RNAi Therapeutic Givosiran for Acute Intermittent Porphyria. N Engl J Med 382, 2289–2301. https://doi.org/10.1056/NEJMoa1913147

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

Campanella, G., Hanna, M.G., Geneslaw, L., Miraflor, A., Werneck Krauss Silva, V., Busam, K.J., Brogi, E., Reuter, V.E., Klimstra, D.S., Fuchs, T.J., 2019. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nature Medicine 25, 1301–1309. https://doi.org/10.1038/s41591-019-0508-1

Chamberlin, S.R., Bedrick, S.D., Cohen, A.M., Wang, Y., Wen, A., Liu, S., Liu, H., Hersh, W.R., 2020. Evaluation of patient-level retrieval from electronic health record data for a cohort discovery task. JAMIA Open 3, 395–404. https://doi.org/10.1093/jamiaopen/ooaa026

Chi, E.A., Chi, G., Tsui, C.T., Jiang, Y., Jarr, K., Kulkarni, C.V., Zhang, M., Long, J., Ng, A.Y., Rajpurkar, P., Sinha, S.R., 2021. Development and Validation of an Artificial Intelligence System to Optimize Clinician Review of Patient Records. JAMA Netw Open 4, e2117391. https://doi.org/10.1001/jamanetworkopen.2021.17391

Cohen, A.M., Chamberlin, S., Deloughery, T., Nguyen, M., Bedrick, S., Meninger, S., Ko, J.J., Amin, J.J., Wei, A.J., Hersh, W., 2020. Detecting rare diseases in electronic health records using machine learning and knowledge engineering: Case study of acute hepatic porphyria. PLoS ONE 15, e0235574. https://doi.org/10.1371/journal.pone.0235574

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

Galloway, C.D., Valys, A.V., Shreibati, J.B., Treiman, D.L., Petterson, F.L., Gundotra, V.P., Albert, D.E., Attia, Z.I., Carter, R.E., Asirvatham, S.J., Ackerman, M.J., Noseworthy, P.A., Dillon, J.J., Friedman, P.A., 2019. Development and Validation of a Deep-Learning Model to Screen for Hyperkalemia From the Electrocardiogram. JAMA Cardiol 4, 428–436. https://doi.org/10.1001/jamacardio.2019.0640

Golas, S.B., Shibahara, T., Agboola, S., Otaki, H., Sato, J., Nakae, T., Hisamitsu, T., Kojima, G., Felsted, J., Kakarmath, S., Kvedar, J., Jethwani, K., 2018. A machine learning model to predict the risk of 30-day readmissions in patients with heart failure: a retrospective analysis of electronic medical records data. BMC Med Inform Decis Mak 18, 44. https://doi.org/10.1186/s12911-018-0620-z

Guermazi, A., Tannoury, C., Kompel, A.J., Murakami, A.M., Ducarouge, A., Gillibert, A., Li, X., Tournier, A., Lahoud, Y., Jarraya, M., Lacave, E., Rahimi, H., Pourchot, A., Parisien, R.L., Merritt, A.C., Comeau, D., Regnard, N.-E., Hayashi, D., 2022. Improving Radiographic Fracture Recognition Performance and Efficiency Using Artificial Intelligence. Radiology 302, 627–636. https://doi.org/10.1148/radiol.210937

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

Haendel, M., Vasilevsky, N., Unni, D., Bologa, C., Harris, N., Rehm, H., Hamosh, A., Baynam, G., Groza, T., McMurry, J., Dawkins, H., Rath, A., Thaxon, C., Bocci, G., Joachimiak, M.P., Köhler, S., Robinson, P.N., Mungall, C., Oprea, T.I., 2020. How many rare diseases are there? Nat Rev Drug Discov 19, 77–78. https://doi.org/10.1038/d41573-019-00180-y

