Axel Kaehne
Julie Feather *Editors*

Handbook of Health Services Evaluation

Theories, Methods and Innovative Practices

Springer

Handbook of Health Services Evaluation

Axel Kaehne • Julie Feather
Editors

Handbook of Health Services Evaluation

Theories, Methods and Innovative Practices

Springer

Editors
Axel Kaehne
Evaluation and Policy Analysis (EPA) Unit
and Medical School
Edge Hill University
Ormskirk, UK

Julie Feather
Evaluation and Policy Analysis (EPA) Unit
Edge Hill University
Ormskirk, UK

ISBN 978-3-031-87868-8 ISBN 978-3-031-87869-5 (eBook)
https://doi.org/10.1007/978-3-031-87869-5

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Artificial Intelligence, Healthcare, and Evaluation: Altering the Landscape

David Fetterman and David Fetterman II

Abstract

Artificial intelligence (AI) has altered the healthcare and evaluation landscapes, offering benefits from time efficiency to increased accuracy and productivity. AI-related healthcare improvements range from increased diagnostic accuracy to more face time with patients and insight into patients' medical needs. It has helped to build patient- and community-based healthcare evaluation capacity; helping individuals and communities become more self-determined concerning their health. Simultaneously, AI is transforming the field of evaluation, increasing evaluator's efficiency and precision. These synergistic changes are taking place in a rapidly evolving environment, making it imperative that AI healthcare products and services be continually evaluated. In addition, regulatory agencies need to be informed about AI-related developments in the healthcare and evaluation arena to help them appropriately guide and monitor these developments and innovations.

Keywords

Artificial intelligence · Healthcare · Evaluation · GPTs · Large language models · Bias

D. Fetterman (✉)
Fetterman and Associates and Claremont Graduate University, Claremont, CA, USA

D. Fetterman II
Northwestern University, Evanston, IL, USA
e-mail: davidfetterman2028@u.northwestern.edu

A. Kaehne, J. Feather (eds.), *Handbook of Health Services Evaluation*,
https://doi.org/10.1007/978-3-031-87869-5_22

1 Overview

In the first section of this chapter, we will clarify AI by defining terms and discussing the logic underlying AI focusing on mathematical probabilities. We will then present common AI chatbots, highlighting their capabilities and limitations. Finally, guidelines are also provided concerning prompts to harness AI's potential effectively.

Central to our discourse is the practical implementation of AI in healthcare and evaluation. Some of the most significant healthcare-related AI applications are in radiology, oncology, dermatology, and gastroenterology. AI applications in evaluation facilitate the development of logic models, theories of change, data analysis, and reporting. In addition, image creation and the construction of Generative Pretrained Transformers (GPTs) are discussed to clarify how AI works. Unpacking GPTs helps to guide the assessment of AI-driven healthcare innovations.

AI-related healthcare products and services can, in large part, be evaluated using traditional and community and program-involvement-evaluation approaches. However, evaluating AI products and services also requires special considerations. This discussion addresses the need to adapt evaluation tools and techniques (including AI-guided evaluation approaches), develop and establish AI-related evaluation guidelines and standards, and establish AI evaluation-informed regulatory frameworks.

Finally, the chapter addresses common AI challenges and concerns. It also responds to the ethical considerations that emerge from integrating AI into healthcare and evaluation practices, underscoring the need for conscientious and informed use of this powerful technology.

2 Defining Terms

Artificial intelligence (AI) enables machines to emulate cognitive functions that humans associate with the human mind, such as learning and problem-solving. Within this field, machine learning (ML) allows computers to learn from data and improve at tasks through experience, without being explicitly programmed for every contingency. ML can categorize data as structured, which includes clearly defined data types such as numbers and dates, or unstructured, which encompasses more complex information like audio and video content.

Deep learning, a more intricate aspect of ML, harnesses neural networks with multiple layers to interpret data with a high level of abstraction. These networks can identify patterns and features within vast amounts of data to carry out complex operations.

Moving beyond analysis and into creation, generative AI (GenAI) is a progression of ML capabilities. While ML extracts patterns to predict outcomes, GenAI synthesizes entirely new content. This innovative AI generates novel outputs such as

text, imagery, sounds, and video that mimic the quality of human-generated content. GenAI applications excel in streamlining processes, enhancing precision in data interpretation, and providing insights by identifying trends and motifs within datasets.

Among the various models underpinning GenAI, large language models (LLMs) stand out, especially in text-based applications. They understand and generate human-like text and can also support a myriad of tasks including translation, summarization, and content creation. Similar foundational models exist for audio and visual content, driving forward the capabilities of generative AI across multiple modalities. (See IBM Data and AI Team (2023) and Zewe (2023) for a more detailed explanation of AI terms.)

3 Understanding the Logic of AI

At the core of AI, particularly Large Language Models (LLMs), is the capacity to process and learn from extensive datasets. These models generate predictions by assessing mathematical probabilities derived from the data they have been trained on. For instance, in the phrase "United States of…," the model predicts that the word "America" will follow based on the likelihood of this sequence in its training data—not less probable continuations like "pizza."

This predictive capability extends to more complex patterns. Consider an LLM prompted with "My favorite drink." It logically infers the subsequent word to be "is," given the common structure of such a statement in English. As the sentence progresses, the LLM continues to anticipate the next most probable word: "Earl" might follow "My favorite drink is," and "Grey" after "My favorite drink is Earl." Ultimately, the model completes the sentence with "My favorite drink is Earl Grey tea," a pattern it recognizes as commonly occurring in its training data.

While LLMs are proficient in pattern recognition and prediction, their reliability is not guaranteed. These models necessitate ongoing training, corrective feedback, and reinforcement to refine their accuracy. An LLM's strength lies in its ability to learn and improve over time, which is why iterative training is a crucial aspect of enhancing its performance and precision in AI evaluations. [See IBM (2024) for a description of large language models.]

