

ARTICLE OPEN



Autonomous artificial intelligence increases real-world specialist clinic productivity in a cluster-randomized trial

Michael D. Abramoff^{1,2,3,4,5}✉, Noelle Whitestone⁶, Jennifer L. Patnaik^{6,7}, Emily Rich^{6,8}, Munir Ahmed⁹, Lutful Husain⁹, Mohammad Yeadul Hassan⁹, Md. Sajidul Huq Tanjil¹⁰, Dena Weitzman², Tinglong Dai^{11,12,13}, Brandie D. Wagner^{7,14}, David H. Cherwek⁶, Nathan Congdon^{6,8,15} and Khairul Islam¹⁰

Autonomous artificial intelligence (AI) promises to increase healthcare productivity, but real-world evidence is lacking. We developed a clinic productivity model to generate testable hypotheses and study design for a preregistered cluster-randomized clinical trial, in which we tested the hypothesis that a previously validated US FDA-authorized AI for diabetic eye exams increases clinic productivity (number of completed care encounters per hour per specialist physician) among patients with diabetes. Here we report that 105 clinic days are cluster randomized to either intervention (using AI diagnosis; 51 days; 494 patients) or control (not using AI diagnosis; 54 days; 499 patients). The prespecified primary endpoint is met: AI leads to 40% higher productivity (1.59 encounters/hour, 95% confidence interval [CI]: 1.37–1.80) than control (1.14 encounters/hour, 95% CI: 1.02–1.25), $p < 0.00$; the secondary endpoint (productivity in all patients) is also met. Autonomous AI increases healthcare system productivity, which could potentially increase access and reduce health disparities. ClinicalTrials.gov NCT05182580.

npj Digital Medicine (2023)6:184; <https://doi.org/10.1038/s41746-023-00931-7>

INTRODUCTION

Lack of access to essential services is a primary cause of health inequity¹. In the United States (US), racial and ethnic minorities, persons with low socioeconomic status, and rural populations are especially affected, and worldwide, an estimated one billion people lack access to essential health services^{2,3}. This inequitable distribution continues to blunt global economic growth and inhibit living standards⁴.

Access can be improved by increasing the overall capacity of the healthcare system. One option is to expand the health workforce⁵; however, training more healthcare professionals at scale requires substantial resources and time, which may not be feasible⁶. Another option is to increase capacity by increasing efficiency⁷. Consistent gains in total factor productivity over the past century, especially in the agricultural and nonfarm industrial sectors, have substantially improved living standards^{8,9}. By contrast, clinic productivity, measured as the number of completed care encounters per hour per physician¹⁰, may actually be declining in the United States (US) (Fig. 1), with similar declines observed in other countries¹¹. This widening healthcare productivity gap has been suggested as a cause of rising healthcare costs⁹.

We hypothesize that autonomous Artificial Intelligence (AI), where a computer rather than a human provider makes the medical decision, can improve clinic productivity as defined above¹². Such autonomous AI systems have recently been approved by the US Food and Drug Administration (FDA), as safe and effective for use in medical care^{13,14} and as reimbursable by Medicare, Medicaid and private insurance payors^{15,16}. However, the potential productivity impact of autonomous AI systems has received scant attention. The

purpose of the present study is to test this hypothesis in a preregistered, randomized controlled (clinical) trial.

RESULTS

All specialists in the clinic ($n = 3$, 100% male, mean 5.17 years of practice (Standard Deviation [SD]: 3.33)) were included. There were 51 clinic days in the intervention group and 54 in the control group. The average number of clinic patients per day was 54.5. The number of clinic patients with diabetes was 2109, of which 1189 and 920 were in the intervention and control groups, respectively (Fig. 2). Among 2109 patients with diabetes, 993 (mean age 50.9 years (SD: 9.86), 47.2% male) were AI eligible, all of whom gave written consent and completed the autonomous AI exam, with 494 patient participants (49.7%) in the intervention group, and 499 (50.3%) in the control group (Table 1).