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
Hannun, A.Y., Rajpurkar, P., Haghpanahi, M., Tison, G.H., Bourn, C., Turakhia, M.P., Ng, A.Y., 2019. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. Nat Med 25, 65–69. https://doi.org/10.1038/s41591-018-0268-3
Hersh, W.R., Cohen, A.M., Nguyen, M.M., Bensching, K.L., Deloughery, T.G., 2022. Clinical Study Applying Machine Learning to Detect a Rare Disease: Results and Lessons Learned. medRxiv. https://doi.org/10.1101/2021.12.07.21267403
Homayounieh, F., Digumarthy, S., Ebrahimian, S., Rueckel, J., Hoppe, B.F., Sabel, B.O., Conjeti, S., Ridder, K., Sistermanns, M., Wang, L., Preuhs, A., Ghesu, F., Mansoor, A., Moghbel, M., Botwin, A., Singh, R., Cartmell, S., Patti, J., Huemmer, C., Fieselmann, A., Joerger, C., Mirshahzadeh, N., Muse, V., Kalra, M., 2021. An Artificial Intelligence-Based Chest X-ray Model on Human Nodule Detection Accuracy From a Multicenter Study. JAMA Netw Open 4, e2141096. https://doi.org/10.1001/jamanetworkopen.2021.41096
Hughes, J.W., Olgin, J.E., Avram, R., Abreau, S.A., Sittler, T., Radia, K., Hsia, H., Walters, T., Lee, B., Gonzalez, J.E., Tison, G.H., 2021. Performance of a Convolutional Neural Network and Explainability Technique for 12-Lead Electrocardiogram Interpretation. JAMA Cardiol. https://doi.org/10.1001/jamacardio.2021.2746
Kevat, A., Kalirajah, A., Roseby, R., 2020. Artificial intelligence accuracy in detecting pathological breath sounds in children using digital stethoscopes. Respir Res 21, 253. https://doi.org/10.1186/s12931-020-01523-9
Liang, H., Tsui, B.Y., Ni, H., Valentim, C.C.S., Baxter, S.L., Liu, G., Cai, W., Kermany, D.S., Sun, X., Chen, J., He, L., Zhu, J., Tian, P., Shao, H., Zheng, L., Hou, R., Hewett, S., Li, G., Liang, P., Zang, X., Zhang, Z., Pan, L., Cai, H., Ling, R., Li, S., Cui, Y., Tang, S., Ye, H., Huang, X., He, Waner, Liang, W., Zhang, Q., Jiang, J., Yu, W., Gao, J., Ou, W., Deng, Y., Hou, Q., Wang, Bei, Yao, C., Liang, Y., Zhang, S., Duan, Y., Zhang, R., Gibson, S., Zhang, C.L., Li, O., Zhang, E.D., Karin, G., Nguyen, N., Wu, X., Wen, C., Xu, J., Xu, W., Wang, Bochu, Wang, W., Li, J., Pizzato, B., Bao, C., Xiang, D., He, Wanting, He, S., Zhou, Y., Haw, W., Goldbaum, M., Tremoulet, A., Hsu, C.-N., Carter, H., Zhu, L., Zhang, K., Xia, H., 2019. Evaluation and accurate diagnoses of pediatric diseases using artificial intelligence. Nat Med 25, 433–438. https://doi.org/10.1038/s41591-018-0335-9
Liu, S., Kawamoto, K., Del Fiol, G., Weir, C., Malone, D.C., Reese, T.J., Morgan, K., ElHalta, D., Abdelrahman, S., 2022. The potential for leveraging machine learning to filter medication alerts. J Am Med Inform Assoc ocab292. https://doi.org/10.1093/jamia/ocab292
Liu, X., Faes, L., Kale, A.U., Wagner, S.K., Fu, D.J., Bruynseels, A., Mahendiran, T., Moraes, G., Shamdas, M., Kern, C., Ledsam, J.R., Schmid, M.K., Balaskas, K., Topol, E.J., Bachmann, L.M., Keane, P.A., Denniston, A.K., 2019. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. The Lancet Digital Health 1, e271–e297. https://doi.org/10.1016/S2589-7500(19)30123-2
Matheny, M., Israni, S.T., Ahmed, M., Whicher, D. (Eds.), 2019. Artificial Intelligence in Health Care: The Hope, the Hype, the Promise, the Peril.
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., Kannan, A., Chen, K., Vardoulakis, L., Chou, K., Cui, C., Dean, J., 2019. Automatically Charting Symptoms From Patient-Physician Conversations Using Machine Learning. JAMA Intern Med 179, 836–838. https://doi.org/10.1001/jamainternmed.2018.8558
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., Chen, E., Banerjee, O., Topol, E.J., 2022. AI in health and medicine. Nat Med 1–8. https://doi.org/10.1038/s41591-021-01614-0
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
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
Romero-Brufau, S., Whitford, D., Johnson, M.G., Hickman, J., Morlan, B.W., Therneau, T., Naessens, J., Huddleston, J.M., 2021. Using machine learning to improve the accuracy of patient deterioration predictions: Mayo Clinic Early Warning Score (MC-EWS). J Am Med Inform Assoc. https://doi.org/10.1093/jamia/ocaa347
Segal, G., Segev, A., Brom, A., Lifshitz, Y., Wasserstrum, Y., Zimlichman, E., 2019. Reducing drug prescription errors and adverse drug events by application of a probabilistic, machine-learning based clinical decision support system in an inpatient setting. J Am Med Inform Assoc 26, 1560–1565. https://doi.org/10.1093/jamia/ocz135
Shen, Y., Shamout, F.E., Oliver, J.R., Witowski, J., Kannan, K., Park, J., Wu, N., Huddleston, C., Wolfson, S., Millet, A., Ehrenpreis, R., Awal, D., Tyma, C., Samreen, N., Gao, Y., Chhor, C., Gandhi, S., Lee, C., Kumari-Subaiya, S., Leonard, C., Mohammed, R., Moczulski, C., Altabet, J., Babb, J., Lewin, A., Reig, B., Moy, L., Heacock, L., Geras, K.J., 2021. Artificial intelligence system reduces false-positive findings in the interpretation of breast ultrasound exams. Nat Commun 12, 5645. https://doi.org/10.1038/s41467-021-26023-2
Shortliffe, E.H., 2019. Artificial Intelligence in Medicine: Weighing the Accomplishments, Hype, and Promise. Yearb Med Inform 28, 257–262. https://doi.org/10.1055/s-0039-1677891
Smith, S.W., Walsh, B., Grauer, K., Wang, K., Rapin, J., Li, J., Fennell, W., Taboulet, P., 2019. A deep neural network learning algorithm outperforms a conventional algorithm for