4 AI Accessibility: Chatbots and Image Generators

Generative Pre-trained Transformers (GPT) are neural network models that use AI to create human-like content. They are trained on large amounts of unlabeled text data. GPT models can respond to questions conversationally. They are also used to produce images and music.

AI chatbots and image generators come in various forms, often featuring both free and premium versions. A variety of chatbots are listed in Box 1.

Box 1 Types of Chatbots

- ChatGPT 3.5: This version is accessible without charge, but its knowledge is static and limited to information up until January 2022.
- ChatGPT 4 and 4o: For a monthly fee of $20, users gain access to the latest updates and a suite of advanced tools, including templates for generating custom GPT responses. [Team is $25].
- Claude: Known for handling extensive texts and supporting document attachments.
- Gemini: Offers a standard free version and an advanced "Ultra" subscription at $20 per month.
- Microsoft Copilot AI: Bundles features from ChatGPT 4 and the image creator DALL-E 3.

Challenges related to chatbots include the quality of the database used. The internet, for example, consists of widely variable quality. Similarly, how a chatbot is trained is rarely disclosed. The training may be systematically skewed or biased resulting in poor quality outputs or responses. The quality of chatbot prompts also determines the quality of the output (discussed later). In addition, chatbots are dependent on the quality of the algorithms employed.

Constructing a personalized GPT is not difficult and provides an additional measure of trust or security in the validity of the data and corresponding outputs. For example, GPT Creator or GPS Builder are free software tools to create a GPT. GPT Creator was used to create my "EmpowerEval Assistant," an AI tool used to guide communities in self-evaluations of healthcare-related programs for continuous improvement. The software provides self-guided instructions including how to create a GPT, recommended GPT names, images for a GPT profile, and most importantly uploading relevant and vetted files (GPT knowledge content). The benefit of creating a GPT is that it provides insight into how GPTs are built, making them easier to evaluate and modify to ensure quality results. The drawback of personalized GPTs is that they require large volumes of data and are time-consuming to maintain.

AI-driven image creation can significantly enhance the impact of evaluation findings. They can be used in evaluation reports, articles, blogs, and social media to punctuate a finding or illuminate a point. (See Box 2 below concerning types of image creators.)

Box 2 Types of AI-Driven Image Creators

- DALL-E 3: A standout in the field, it generates images from detailed prompts. Users may need to request multiple iterations for precise results. Its outputs are characterized by their artistic and stylized nature.
- KREA: Similar to DALL-E 3, but tends to produce more lifelike and realistic images, such as portraits.
- Midjourney: Another popular choice for generating images, noted for its distinct style and capabilities.

While these AI tools offer impressive features, they are not without limitations. For instance, image creators may sometimes render anatomical elements, like hands, with inaccuracies. Pictures of ethnic and racial group members' faces may be systematically distorted (Small 2023).

The ongoing debate questions whether AI-generated images sufficiently acknowledge and credit the original artists they reference. Additionally, there are concerns that AI may supplant human graphic designers. While AI can produce impressive results, it still needs human oversight and adjustments to reach the subtle expertise of professional designers. Despite these challenges, AI image generators are celebrated for their ability to deliver quality visuals quickly and at a lower cost compared to traditional methods.

In summary, the selection of an AI chatbot or image generator depends on specific requirements and preferences. These technologies are evolving rapidly, with new iterations entering the market as competition intensifies and user needs develop.

5 Prompts

Crafting effective prompts is essential for leveraging AI's capabilities, whether for text generation or image creation. A well-constructed prompt should be direct and specific and include any required style, such as academic or creative. It can also request adherence to a specific school of thought. If necessary, it should also ask for a specific format and sources to be cited.

The process often involves a series of iterative refinements. Take, for instance, the quest for an appropriate image to accompany a tuberculosis prevention evaluation report. The initial prompt—"a picture of a hypodermic needle eliminating tuberculosis"—may yield an image that is metaphorically correct but contextually inappropriate, perhaps showing a needle in an alarming scenario. A revised prompt aiming for sensitivity might be, "show a hypodermic needle used for lung disease prevention in rural India," but this too might return an image that, while closer to the goal, still misses the mark by portraying an unsettling scene.

Finally, a prompt that specifies the context and the action, such as "an illustration of a vaccine being administered in a rural Indian setting for lung disease prevention," is more likely to result in a suitable and effective image. It may depict a person receiving a vaccination with a visual cue indicating lung health, aligning with the report's communicative intent (see Fig. 1: iterative prompts and generated pictures).

This iterative process underscores the importance of precise and evolving prompts to guide AI toward producing the most relevant and effective outcomes.

6 Practical Applications of AI

Artificial intelligence (AI) boasts a wide range of practical uses across various industries. It streamlines tasks by generating outlines, summarizing documents, creating images, and composing various types of written content,

Iterative Prompts	Generated Pictures
First Prompt: "A picture of a hypodermic needle eliminating tuberculosis."	
Second Prompt: "A hypodermic needle eliminating lung disease in rural India."	
Third Prompt: "A hypodermic needle in a person's arm to eliminate lunch disease in rural India."	

Fig. 1 Iterative prompts (generated using DALL-E 3)

including letters, blog entries, social media updates, and comprehensive reports. AI's analytical prowess extends to parsing large databases and writing programming code.

Deep Learning, a form of pattern recognition, harnesses neural networks to teach a computer how to recognize certain symptoms or radiological images. One of the classic examples is Google's project to recognize cats based on ten million YouTube videos with a 75% accuracy rate (Clark 2019).