Primary outcome

The primary outcome, productivity λ_d among patients with diabetes, was significantly greater in the intervention group ($\lambda_{d,AI}$: 1.59 completed care encounters per hour per specialist physician, 95% confidence interval [CI]: 1.37–1.80) than in the control group ($\lambda_{d,C}$: 1.14, 95% CI: 1.02–1.25), Student's t -test $p < 0.001$ (Table 2). This corresponds to an increase of 0.45 completed care encounters per hour per specialist physician, or 39.5%.

The linear regression model showed a significant association between membership in the intervention group and productivity in univariate analysis (beta = 0.449 (SE: 0.120), $p < 0.001$). Results from the sensitivity analysis that included adjustment for age, sex,

¹University of Iowa, Iowa City, Iowa, USA. ²Digital Diagnostics Inc, Coralville, Iowa, USA. ³Iowa City Veterans Affairs Medical Center, Iowa City, Iowa, USA. ⁴Department of Biomedical Engineering, The University of Iowa, Iowa City, USA. ⁵Department of Electrical and Computer Engineering, The University of Iowa, Iowa City, Iowa, USA. ⁶Orbis International, New York, New York, USA. ⁷Department of Ophthalmology, University of Colorado School of Medicine, Aurora, Colorado, USA. ⁸Centre for Public Health, Queen's University Belfast, Belfast, UK. ⁹Orbis Bangladesh, Dhaka, Bangladesh. ¹⁰Deep Eye Care Foundation, Rangpur, Bangladesh. ¹¹Carey Business School, Johns Hopkins University, Baltimore, Maryland, USA. ¹²Hopkins Business of Health Initiative, Johns Hopkins University, Baltimore, Maryland, USA. ¹³School of Nursing, Johns Hopkins University, Baltimore, Maryland, USA. ¹⁴Department of Biostatistics and Informatics, Colorado School of Public Health, Aurora, Colorado, USA. ¹⁵Zhongshan Ophthalmic Center, Sun Yat-sen University, Guangzhou, China. ✉email: michael-abramoff@uiowa.edu

day of week, and autonomous AI output confirmed this association ($\beta = 0.461$ (SE: 0.118), $p < 0.001$) (Table 3). The diagnostic output of autonomous AI and the day of the week were

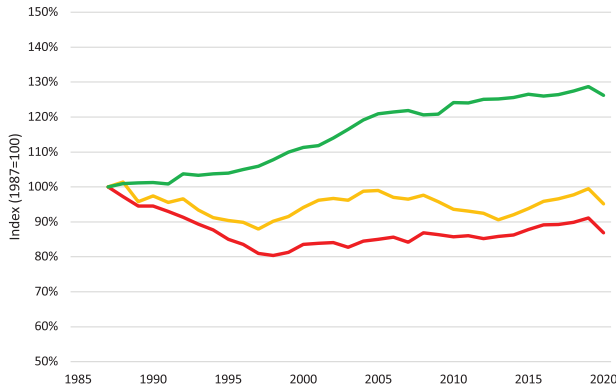


Fig. 1 Productivity changes: 1987–2020. US healthcare productivity declined over the last three decades, at the same time that US productivity overall, commonly expressed as “private nonfarm business total factor productivity,” increased by 26.2% between 1987 and 2020. Over this same period, productivity in ambulatory healthcare services declined by 13.2%. One contributor to this growing difference is the loss of labor productivity in ambulatory healthcare services during this same period (with a total decline of 4.9%). (The productivity data was provided by the US Department of Labor, Bureau of Labor Statistics, and graphed with assistance from the Office of Productivity, Bureau of Labor Statistics. Industry data prior to 1987 is unavailable on a consistent classification basis). The red line is Ambulatory Healthcare Total Factor Productivity. The yellow line is Ambulatory Healthcare Labor Productivity. The green line is Private Nonfarm Business Total Factor Productivity.

associated with the primary outcome but had minimal impact on the primary measure of association.