emergency department electrocardiogram interpretation. J Electrocardiol 52, 88–95. https://doi.org/10.1016/j.jelectrocard.2018.11.013
Steele, A.J., Denaxas, S.C., Shah, A.D., Hemingway, H., Luscombe, N.M., 2018. Machine learning models in electronic health records can outperform conventional survival models for predicting patient mortality in coronary artery disease. PLoS One 13, e0202344. https://doi.org/10.1371/journal.pone.0202344
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
Topol, E., 2019. Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again, Illustrated Edition. ed. Basic Books, New York.
Tschandl, P., Rinner, C., Apalla, Z., Argenziano, G., Codella, N., Halpern, A., Janda, M., Lallas, A., Longo, C., Malvehy, J., Paoli, J., Puig, S., Rosendahl, C., Soyer, H.P., Zalaudek, I., Kittler, H., 2020. Human-computer collaboration for skin cancer recognition. Nat Med 26, 1229–1234. https://doi.org/10.1038/s41591-020-0942-0
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, S., Liu, S., Wang, Y., Timmons, T., Uppili, H., Bedrick, S., Hersh, W., Liu, H., 2017. Intrainstitutional EHR collections for patient-level information retrieval. Journal of the Association for Information Science and Technology 68, 2636–2648. https://doi.org/10.1002/asi.23884
Yao, X., Rushlow, D.R., Inselman, J.W., McCoy, R.G., Thacher, T.D., Behnken, E.M., Bernard, M.E., Rosas, S.L., Akfaly, A., Misra, A., Molling, P.E., Krien, J.S., Foss, R.M., Barry, B.A., Siontis, K.C., Kapa, S., Pellikka, P.A., Lopez-Jimenez, F., Attia, Z.I., Shah, N.D., Friedman, P.A., Noseworthy, P.A., 2021. Artificial intelligence-enabled electrocardiograms

for identification of patients with low ejection fraction: a pragmatic, randomized clinical trial. Nat Med 27, 815–819. https://doi.org/10.1038/s41591-021-01335-4

Zhang, J., Wang, H.-S., Zhou, H.-Y., Dong, B., Zhang, L., Zhang, F., Liu, S.-J., Wu, Y.-F., Yuan, S.-H., Tang, M.-Y., Dong, W.-F., Lin, J., Chen, M., Tong, X., Zhao, L.-B., Yin, Y., 2021. Real-World Verification of Artificial Intelligence Algorithm-Assisted Auscultation of Breath Sounds in Children. Front Pediatr 9, 627337. https://doi.org/10.3389/fped.2021.627337

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

Translational Artificial Intelligence (AI): The Need to Translate from Basic Science to Clinical Value

William Hersh
Professor and Chair
Department of Medical Informatics & Clinical Epidemiology
Oregon Health & Science University
May 16, 2022 – PDF of slides and references at www.billhersh.info or from @williamhersh

1

Translational AI – outline

- Promise of artificial intelligence (AI) and machine learning (ML) in medicine
- Current state of clinical impact of AI prediction tools
- Results of research aiming to diagnose rare disease using ML