7 AI in Medicine

AI is already embedded in medicine. The virtual representation of AI in medicine includes online scheduling of appointments, online check-ins, digitized medical records, listening, transcribing, and writing progress notes services (i.e., AI Notetaker, NextGen Ambient Assist, Scribe, and TheraPro), telemedicine platforms, responses to patient inquiries (see Small 2023), as well as neural network-based

guidance to facilitate differential diagnoses. The physical stream of AI in medicine includes robotic-assisted surgery and intelligent prostheses (Hamlet and Tremblay 2017). Great strides are being made in the mainstream medical specialties.

Ophthalmologists leverage AI to examine hundreds of thousands of retinal scans, enhancing the precision of diagnoses related to conditions like diabetes, hypertension, kidney disease, and neurological disorders, such as Parkinson's disease and Alzheimer's (Cao et al. 2022; Sayres et al. 2018; Gulshan et al. 2016).

Similarly, gastroenterologists employ AI to sift through vast numbers of colonoscopy images, improving the detection of polyps (Wallace et al. 2022; Zhao et al. 2022). AI systems have also been able to outperform dermatologists in correctly classifying suspicious skin lesions (Esteva et al. 2017).

One of AI's most notable contributions is in radiology. AI provides faster diagnoses (Esteva et al. 2017; Teare et al. 2017; Bar et al. 2017; Makeeva, 2022), improved visualizations (Gulshan et al. 2016; Laukamp et al. 2018; Bejnordi et al. 2017; Li et al. 2012, 2014), and faster identification of alert situations (Prevedello et al. 2017; United States Food and Drug Administration 2018; Barreira et al. 2018), and is responsive to the shortage of experienced radiologists (Rosenkrantz et al. 2016; Nakajima et al. 2008).

In addition, it has been known to outperform experienced radiologists in diagnosing from medical images (Allen et al. 2021; Stiepan 2022; Somashekhar et al. 2018). The technology's edge lies in its ability to review far more images than human professionals can, operating continuously without fatigue and maintaining consistent performance. [See Eric Topol's TED Talk (2023) Can AI Catch What Doctors Miss?; see also Mayo and Leung, 2018 concerning the use of computer-assisted diagnosis.] A list of AI tools used in medicine is provided in Box 3.

Box 3 AI Tools in Medicine

AI Notetaker: listens, transcribes, and writes automated progress notes.

AI Therapy: online course for people with social anxiety using Cognitive Behavioral Therapy.

Einstein Copilot: Health Actions: use conversational AI to enable qualified healthcare professionals to trigger workflows, including capturing and summarizing patient details from different sources, sending referrals, booking appointments, and revising care plans.

eMed: telemedicine platform where patients can meet doctors through mobile apps and chat.

NextGen Ambient Assist: transforms patient-provider conversations on mobile devices into temporary transcripts, generating structured SOAP notes (allows physicians to spend more time observing body language instead of typing while listening to the patient).

NYU Langone's Electronic Health Record GPT-4: AI chatbot that responds to patient email inquiries to physicians, providing an empathetic response similar in quality to human providers (see Small et al. 2024).

Scribe: collects symptoms from patients online, assists with potential diagnosis of symptoms (in real-time), helps create a plan for the case, and turns conversations into medical notes.

TheraPro: creates therapy session summaries, enabling therapists to focus on their patients instead of taking notes.

Youper: mental health chatbot app that uses cognitive behavioral therapy and positive psychology techniques to help users deal with anxiety and depression (clinically effective).

8 Patient Use of AI

The online application eMed is used by patients to consult with their physician, inquire about symptoms, monitor health, and order tests. AI-driven therapy services range from online courses to consultations designed to help patients use cognitive behavior therapy and positive psychology to treat social anxiety (McCall et al. 2018).

Fitbit, Apple, and Qardio are health trackers. They monitor heart rate, sleep levels, steps, environmental sound levels, blood oxygen, blood pressure, cardio fitness, and sexual activity.

9 AI Applications in Evaluation

AI helps make evaluation more efficient and effective. It is used to organize thinking, streamline tasks, improve writing quality, analyze data, and build capacity. A few examples drawn from community health evaluations are provided below, such as a logic model, theory of change, and interview questions.

10 Logic Model

A logic model is a visual tool that shows how a program works. It breaks down the program into different parts:

1. Inputs: The things needed to make the program happen, like money, people, and materials.
2. Activities: The things the program does, like workshops or events.
3. Outputs: The results of these activities, like reports or numbers of people reached.
4. Outcomes: The changes or benefits the program aims to achieve, both short-term and long-term.
5. Impact: The bigger, long-lasting changes the program hopes to make in society.

Overall, a logic model helps us understand how a program is supposed to work and what it aims to achieve. It's like a roadmap for planning, running, and

evaluating programs. The example below shows how a logic model for a program was created.

The AI prompt requested a logic model for a tobacco prevention program, and it provided all the necessary components, including inputs, activities, outputs, outcomes, and impacts. Additionally, it included relevant details specific to the program's logic model. The AI response generated is presented below.

Inputs
- Funding and resources
- Trained staff and volunteers
- Partnerships with schools, community organizations, and healthcare providers

Activities
1. Research and planning
2. Community outreach
3. Education and awareness
4. Policy advocacy
5. Access to resources

Outputs
- Number of workshops conducted
- Number of presentations delivered
- Number of educational materials distributed
- Number of media campaigns launched
- Number of community partnerships established

Outcomes
- Increased knowledge and awareness of tobacco-related risks and consequences.
- Improved attitudes and perceptions toward tobacco use
- Increased adoption of tobacco-free policies in schools and public spaces
- Increased utilization of cessation resources and support services

Impact
- Decreased tobacco use among the target population
- Improved overall health and well-being of the community
- Reduced tobacco-related illnesses and associated healthcare costs

This logic model outlines the key components of a tobacco prevention program, starting with inputs like funding, staff, and partnerships. These inputs fuel activities such as research, outreach, education, advocacy, and resource provision. The program then generates outputs like workshops, presentations, and materials, which

contribute to outcomes such as increased knowledge, changed attitudes, policy adoption, and resource utilization. Ultimately, the impact is seen in reduced tobacco use, improved community health, and lower healthcare costs. The model emphasizes the importance of setting clear goals, understanding the target population, and utilizing available resources effectively. AI can be used in a similar way to generate theories of change, which help identify causal connections between program interventions and program outcomes.