Secondary outcomes

The secondary outcome of productivity λ over all patients (with and without diabetes) was also significantly greater in the intervention group ($\lambda_{AI} 4.05$, 95% CI: 3.66–4.43) than in the control group ($\lambda_{d,c} 3.36$, 95% CI: 3.08–3.63), Student’s t -test $p = 0.004$.

Specialist productivity adjusted for patient complexity, for diabetes patients, was also significantly greater in the intervention group ($\lambda_{caAI} = 3.15$) than in the control group ($\lambda_{ca,d,c} = 1.19$). Table 2 corresponds to an increase by a factor of 2.65.

Patient participants were satisfied with the appointment waiting time (100% satisfied or very satisfied) and the interaction with the healthcare team (499/499 = 100% in the control group and 493/494 = 99.8% in the intervention group). Among patient participants in the intervention group who completed their care encounter through autonomous AI only ($n = 331$, 67.0%), 100% were satisfied or very satisfied with the time to receive results, and 100% were satisfied with receiving results from an autonomous AI system. Among the specialist participants, all “agreed” or “strongly agreed” that autonomous AI saved time in their clinics, and all “agreed” or “strongly agreed” that autonomous AI allowed them to focus their time on appropriate patients.

The number of DED treatments scheduled per day did not differ between the control (0.70, 95% CI: 0.47–0.93) and intervention (0.61, 95% CI: 0.38–0.83, Wilcoxon rank sum test $p = 0.532$) groups, nor did the patient complexity score (mean score $1.06 \pm SD: 2.36$ vs $0.949 \pm SD: 2.26$, Wilcoxon rank sum test $p = 0.288$). When analyzing complexity for only those patient participants who required a specialist examination after completion of the autonomous AI exam, the mean complexity score was significantly

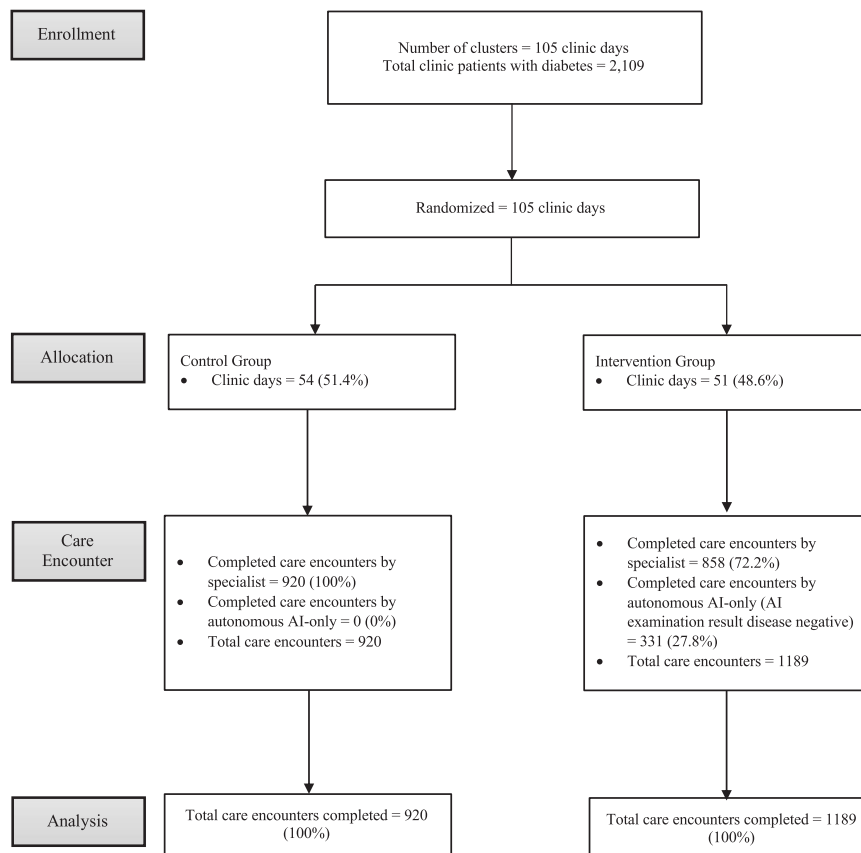


Fig. 2 Cluster randomization flow chart showing clinic days and patients in the control or intervention group, according to CONSORT.