2

One-slide history of AI and ML in medicine

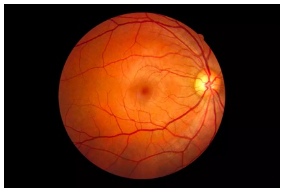
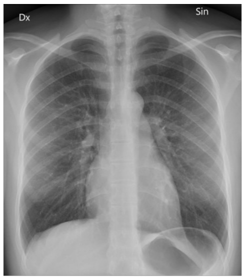
- A major activity of clinical informatics has been application of AI with aim of improving patient care (Shortliffe, 2019)
- First generation in 20th century
  - Focus on hand-crafted knowledge bases
  - Computers lacking data, power, GUIs, Internet, etc.
  - Led to “AI winter” in late 1980s and beyond
- Resurgence in 21st century
  - Driven by advances in ML, especially deep learning
  - Based on large amounts of data and plentiful computer power and networks
  - Overviews – Topol (2019), NAM (2019), Rajpurkar (2022)
  - Still modest impact (as of 2022) in clinical setting

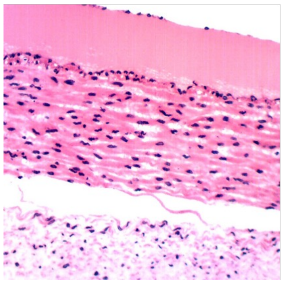
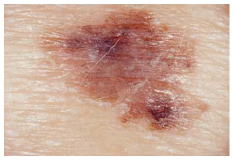
Promise of ML and AI in medicine

- Imaging
- Clinical prediction
- Biological processes
- Assisting humans

Imaging

- Early studies
  - Diabetic retinopathy (DR) (Gulshan, 2016; Ting, 2017)
  - Histology of cancer (Bejnordi, 2017) and metastases (Veta, 2019)
  - Tuberculosis (Lakhani, 2017) and pneumonia (Rajpurkar, 2018)
  - Skin cancer (Esteva, 2017; Haenssle, 2018; Tschandi, 2019)
- Systematic review (Liu, 2019)
- State of the art (Esteva, 2021)

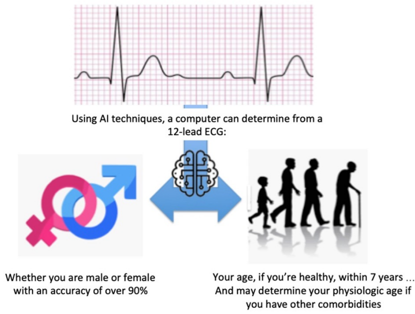



5

Other pattern-recognition areas

- Wave forms – use of ECGs
  - Age and sex determination (Attia, 2019)
  - Cardiac arrhythmia detection comparable to cardiologists (Hannun, 2019)
  - Interpretation better than conventional algorithm (Smith, 2019; Hughes, 2021)
  - Detecting hyperkalemia from 2 (of 12) leads (Galloway, 2019)
  - Early diagnosis of low ejection fraction in patients in setting of routine primary care (Yao, 2021)
- Sounds
  - Detecting pathological breath sounds in children with digital stethoscopes (Kevat, 2020; Zhang, 2021)



6

Clinical prediction

- Length of stay, mortality, readmission, and diagnosis at two large medical centers (Rajkomar, 2018)
- 30-day readmission in heart failure (Golas, 2018)
- ML-selected variables outperformed expert-selected variables in predicting patient mortality from coronary artery disease (Steele, 2018)
- Age and sex determination from retinal images (Poplin, 2018)
- Wide variety of pediatric diagnoses from EHR data at major referral center (Liang, 2019)
- Dementia from EHR data up to two years before clinical diagnosis (Wang, 2019)
- Improve accuracy of patient deterioration predictions (Romero-Brufau, 2021)
- Prediction models for mechanical ventilation, renal replacement therapy, and readmission in COVID-19 (Rodriguez, 2021)

OHSU

Assisting humans

- Automatically charting symptoms from patient-physician conversations (Rajkomar, 2019)
- "Weakly supervised" (using clinical diagnoses) interpretation of pathology slides would allow pathologists to exclude 65–75% of slides while retaining 100% sensitivity (Campanella, 2019)
- Learning outlier clinical alerts to reduce drug prescribing errors and adverse events (Segal, 2019)
  - 85% confirmed clinically valid, 80% considered clinically useful
  - Alert burden low – 0.4% of all medication orders
- Assisting dermatologists improved accuracy but poor ML worsened human performance (Tschandl, 2020)