11 Theory of Change

Weiss (1995), a leading figure in evaluation research, stated, there is "nothing as practical as a good theory". This statement underscores the significance of theory in informing practical decision-making and actions across different domains, such as program evaluation and policy analysis. It implies that robust theoretical frameworks offer valuable insights and tools for comprehending intricate phenomena and crafting successful interventions.

A theory of change is an explanation of how and why a desired transformation is anticipated to unfold within a specific context. It lays out the fundamental assumptions, causal pathways, and mechanisms through which interventions or activities are expected to generate particular outcomes or impacts.

By explaining the reasoning behind how interventions will cause change, a theory of change acts as a guiding framework for designing, implementing, and evaluating healthcare programs or initiatives. It facilitates stakeholders' understanding of the cause-and-effect relationships, thereby empowering them to make informed decisions and adaptations throughout the intervention process.

A *theory of change* for a tobacco prevention program outlines the causal pathway from program activities to desired outcomes. In this specific case, AI produced the following result:

1. Resource mobilization:
   (a) Funding: Secure financial resources to sustain program activities.
   (b) Human capital: Recruit and train skilled staff, educators, and counselors.
   (c) Partnerships: Collaborate with schools, healthcare providers, NGOs, and community leaders.
2. Awareness and advocacy:
   (a) Education campaigns: Conduct targeted campaigns in schools, workplaces, and public spaces.
   (b) Policy advocacy: Engage policymakers to enforce smoke-free laws and regulations.
   (c) Media outreach: Utilize TV, radio, and social media to disseminate anti-tobacco messages.

3. Empowerment and skill-building:
   (a) Youth programs: Develop age-appropriate curricula on tobacco risks.
   (b) Life skills training: Equip youth with refusal skills, stress management, and coping mechanisms.
   (c) Counseling services: Offer personalized support for smokers aiming to quit.
4. Behavioral change:
   (a) Preventing initiation: Educate youth about the dangers of tobacco use.
   (b) Promoting compliance: Encourage adherence to smoke-free policies.
   (c) Facilitating cessation: Provide evidence-based interventions for smokers.
5. Intermediate outcomes:
   (a) Shift in attitudes: Reduced social acceptance of smoking.
   (b) Increased knowledge: Improved understanding of health risks.
   (c) Behavioral shift: Decreased experimentation and initiation among youth.
6. Long-term impact:
   (a) Health improvements: Reduced tobacco-related diseases (cancer, respiratory issues).
   (b) Economic benefits: Lower healthcare costs and increased productivity.
   (c) Social norms transformation: Tobacco use becomes less normalized.

The theory represents a useful starting point to ensure the major steps are included. It can be refined with more detail about the specific tobacco prevention and cessation program.

12 Interview Questions and Surveys

Another useful application of AI in evaluations is in designing data collection instruments. Interview questions and surveys are essential in evaluations. AI can generate various types of questions, including informal, formal, structured, and semi-structured ones. These questions are logically organized and avoid compound or leading formats. Additionally, AI can create targeted professional surveys and questionnaires.

AI can be useful for specific community-based healthcare evaluation approaches, such as empowerment evaluations,[1] in which community and program members are responsible for the assessment process (Fetterman et al. 2010, 2015; Fetterman 2023a, b). AI enables community members to generate guides and steps needed when they lack the funds to employ experienced evaluation teams. Recommended empowerment evaluation steps based on an AI inquiry include a three-step and ten-step approach. AI can also generate a list of interview questions for this purpose, and The Empowerment Evaluation Interview Questions below are an example of an AI generated qualitative data collection instrument.

[1] See Fetterman et al. (2010), in which empowerment evaluation was used to transform the curriculum and prepare for and pass an accreditation review at Stanford University's School of Medicine.

Empowerment Evaluation Interview Questions and Guidelines

Understanding Needs and Perceptions
- What were your initial hopes and goals for this program/initiative?
- How has this program/initiative impacted your ability to address those goals?
- What aspects of your daily work or life do you feel most empowered by?
- Conversely, what aspects leave you feeling powerless or limited?

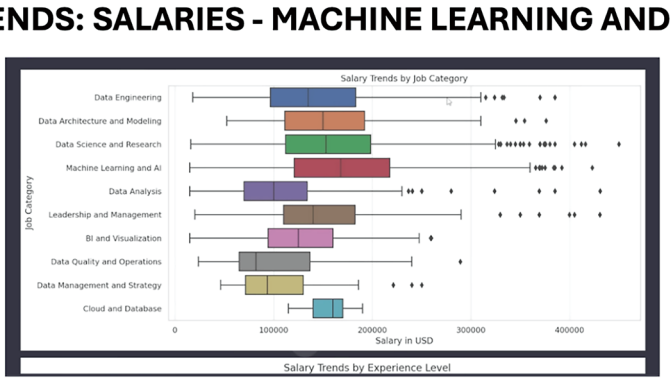
Assessing Capacity Building
- Have you gained any new skills or knowledge through your involvement?
- How comfortable do you feel using these skills to advocate for yourselves or others?
- Does this program/initiative provide opportunities for you to share your knowledge with others?