higher in the intervention group ($2.80 \pm \text{SD: } 3.19$) than in the control group ($1.06 \pm \text{SD: } 2.36$, Wilcoxon rank sum test $p < 0.0001$). The estimated sensitivity of the Autonomous AI system compared

to the level 4 reference standard (human graders) was 93.9% (95% CI: 90.5, 97.2), and the estimated specificity was 84.0% (95% CI: 81.4, 86.7).

DISCUSSION

B-PRODUCTIVE confirmed our primary hypothesis: the use of autonomous AI systems significantly improves clinic productivity (λ) in the real world¹². The healthcare productivity gap has been underappreciated as an issue in healthcare, leading to health inequities along racial, ethnic and geographic lines; reduced access to high-quality care; and increasing healthcare expenditures, despite cost-saving measures, such as rationing, which may, in turn, diminish the quality of care^{17,18}.

The importance of increasing productivity as a potential solution to these issues has also received scant attention¹⁹. Increasing worker productivity has been highly successful in other sectors of the economy⁸ but has been challenging in healthcare²⁰. For example, while information technology has facilitated substantial productivity growth in other sectors²¹, there is evidence that innovations such as electronic medical records may lower healthcare labor productivity in some cases²². Other causes of the productivity gap may be increased regulatory requirements and the resulting documentation burden, as well as the increasing complexity of clinical information systems, though these are beyond the scope of this study.

The autonomous AI system used in the current study was developed and validated under a strict ethical framework^{23,24}, outperformed physician accuracy to the same prognostic standard in clinical trials¹³, shows no racial or ethnic bias^{13,25} as also demonstrated in Hansen et al.²⁶, is explainable¹², is highly effective for outcomes²⁷, is supported by all US healthcare stakeholders¹⁵, and can reduce the cost of care^{15,27}. While the accepted reference standard for validating specialist clinicians and autonomous AI is the prognostic ETDRS and DRCR standards (a level 1 reference standard)²⁴, confirming that the autonomous AI used in this study has much higher accuracy than human specialists in the US population^{13,24}, the present results show the AI's high accuracy in this Bangladeshi study population, where AI had not been tested previously. The existing evidence, combined with the present findings, show that autonomous AI can increase clinic productivity at equivalent or higher quality of care, in contrast to other cost-saving measures such as rationing or substitution¹⁷.

In B-PRODUCTIVE, specialists reported that autonomous AI allowed them to focus their time on more complex cases, as reflected in the mean complexity score in the intervention group,

Table 1. Characteristics of clinic day clusters and patient participants by study group.

Characteristic	Control group	Intervention group
Clinic days		
Number	54	51
Patients total (including non-participants)		
Number	2708	3013
Patient participants		
Number	499	494
Age, years		
22–40	81 (16.2%)	86 (17.4%)
41–50	183 (36.7%)	167 (33.8%)
51–60	158 (31.7%)	170 (34.4%)
>60	77 (15.4%)	71 (14.4%)
Mean (SD)	51.0 (10.0)	50.8 (9.70)
Male sex, <i>n</i> (%)	234 (46.9%)	235 (47.6%)
Education		
No education	137 (27.4%)	114 (23.1%)
Non-graded religious school	102 (20.4%)	123 (24.9%)
Primary school only	237 (47.5%)	231 (46.8%)
Secondary school	23 (4.61%)	26 (5.26%)
Monthly income ^a		
\$50–\$150	70 (14.0%)	64 (13.0%)
\$151–\$250	122 (24.4%)	137 (27.7%)
\$251–\$500	196 (39.3%)	181 (36.6%)
>\$500	111 (22.2%)	112 (22.7%)
Patient autonomous AI output		
DED present	167 (33.5%)	140 (28.3%)
DED absent	321 (64.3%)	331 (67.0%)
Insufficient quality	11 (2.20%)	23 (4.66%)

DED : referable Diabetic Eye Disease: ETDRS level 35 or higher, clinically significant macular edema, and/or center-involved macular edema.