OHSU

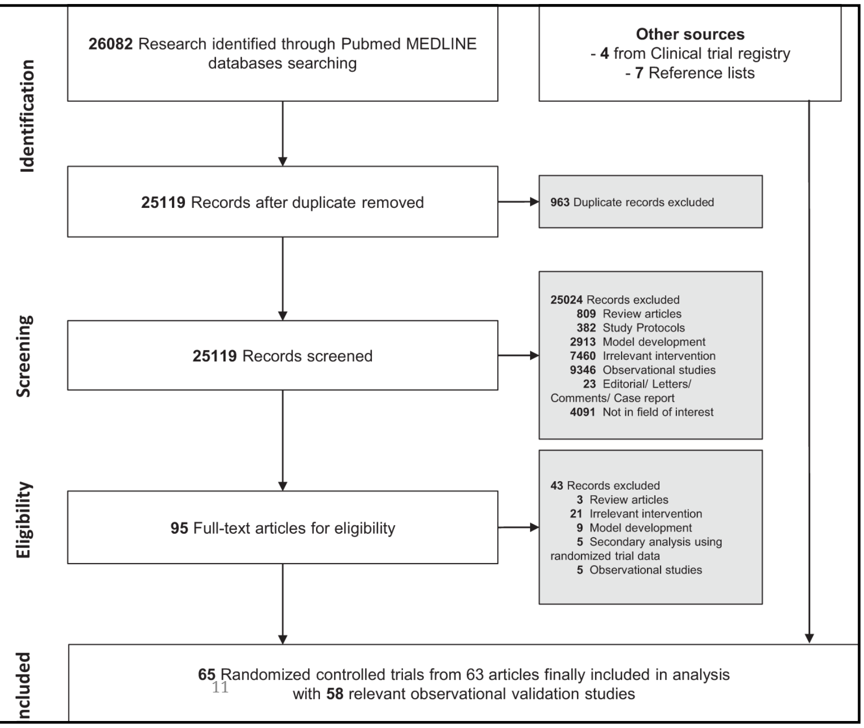
Assisting humans (cont.)

- Aiding radiologists
  - In breast ultrasound, reduced false-positive rates by 37.3% and requested biopsies by 27.8% while maintaining same level of sensitivity (Shen, 2021)
  - In interpreting CXRs, increased sensitivity for junior radiologists and specificity for senior radiologists (Homayounieh, 2021)
  - In fracture assessment, improved sensitivity without increasing reading time (Guermazi, 2022)
- AI system helped physicians extract relevant patient information in a shorter time while maintaining high accuracy (Chi, 2021)
- Identify features in CDS medication alerts to reduce volume by half while still maintaining 99% sensitivity (Liu, 2022)

How effective are interventions using AI clinical prediction tools?

- Zhou et al., 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
- Systematic 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
- TS and ML tools focused on assistive treatment decisions, assistive diagnosis, and risk stratification, whereas DL tools only focused on assistive diagnosis

65 RCTs from 26K+ publications – not uncommon



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, 0.90]
AUC in internal validation (n^a-18, median [IQR])		0.78 [0.73, 0.78]
AUC in external validation (n^a-20, median [IQR])		0.83 [0.79, 0.97]

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

Characteristics by tool type varied

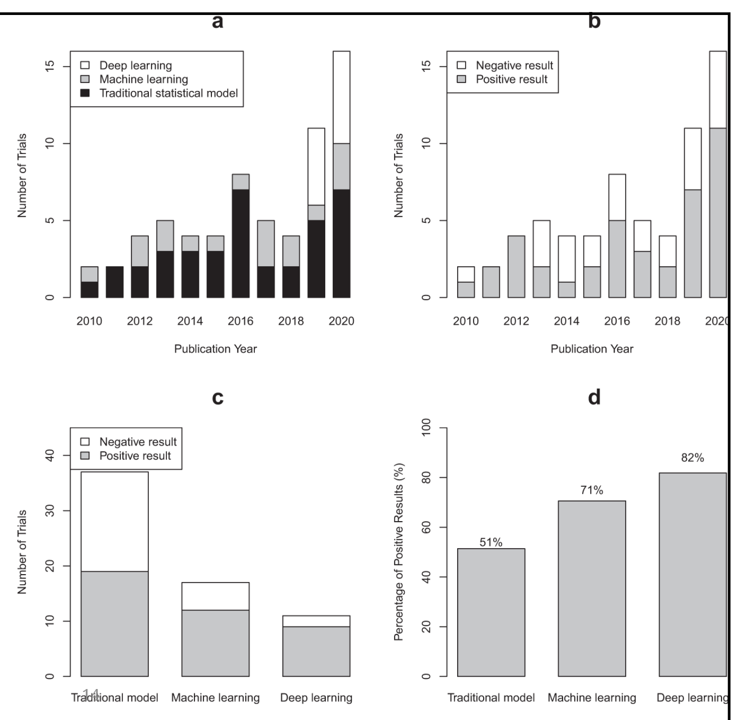
- Model input – clinical quantitative data for TS/ML, images for DL
- Disease category – varied for TS, chronic disease for ML, cancer for DL
- Tool function – risk stratification and treatment for TS, treatment for ML, diagnosis for DL
- Results – mixed for TS, more positive for ML/DL