Gauging Participation and Ownership
- How involved have you been in shaping the direction of this program/initiative?
- Do you feel your voice is heard and valued in decision-making processes?
- What suggestions do you have for increasing community and program staff, and participant member participation in the future?

Identifying Strengths and Weaknesses
- What are the biggest strengths of this program/initiative from your perspective?
- Conversely, what are some areas where you see room for improvement?
- How can we better address any challenges or barriers you've encountered?

Promoting Reflection and Learning
- What have you learned the most from your participation in this program/initiative?
- How can we use this evaluation process to improve our work moving forward?
- What are your hopes for the future of this program/initiative?

Remember
- Empowerment evaluation emphasizes open-ended questions that encourage community and program staff members to share their experiences and perspectives.
- Wording should be clear and easy to understand, avoiding jargon or overly technical terms.
- Actively listen and acknowledge the responses of interviewees.
- The goal is to create a collaborative space for learning and improvement.



1. Salary Trends by Job Category – boxplots show a wide range of salaries across different job categories
2. Salary Trends by Experience Level – clear trend showing higher salaries for more experienced levels
3. Salary Distribution by Employment Type – plot indicates the salary distribution across different types
4. Number of Jobs by Company Size – bar graph shows the distribution of job types across companies with different sizes

Fig. 2 Using AI to analyze data sets (generated using ChatGPT 4o)

13 Data Analysis

Data analysis is a critical step after data collection in most evaluations. It typically involves triangulation, identification of patterns, content analysis, and statistics (Fetterman 2020). AI can streamline much of this process, saving time and providing rapid results. AI chatbots can interpret, clean, and analyze data efficiently.

For instance, an analysis of salary trends in computing by job category is useful when hiring a team in a healthcare system, e.g., management, data engineering, cloud and database management, and machine learning and AI. To determine the appropriate compensation when hiring a computing staff, consider a publicly available dataset on salary trends in computing from Kaggle (free, public data collections for research and training). The initial prompt asked AI to interpret the data, and it promptly provided insights on salary trends by job category, experience level, employment type, and company size. Next, AI cleaned the data by removing duplicate rows, ensuring better accuracy. Typically, manual data cleaning takes days, but AI accomplishes it in seconds. Finally, when asked about trends, AI swiftly generated salary trend visualizations. [See Fig. 2. Using AI to Analyze Data Sets.]. The same approach applies to any healthcare database.

14 Reporting and Images

Disseminating findings is as important as producing evaluation results. AI can assist in organizing and presenting these findings logically, persuasively, and with visual impact. It can summarize extensive documents and revise draft reports, eliminating typos and unnecessary jargon for better readability. Styles can vary from

professional and balanced to marketing-oriented or even poetic. AI also aids in creating social media posts, presentation materials, and scholarly journal manuscripts. While the evaluator remains responsible for data and manuscript drafts, AI significantly enhances the quality of text and the impact of presentations.

For instance, AI-generated images effectively conveyed the global impact of smoking on public health. By requesting an image of a hand holding a globe with a cigarette (using DALL-E 3), the message was powerfully communicated: cigarette smoking is a worldwide health issue. These visuals captured the attention of policymakers and community members, drawing them to the evaluation findings and recommendations.

15 Evaluating AI

Much of AI can be evaluated using traditional evaluation approaches. Traditional evaluation tools include both qualitative and quantitative approaches, such as interviews, observations, online surveys, and statistical analysis (Davidson 2004; Fetterman 1988, 2020; Patton and Campbell-Patton 2021). However, evaluating AI products and services also requires special considerations, including adapting evaluation tools and techniques (including using AI-guided evaluation approaches), developing and establishing AI-related evaluation guidelines and standards, and establishing AI evaluation-informed regulatory frameworks. An outline of recommended steps is presented below:

16 Apply Traditional Evaluation Tools and Techniques

The majority of the existing evaluation theories, concepts, techniques, and tools apply to evaluating AI. This includes systematically applying existing evaluation approaches, methods, tools, and techniques. It also extends to stakeholder involvement approaches to ensure involvement and participation in the evaluation of community healthcare initiatives, e.g., collaborative, participatory, and empowerment evaluation approaches (Fetterman et al. 2018). In addition, evaluation standards, e.g., utility, feasibility, propriety, accuracy, usability, and accountability, are applicable to AI in medicine or health care services.

And while ethical and legal requirements, e.g., informed consent, transparency, and confidentiality, are certainly relevant, the nature and quality of AI supported activities in health services require careful calibration and discussion of AI's impact on them. This remains a topic of heated debate, which is likely to continue as AI is being increasingly incorporated into medical and health care processes. We set out below how current evaluation approaches may be adapted in this newly emerging context.

17 Adapt Existing Evaluation Tools and Techniques

Although most evaluation approaches, tools, and techniques apply to evaluating AI, new rapid formative evaluation approaches and methods are needed. Evaluation and evaluators need to evolve and adapt to the "moving targets" nature of the development of AI products and services, e.g., changes in the community, market, and AI product or service. In particular, the rapid timeline associated with the development and deployment of AI-driven products and services, e.g. "fail quickly" mindset; "beta" quality testing; market, company, patient or client, and stockholder timeline considerations are of significant importance. In addition, there is a variability concerning the quality and efficacy of AI products, services, and solutions, which requires close scrutiny by evaluators using AI tools. It remains critical to be aware of the difficulties in interpreting and validating the outputs of AI systems in complex healthcare settings and the need to cross-validate evaluation findings. Last but not least, the field needs to develop, test, and refine the use of AI-guided evaluation approaches.