^aMonthly income reported in taka and converted to USD at a conversion rate of 0.01057 as of July 29, 2022 (<https://www.xe.com/currencycharts/?from=BDT&to=USD>).

Table 2. Productivity outcomes by study group.

	Control group mean (95% CI)	Intervention group mean (95% CI)
Completed care encounters among clinic patients with diabetes		
Care encounter involved specialist	920	858
Care encounter completed by AI-only	0	331
Total	920	1189
Total number of specialist hours in clinic	819	763
Clinic productivity (95% CI) for diabetes patients: number of completed care encounters per hour per specialist physician ^a	$\lambda_{d,c} = 1.14$ (1.02, 1.25)	$\lambda_{d,AI} = 1.59$ (1.3, 1.80)
Clinic productivity (95% CI) for all patients number of completed care encounters per hour per specialist physician ^b	$\lambda_c = 3.36$ (3.08, 3.63)	$\lambda_{AI} = 4.05$ (3.66, 4.43)
Specialist productivity adjusted for patient complexity for diabetes patients	$\lambda_{ca,d,c} = 1.19$	$\lambda_{caAI} = 3.15$

^aStudent's *t*-test $p < 0.001$ for between-group difference.

^bStudent's *t*-test $p = 0.004$ for between-group difference.

Table 3. Potential predictors of main outcome, provider productivity assessed as number of completed clinic visits among patients with diabetes per specialist per hour.

Potential predictor	β (SE) ^a	<i>p</i> -value
Membership in intervention group	0.449 (0.120)	0.0002
Patient-level factors		
Patient age, years	−0.00000003 (0.00000006)	0.607
Patient sex, female	0.000001 (0.000001)	0.400
Patient no education	0.0000003 (0.000002)	0.826
Patient monthly income, USD	−0.000000007 (0.000000002)	0.727
Clinic-level factors		
Day of the week		
Sunday	−0.33 (0.24)	0.174
Monday	−0.20 (0.26)	0.433
Tuesday	−0.18 (0.27)	0.504
Wednesday	−0.49 (0.24)	0.040
Thursday	−0.67 (0.23)	0.003
Friday	Closed	—
Saturday	Reference	—
AI diagnostic output		
DED present	0.0000005 (0.000001)	0.697
DED absent	Reference	—
Insufficient quality	−0.000003 (0.000002)	0.059
Complexity sum	0.0000004 (0.0000002)	0.134

^aBeta coefficients and standard errors (SE) from linear regression model with generalized estimating equations that included clustering effects of clinic days.

which was significantly higher than in the control group. Given the large proportion of patients who were able to avoid the wait to see a specialist as a result of receiving their examination from the AI, the net effect of the autonomous AI visible to patients was to reduce wait time. This benefit would be especially likely to lead to improved satisfaction in settings where wait times comprise some of the most common patient complaints.

Some three-quarters of patient participants in the intervention group completed their care encounters through the autonomous AI system only. Productivity ($\lambda_{d,AI}$) increased by 40% in the intervention group because non-AI-eligible diabetes patients filled the 331 clinic spots that became available when eligible patients were identified by autonomous AI as “DED absent” and thereby completed their clinic encounter. Productivity $\lambda_{d,AI}$ did not achieve its upper bound, as the average complexity of the patients evaluated by specialists increased. If the prevalence π of DED in the patient participants had been lower, or the proportion of diabetes patients who were AI-eligible (α) had been higher, productivity $\lambda_{d,AI}$ would have increased even further. When the shift to more complex patients for the specialist due to the use of autonomous AI was taken into account, by calculating specialist productivity adjusted for complexity, autonomous AI increased productivity by a factor of 2.65.