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)

By publication year

- Increasing per year
- Increasing DL per year

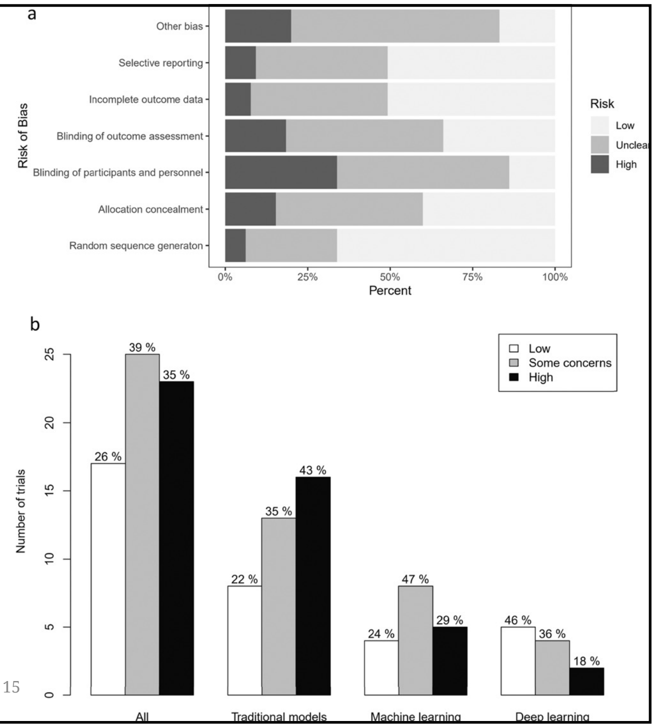
By tool type, more positive for DL > ML > TS



Only 17 of 65 trials with low risk of bias

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

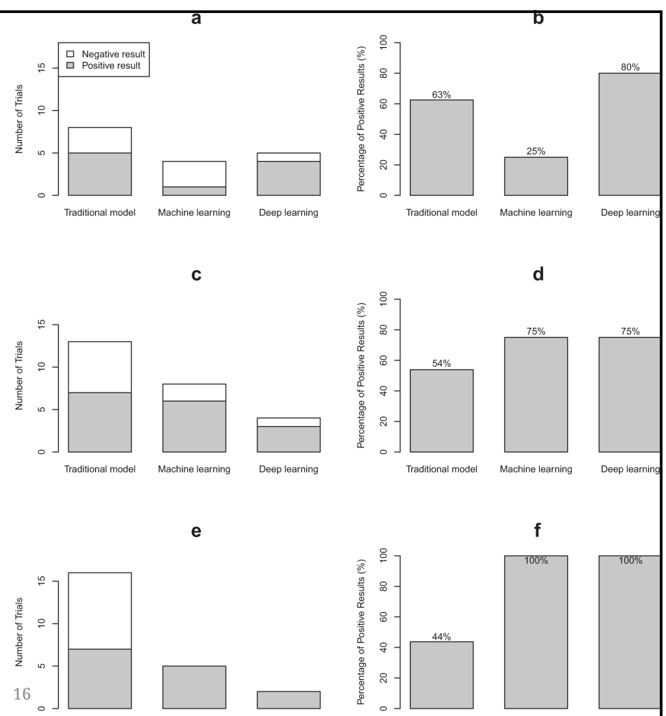
Suboptimal use of CONSORT, sample size pre-estimation, randomization, and intent-to-treat analysis