18 Develop and Establish AI-related Evaluation Guidelines and Standards

Evaluators need to develop and establish quality, defensible, and realistic guidelines and standards concerning AI-related development, performance, and evaluation. Important considerations include time and resource constraints, an understanding of the dimension and urgency of the problem, the impact of the dimensions of fairness, interpretability, and bias of AI algorithms as well as at what stage a specific AI product or service development is, e.g., its conceptualization, product development, and/or implementation. Moreover, evaluators should be aware of the security, validity, and cost-effectiveness of AI tools and develop multiple "gold standards" depending on the task and application with a focus on end-user of technology to help determine effectiveness of AI tools. Arguably, this requires the inclusion of a wide spectrum of AI developers and relevant healthcare stakeholders (including healthcare providers and patients) in the development of AI evaluation guidelines and standards as well as open data repositories to save time and expense and foster collaboration.

Checklists are useful tools to remind evaluators to adhere to existing and newly developed guidelines and standards. An abbreviated example is provided below concerning a tuberculosis "cough" sound-based AI early detection tool evaluation, which was part of a USAID and Tuberculosis Implementation Framework Agreement (TIFA) funded project in India.

Tuberculosis "Cough" Sound-Based AI Early Detection Tool Evaluation Evaluation Checklist

This is an abbreviated evaluation checklist designed to guide the evaluation of an AI-driven tuberculosis "cough" app in India. People cough into the speaker of a phone and the app uses the sound of the cough to help determine if the person has

tuberculosis and needs to come to the clinic for testing. The checklist highlights how evaluation is applied to the task and how, in some cases, it needs to adapt to properly evaluate AI-related healthcare products and services.

1. Does this app respond to a societal need?
   Answer: Yes. TB is among the top ten causes of death globally and India has 27% of the total TB burden.
2. Does the app do what it claims to do? If so, how well?
   Answer: Yes, it is an early detection tool. The effects vary as per context and feasibility of confirmatory tests. It has shown exclusive results in identifying TB at very early stages during deployments.
   Degree of difficulty factor: The app is designed to reach people living on the street who are difficult to reach with conventional healthcare approaches after appropriate regulatory approvals.
   Recommendations: More resources will benefit to adjust algorithms and sound sensitivity and test on a larger population
3. Are there any ethical considerations associated with the use of the app?
   Answer: Yes. The healthcare provider may get infected if they hold the phone in front of the prospective patient as they cough.

Solutions:

1. The healthcare provider can stand behind the prospective patient as they cough into the phone.
2. The healthcare provider can use a selfie stick to hold the phone away from themselves.
3. The prospective patient can hold the phone.
4. The health care worker and patient, both, wears the mask (this is the current practice).

Source: Fetterman et al. (2015). Evaluation Checklist Based on Learning Meeting on Cough Sound-Based AI Tools for Improved TB Care.

19 Establish AI-Related Evaluation Regulatory Frameworks

Authorizing, governing, and regulatory bodies need assistance evaluating AI applications and solutions in healthcare settings. They need guidance to inform the development of regulations, regulatory decision-making, and enforcement. Their work should be guided by a variety of considerations, including accessibility, accuracy, cost, ethics, inclusiveness, quality, responsibility, responsiveness, safety, and sustainability.

Specifically, they need to be aware of relevant criteria to evaluate AI solutions, specific criteria to determine if AI solutions align with social values and the need to update regulations to adapt to advancements in AI technology. This also requires updating and incorporating ethical and legal considerations ranging from handling

sensitive patient information to copyright protection and IP concerns. Ideally this should result in streamlined evaluation processes and recognized approval frameworks based on credible assessments.

20 Common Challenges and Concerns Using Artificial Intelligence

Artificial intelligence (AI) is a powerful tool with substantial potential, but it is not infallible. Much like a developing child, AI requires ongoing nurturing in the form of continuous training and feedback to enhance its accuracy. Several common issues can arise if AI systems are not carefully managed and updated.

One such issue is the handling of sources and references. AI does not inherently know to cite sources unless this practice is built into its design, or it is specifically prompted to do so. Microsoft Copilot AI, by design, is an example in which AI automatically includes links to references to facilitate additional inquiry concerning the topic and make it easier to assess the authenticity and validity of the citation. However, at this stage of AI development, users must continue to be diligent in verifying the information AI provides.

Bias within AI systems is a critical issue that can skew results, reinforcing and amplifying social inequalities. Known as an "alignment problem," it arises when AI's outputs are not aligned with intended goals and societal norms (Watkins 2024). For instance, AI can inadvertently uphold stereotypes by unfairly correlating certain races or genders with particular professional roles (UNESCO 2024; Ahn and Costigan 2019; Mwangi and Kimani 2023). Such biases can adversely affect decisions in employment, healthcare, and law enforcement (Banaji et al. 2021; Swarns 2023). For example, predictive policing algorithms have produced racist results, including unjust incarceration (see ShotSpotter example: Burke et al. 2022; Heaven 2020). Addressing AI bias requires expertise specific to the affected domains, such as deep understanding and experience with minority communities, to lessen or prevent harmful social impacts.

Confidentiality presents a significant challenge as well. When AI systems are trained on proprietary (or private) data, there's a risk that sensitive information could inadvertently be disseminated. To address this, some organizations opt for private AI models or enterprise solutions that can be customized and controlled to safeguard proprietary information.

One of the most commonly reported issues with AI is the occurrence of "hallucinations," or the generation of factually incorrect information. This is not a result of deliberate deceit but rather a limitation of the AI's understanding. For example, if AI is asked to produce images of the far side of the moon, a side that is largely unphotographed by humans, it may generate plausible but not factual images based on available data. It is not an act of deception but an attempt to fulfill the request with the available information, often drawing on what it has learned about similar subjects. This problem persists; however, it is much less frequent because LLMs are learning on larger data systems.