We based our study design on our mathematical productivity model, using a concealed cluster-randomized design, in a clinic context where demand is overwhelmingly greater than clinic capacity ($\Lambda \gg \mu$). This was done to minimize bias by schedulers, clinic staff, patients, or specialists. This productivity hypothesis testing study design would not have been possible in a scheduled outpatient clinic context: in such clinics, the schedulers would fix any measured productivity gains, as they would have to make

additional slots available on intervention dates, the number of additional slots determined by their expectation of gains in λ_{AI} , not by true λ_c . Similarly, masking clinic staff, patients or specialists, whether or not the AI diagnosis was being used, mitigated bias from those sources.

Potentially, while specialists were masked to whether or not a day was an intervention day, they could potentially have determined that from their perceived average patient complexity on that day. If that was the case, it would have biased against AI, as specialists would spend more time with these more complex patients.

Autonomous AI systems have particular advantages in under-resourced settings, most obviously, the benefit of improved productivity where trained personnel is scarce. While telemedicine platforms have been implemented in some cases, these do not allow instantaneous, point-of-care diagnosis, so that the care encounter cannot be completed in the same visit. The reason is that while the patient images can be taken in the clinic, the diagnostic result will only be available after the patient has already left the clinic, resulting in care completion rates of 30%, at lower diagnostic accuracy²⁷. Implementation of the AI system, including operator training, was delivered remotely. This suggests these AI systems are scalable and sustainable, especially in low- and middle-income countries, further strengthened by the high participant and provider satisfaction.

Limitations of the current study are that B-PRODUCTIVE was conducted in a single health system, in a low-income country, with only three physician specialists, and using an autonomous AI designed to diagnose only a single disease, DED, in patients without symptoms or a history of DED. While it was conducted in a single health system, the results from our mathematical model of healthcare productivity have implications for other health systems that are characterized by a ‘saturated queue’ (i.e., without schedules or appointment slots). While the autonomous AI diagnosed only DED, this complication of diabetes is of particular economic importance as the leading cause of vision loss among working-age people worldwide, including in Bangladesh²⁸. The autonomous AI system, in addition to being validated by the US FDA, EU CE mark, and various other national regulatory agencies, with respect to its safety, efficacy and lack of racial bias, was also evaluated on the Bangladeshi patient population by comparison of the AI output to a UK NHS-certified retina expert. While the autonomous AI is only validated for patients without symptoms or a history of DED, the majority of patients visiting the retina specialist fall into this category. Application of these results to other settings, conditions and AI systems must be made with caution, and further studies are needed to extend these findings more broadly. Application of these results to other settings, conditions and AI systems must be made with caution, and further studies are needed to extend these findings more broadly.

Strengths of the current study include the model-based hypothesis testing; the preregistered, randomized design; real-world²⁹ implementation in a lower-income country where productivity gains among scarce specialists are particularly relevant; and the collection of data on patient and provider satisfaction.

In summary, the use of an autonomous AI system improved clinic productivity by 40% in the B-PRODUCTIVE trial. Autonomous AI systems can play an important role in addressing global health disparities by improving access to affordable, high-quality care, especially in low- and middle-income countries.

METHODS

Theoretical foundation of unbiased estimation of healthcare productivity

To test our central hypothesis—that autonomous AI improves healthcare system productivity—in an unbiased manner, we

developed a healthcare productivity model based on rational queueing theory³⁰, as widely used in the healthcare operations management literature³¹. A healthcare provider system, which can be a hospital, an individual physician providing a service, an autonomous AI providing a service at a performance level at least or higher than a human expert, a combination thereof, or a national healthcare system, are all modeled as an “overloaded queue,” facing a potential demand that is greater than its capacity; that is, $\Lambda \gg \mu$, where Λ denotes the total demand on the system—patients seeking care—and μ denotes the maximum number of patients the system can serve per unit of time. We define system productivity as

$$\lambda = \frac{n_q}{t}, \quad (1)$$

where n_q is the number of patients who completed a care encounter with a quality of care that was non-inferior to q , and t is the length of time over which n_q was measured, allowing for systems that include autonomous AI in some fashion. While the standard definitions of healthcare labor productivity, such as in Camasso et al.⁷, ignore quality of care, q assumes quality of care non-inferior to the case when care is provided by a human expert, such as a retina specialist, to address potential concerns about the safety of healthcare AI⁸. Our definition of λ , as represented by Eq. (1), guarantees that quality of care is either maintained or improved.