Proportion of trials and results for

- Low risk of bias – a-b
- Some concerns – c-d
- High risk of bias – e-f

For low risk of bias trials, positive outcomes in TS 63%, ML 25%, DL 80%



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 Arrest Registry	Negative

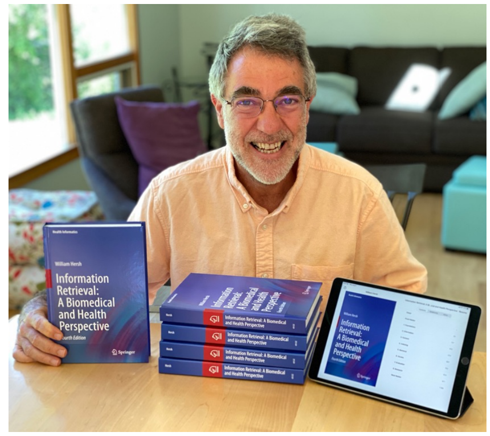
Abbreviations: AI = Artificial intelligence; DL = Tools using deep learning algorithms; ML = Tools using machine learning algorithms; CNN = Convolutionalneural networks; DCNN = Deep convolutional neural networks; CADe = Computer-aided detection; EGD = Esophagogastroduodenoscopy; OHCA = Out-of-hospital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

Conclusions about review

- 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
  - Trials of deep learning methods highly 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

My work: applying information retrieval (search) to EHR data

- Use cases
  - Cohort discovery
  - Detection of rare diseases
- Data set
  - OHSU Research Data Warehouse (fully identifiable)
- Funded by grants from
  - NLM 1R01LM011934
  - Alnylam Pharmaceuticals
- With help from OHSU collaborators
  - Steven Bedrick
  - Steven Chamberlin
  - Aaron Cohen
  - Tom Deloughery



(Hersh, 2020)

Cohort discovery

- Methods (Wu, 2017) and results (Chamberlin, 2020) for collection of 100K records
- Heterogeneous records make retrieval challenging
- R01 recently renewed and updating data, systems, and method
- A major challenge – privacy concerns

Adults with IBD who haven't had GI surgery	Adults with inflammatory bowel disease who haven't had surgery involving the small intestine, colon, rectum, or anus.
Adults with a Vitamin D lab result	Adults with a lab result for 25-hydroxy Vitamin D collected between May 15 and October 15.
Postherpetic neuralgia treated with topical and systemic medication	Adults with postherpetic neuralgia ever treated by concurrent use of topical and non-opioid systemic medications.
Children seen in ED with oral pain	Children who were seen in the emergency department with herpetic gingivostomatitis, herpangina or hand, foot, and mouth disease, tonsillitis, gingivitis, or ulceration (aphthae, stomatitis, or mucositis) not due to chemotherapy or radiation.
3rd trimester prenatal visit with midwife or Ob/Gyn	Women who had a pregnancy with a 3rd trimester outpatient prenatal visit with an obstetrician and gynecologist or midwife.

Rare disease detection

- Over 1200 known rare disorders that affect < 1 in 200K patients worldwide, many under-diagnosed (https://rarediseases.org/; Haendel, 2020)
- Acute Intermittent Porphyria (AIP, aka Acute Hepatic Porphyria)
  - Rare genetic disease of heme biosynthesis – variable penetrance
  - Incidence 1 per 100K in population
  - Often undiagnosed for long time
  - Significant morbidity and effect on quality of life
  - "Neurovisceral" symptoms common with other diseases
    - Abdominal pain
    - Nausea and vomiting
    - Psychiatric changes
  - Diagnosed by inexpensive urine porphobilinogen test
  - New highly effective (and highly expensive) treatment available – RNA-silencing molecule givosiran (Balwani, 2020)

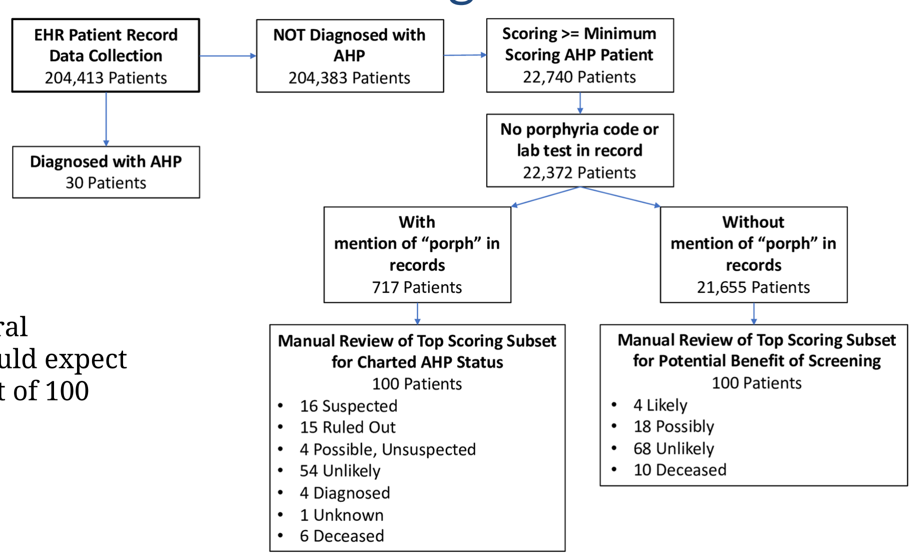
Can we detect rare diseases earlier using population-based techniques with EHR data?