All of these issues are escalated in healthcare contexts because the risks are higher when human lives and the quality of care are at stake in a litigious environment. Understanding these limitations is crucial in managing expectations of AI's capabilities. It's important to recognize that AI operates within the scope of its programming and training, always aiming to provide the most informative response possible. [See Villasenor 2019 for an additional discussion about artificial intelligence and bias.]

21 Conclusion

AI has significantly enhanced healthcare and healthcare evaluation practices. It efficiently analyzes large datasets and produces compelling, accurate texts and visual presentations. In medicine, AI is already embedded in the healthcare system, including notation and scheduling programs. It is also helping to improve patient care ranging from early detection and diagnosis to reducing morbidity and mortality. AI is even responding to workforce issues by providing prompt and accurate diagnostic responses to radiological inquiries, despite a reduced labor force of radiologists.

However, it is not flawless. AI makes errors. Moreover, it can perpetuate existing biases found in the data it is trained on, potentially amplifying historical biases on a large scale automatically and rapidly. The problem of AI misalignment with our social values is of paramount significance. It is fundamental to our vision of a fair, equitable, and democratic society. The best approach to improvement is to include stakeholders with domain-specific expertise in AI development decision-making.

Oversight and regulations are also needed to monitor the development and use of AI healthcare products and innovations. Human subjects issues that need to be taken into consideration range from data sharing and data security to patient confidentiality. The safety and security of AI-informed medical devices also require review. Do they provide accurate diagnostic information about health (as claimed)? Are the devices calibrated? Are they monitored continually as they continue to develop and improve (Dennis 2022)?

AI will continue to expand its boundaries in medicine and evaluation. The best way to monitor and improve it is to use it. AI requires consistent usage, continuous feedback, and ongoing learning to make it a surgically precise tool aligned with our social values and a commitment to democratic ideals. With sufficient care and oversight, AI can continue to enhance healthcare and healthcare evaluation practices.

References

Ahn S, Costigan A (2019) How AI reinforces gender stereotypes (trend brief). https://www.catalyst.org/research/ai-gender-stereotypes/

Allen B, Agarwal S, Coombs L, Wald C, Dreyer K (2021) 2020 ACR data science institute artificial intelligence survey. J Am Coll Radiol 18(8):1153–1159

Banaji MR, Fiske ST, Massey DS (2021) Systemic racism: individuals and interactions, institutions and society. Cogn Res 6:82. https://doi.org/10.1186/s41235-021-00349-3

Bar A, Wolf L, Amitai O et al (2017) Compression fractures detection on CT. Med Imaging 2017:10134

Barreira CMA, Bouslama M, Lim J et al (2018) ALADIN study: automated large artery occlusion detection in stroke imaging study- a multi-center experience. Eur Stroke J 3:75

Bejnordi BE, Veta M, van Diest PJ et al (2017) Diagnostic assessment of deep learning algorithms for detection of lymph node metastasis in women with breast cancer. JAMA 318:2199–2210

Burke G, Mendoza M, Linderman J, Tarm M (2022) How AI-powered tech landed man in jail with scant evidence. https://apnews.com/article/artificial-intelligence-algorithm-technology-police-crime-7e3345485aa668c97606d4b54f9b6220

Cao J, Chang-Kit B, Katsnelson G et al (2022) Protocol for a systematic review and meta-analysis of the diagnostic accuracy of artificial intelligence for grading of ophthalmology imaging modalities. Diagn Progn Res 6:15. https://doi.org/10.1186/s41512-022-00127-9

Clark L (2019) Google's artificial brain learns to find cat videos. Wired UK Science. http://www.wired.com/2012/06/google-xneural-network

Davidson J (2004) Evaluation methodology basics: the nuts and bolts of sound evaluation. Sage, Thousand Oaks

Dennis A (2022) Defining artificial intelligence (AI) and why it matters for medical device regulation. Digital Health. Taylor and Wessing. https://www.taylorwessing.com/en/insights-and-events/insights/2022/06/defining-artificial-intelligence-and-why-it-matters-for-medical-device-regulation

Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM et al (2017) Dermatologist-level classification of skin cancer with deep neural networks. Nature 542:115–118

Fetterman DM (1988) Qualitative approaches to evaluation in education: the silent scientific revolution. Praeger, New York

Fetterman DM (2020) Ethnography: step by step. Sage, Thousand Oaks

Fetterman DM (2023a) Empowerment evaluation and social justice: confronting the culture of silence. Guilford Press, New York

Fetterman DM (2023b) Searching for self-determination and results. YouTube. https://www.youtube.com/watch?v=lsY1APZVUwI

Fetterman DM, Deitz J, Gesundheit N (2010) Empowerment evaluation: a collaborative approach to evaluating and transforming a medical school curriculum. Acad Med 85(5):813–820

Fetterman DM, Kaftarian S, Wandersman A (2015) Empowerment evaluation: knowledge and tools for self-assessment, evaluation capacity building, and accountability. Sage, Thousand Oaks

Fetterman, D.M., Rodriquez-Campos, L. Zukoski, A.P., and Contributors, (2018). Collaborative, Participatory, and Empowerment Evaluation: Stakeholders Involvement Approaches. New York: Guilford Press.