β denotes the proportion of patients who receive and complete the care encounter in a steady state, where the average number of patients who successfully complete the care encounter is equal to the average number of patients who gain access to care, per unit of time, in other words, $\lambda = \beta \cdot \Lambda$. See Fig. 3. Remember that in the overloaded queue model, there are many patients $1-\beta \cdot \Lambda$ who do not gain access. β is agnostic about the specific manner in which access is determined: β may take the form of a hospital administrator who establishes a maximum number of patients admitted to the system or in the form of barriers to care—such as an inability to pay, travel long distances, take time off work or other sources of health inequities—limiting a patient gaining access to the system. As mentioned, λ is agnostic on whether the care encounter is performed and completed by an autonomous AI, human providers, or a combination thereof, as from the patient perspective, we measure the number of patients that complete the appropriate level of care per unit time at a performance level at least or higher than human physician. Not every patient will be eligible to start their encounter with autonomous AI, and we denote by α , $0 < \alpha < 1$ the proportion of eligible patients, for example, because they do not fit the inclusion criteria for the

autonomous AI; not every patient will be able to complete their care encounter with autonomous AI when the autonomous AI diagnosed them with disease requiring a human specialist, and we denote by γ , $0 < \gamma < 1$, the proportion of patients who started their care encounter with AI, and still required a human provider to complete their encounter. The proportion $\alpha(1-\gamma)$ are diagnosed as “disease absent” and start and complete their encounter with autonomous AI only, without needing to see a human provider. For all permutations, productivity λ is measured as the number of patients who complete a provided care encounter per unit of time, with λ_c , the productivity associated with the control group, where the screening result of the AI system is not used to determine the rest of the care process, and λ_{AI} , the productivity associated with the intervention group, where the screening result of the AI system is used to determine the rest of the care process, and where the AI performance is at least as high as the human provider.

Because an autonomous AI that completes the care process for patients without disease—typically less complex patients—as in the present study, will result in relatively more complex patients to be seen by the human specialist, we calculate complexity-adjusted specialist productivity as

$$\lambda_{ca} = \frac{\bar{c}n_q}{t}, \quad (2)$$

with \bar{c} the average complexity, as determined with an appropriate method, for all n patients that complete the care encounter with that specialist. This definition of λ_{ca} , as represented by Eq. (2), corrects for a potentially underestimated productivity because the human specialist sees more clinically complex patients requiring more time than without the AI changing the patient mix.

We focus on the implication $\Lambda \gg \mu$; in other words, that system capacity is limited relative to potential demand, as that is the only way in which λ_c and λ_{AI} can be measured without recruitment bias, i.e., in a context where patients arrive throughout the day without appointment or other filter, as is the case in Emergency Departments in the US, and almost all clinics in low- and middle-income countries (LMICs). This is not the case, however, in contexts where most patient visits are scheduled, and thus β cannot be changed dynamically, and measuring λ in such a context would lead to bias. Thus, we selected a clinic with a very large demand (Λ), Deep Eye Care Foundation (DECF) in Bangladesh, for the trial setting in order to avoid recruitment bias.