- Funding from Alnylam Pharmaceuticals
- Expanded EHR data set to 200+ K patients
  - Updated base data set to 200K patients
    - Including from post-2015 era of ICD-10-CM coding
  - Enriched with 5,571 additional patients having "porph" in diagnoses, lab tests, and notes
- Preparation for machine learning
  - Positive training cases from ICD-10-CM E80.21 (47) with manual review to verify (30)
  - Negative training cases were the rest

Machine learning approach (Cohen, 2020)

- Parsed EHR record into features – scored by frequency of appearance, labeled features by the EHR source document
- Univariate feature analysis – manually choose features not directly tied to provider attributes or suspecting patient had porphyria
  - e.g., “DeLoughery” and “cimetidine”
- Trained on full dataset, with best performance using support vector machine (SVM) with radial basis function (RBF) kernel
- Applied trained model back to full data set – ranked patients by margin distance

Aimed to identify patients with symptoms but no consideration of diagnosis of AIP

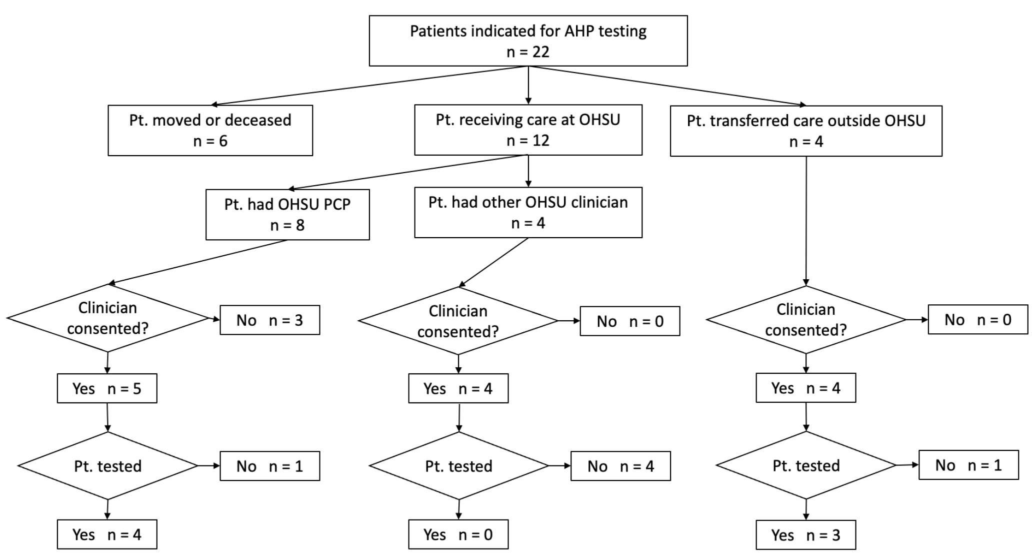


Note with natural prevalence, would expect 0.0005 cases out of 100

Clinical study (Hersh, 2022)

- Hindered by prolonged IRB process and COVID-19 pandemic, study was launched in late 2020
- IRB protocol required initial contact with primary care physician and, if they approved, offering the patient urine porphobilinogen testing
- Aimed to contact and enroll all 22 patients with AIP symptomatology but "unrecognized"

Clinician and patient participation



And the results showed…

- All 7 patients who came for testing had normal urine porphobilinogen

- Lessons learned
  - Clinical validation of machine learning models essential
  - Two-step approval required for patients not under our care but complicated
  - Rare diseases are rare
  - For other diseases, testing may be expensive and/or harmful

Conclusions

- As in all of medicine, results of basic science advances in AI must achieve clinical validation
- Many ML models have achieved basic science success but must be demonstrated to provide clinical value
- Much opportunity for research and researchers in this area

Thank you!

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: http://informaticsprofessor.blogspot.com
Twitter: @williamhersh

29