Gulshan V, Peng L, Voram M et al (2016) Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA 16:2402–2410

Hamlet P, Tremblay J (2017) Artificial intelligence in medicine. Metabolism 69:36–40

Heaven W (2020) Artificial intelligence: predictive policing algorithms are racist. They need to be dismantled. MIT Technology Review, Cambridge. https://www.technologyreview.com/2020/07/17/1005396/predictive-policing-algorithms-racist-dismantled-machine-learning-bias-criminal-justice/

IBM (2024) What are large language models (LLMs)? https://www.ibm.com/topics/large-language-models

IBM Data and AI Team (2023) AI vs. machine learning vs. deep learning vs. neural networks: what's the difference. https://www.ibm.com/blog/ai-vs-machine-learning-vs-deep-learning-vs-neural-networks/

Laukamp KR, Thiele F, Shakirin G et al (2018) Fully automated detection and segmentation of meningiomas using deep learning on routine multiparametric MRI. Eur Radiol 29:124–132

Li YH, Zhang L, Hu QM et al (2012) Automatic sub-arachnoid space segmentation and hemorrhage detection in clinical head CT scans. Int J CARS 7:507–516

Li R, Zhang W, Suk HI et al (2014) Deep learning based imaging data completion for improved brain disease diagnosis. Med Image Comput Assist Interv 17:305–312

Makeeva A (2022) An essential roadmap for AI in radiology. ACR Bulletin https://www.acr.org/Practice-Management-Quality-Informatics/ACR-Bulletin/Articles/September-2022/An-Essential-Roadmap-for-AI-in-Radiology

Mayo RC, Leung J (2018) Artificial intelligence and deep learning-radiology's next frontier? Clin Imaging 49:87–88

McCall HC, Richardson CG, Helgadottir FD, Chen FS (2018) Evaluating a web-based social anxiety intervention: a randomized controlled trial among university students? J Med Internet Res 20:e91. https://doi.org/10.2196/jmir.8630

Mwangi L, Kimani E (2023) The hidden burden: gender bias and stereotyping as a barrier to women's leadership. IDInsight https://www.idinsight.org/article/the-hidden-burden-gender-bias-and-stereotypes-as-a-barrier-to-womens-leadership/

Nakajima, Y., Yamada, K., Imamura, K., Kobayashi, K. (2008). Radiologist supply and workload: international comparison—Working Group of Japanese College of Radiology. Radiat Med (2008) 26:455–465. https://doi.org/10.1007/s11604-008-0259-2

Patton MQ, Campbell-Patton C (2021) Utilization-focused evaluation, 5th edn. Sage, Thousand Oaks

Prevedello LM, Erdal B, Ryu JL et al (2017) Automated critical test findings identification and online notification system using artificial intelligence. Radiology 285:923–931

Rosenkrantz AB, Hughes DR, Duszak R Jr. The U.S. Radiologist Workforce: An Analysis of Temporal and Geographic Variation by Using Large National Datasets. Radiology. 2016 Apr;279(1):175-84. https://doi.org/10.1148/radiol.2015150921. Epub 2015 Oct 28. PMID: 26509294.

Sayres R, Taly A, Rahimy E, Blumer K, Coz D, Hammel N, Krause J, Narayanaswamy A, Rastegar Z, Wu D, Xu S, Barb S, Joseph A, Shumski M, Smith SA, Corrado G, Peng L, Webster D (2018) Using a deep learning algorithm and integrated gradients explanation to assist grading for diabetic retinopathy. Am Acad Ophthalmol 126:552–564. https://doi.org/10.1016/j.ophtha.2018.11.016

Small Z (2023) Black artists see clear bias in A.I. New York times (New York edition), July 5, section C, page 1

Small WR, Wiesenfeld B, Brandfield-Harvey B et al (2024) Large language model–based responses to patients' in-basket messages. JAMA Netw Open 7(7):e2422399. https://doi.org/10.1001/jamanetworkopen.2024.22399

Somashekhar SP, Sepulveda MJ, Puglielli S et al (2018) Watson for oncology and breast cancer treatment recommendations: agreement with an expert multidisciplinary tumor board. Ann Oncol 29:1–6

Stiepan D (2022) Using AI in radiology clinical practice. Mayo Clinic, September 14, 2022

Swarns C (2023) When artificial intelligence gets it wrong. Innocence Project. https://innocenceproject.org/when-artificial-intelligence-gets-it-wrong/

Teare PFM, Benzaquen O, Toledano E, Elnekave E (2017) Malignancy detection on mammography using dual deep convolutional neural networks and genetically discovered false color input enhancement. J Digit Imaging 30:499–505

Topol E (2023) Can AI catch what doctors miss? TED Talk. YouTube. https://www.youtube.com/watch?v=ll5LY7wI_Xc&t=306s

UNESCO (2024) Challenging systematic prejudices: an investigation into bias against women and girls in large language models. https://unesdoc.unesco.org/ark:/48223/pf0000388971

United States Food and Drug Administration (2018) FDA permits marketing of clinical decision support software for alerting providers of a potential stroke in patients

Villasenor J (2019) Artificial intelligence and bias: four key challengers. Brookings. https://www.brookings.edu/articles/artificial-intelligence-and-bias-four-key-challenges/

Wallace M et al (2022) Impact of artificial intelligence on miss rate of colorectal neoplasia. Gastroenterology 163:295

Watkins SC (2024) Dr. S. Craig Watkins on why AI's potential to combat or scale systemic injustice still comes down to humans. https://chartable.com/podcasts/unlocking-us-with-brene-brown/episodes/165526994-dr-s-craig-watkins-on-why-ais-potential-to-combat-or-scale-systemic-injustice-still-comes-down-to-humans

Weiss CH (1995) Nothing as practical as good theory. Aspen Institute, Washington

Zewe A (2023) Explained: generative AI. How do powerful generative AI systems like ChatGPT work, and what makes them different from other types of artificial intelligence? MIT News. https://news.mit.edu/2023/explained-generative-ai-1109

Zhao Q, Jia Q, Chi T (2022) Deep learning as a novel method for endoscopic diagnosis of chronic atrophic gastritis: a prospective nested case–control study. BMC Gastroenterol 22:352. https://doi.org/10.1186/s12876-022-02427-2