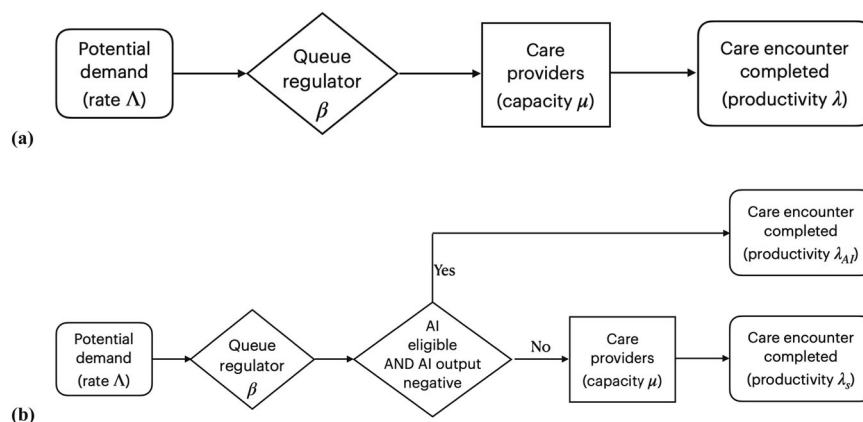


Fig. 3 Healthcare productivity model based on rational queueing theory. **a** Mathematical model of ‘overloaded queue’ healthcare system in order to estimate productivity as $\lambda = \beta \Lambda$. without observer bias. **b** Model of ‘overloaded queue’ healthcare system where autonomous AI is added to the workflow.

Trial design

The B-PRODUCTIVE (Bangladesh-PRODUCTIVITY in Eyecare) study was a preregistered, prospective, double-masked, cluster-randomized clinical trial performed in retina specialist clinics at DECF, a not-for-profit, non-governmental hospital in Rangpur, Bangladesh, between March 20 and July 31, 2022. The clusters were specialist clinic days, and all clinic days were eligible during the study period. Patients are not scheduled; there are no pre-scheduled patient visit times or time slots, instead access to a specialist clinic visit is determined by clinic staff on the basis of observed congestion, as explained in the previous Section.

The study protocol was approved by the ethics committees at the Asian Institute of Disability and Development (Dhaka, Bangladesh; # Southasia-hrec-2021-4-02), the Bangladesh Medical Research Council (Dhaka, Bangladesh; # 475 27 02 2022) and Queen's University Belfast (Belfast, UK; # MHLS 21_46). The tenets of the Declaration of Helsinki were adhered to throughout, and the trial was preregistered with ClinicalTrials.gov, #NCT05182580, before the first participant was enrolled. The present study included local researchers throughout the research process, including design, local ethics review, implementation, data ownership and authorship to ensure it was collaborative and locally relevant.

Autonomous AI system

The autonomous AI system (LumineticsCore (formerly IDX-DR), Digital Diagnostics, Coralville, Iowa, USA) was designed, developed, previously validated and implemented under an ethical framework to ensure compliance with the principles of patient benefit, justice and autonomy, and avoid "Ethics Dumping"¹³. It diagnoses specific levels of diabetic retinopathy and diabetic macular edema (Early Treatment of Diabetic Retinopathy Study level 35 and higher), clinically significant macular edema, and/or center-involved macular edema³², referred to as "referable Diabetic Eye Disease" (DED)³³, that require management or treatment by an ophthalmologist or retina specialist, for care to be appropriate. If the ETDRS level is 20 or lower and no macular edema is present, appropriate care is to retest in 12 months³⁴. The AI system is autonomous in that the medical diagnosis is made solely by the system without human oversight. Its safety, efficacy, and lack of racial, ethnic and sex bias were validated in a pivotal trial in a representative sample of adults with diabetes at risk for DED, using a workflow and minimally trained operators comparable to the current study¹³. This led to US FDA De Novo authorization ("FDA approval") in 2018 and national reimbursement in 2021^{13,15}.

Autonomous AI implementation and workflow

The autonomous AI system was installed by DECF hospital information technology staff on March 2, 2022, with remote assistance from the manufacturer. Autonomous AI operators completed a self-paced online training module on basic fundus image-capture and camera operations (Topcon NW400, Tokyo, Japan), followed by remote hands-on training on the operation by representatives of the manufacturers. Deployment was performed locally, without the physical presence of the manufacturers, and all training and support were provided remotely.

Typically, pharmacologic pupillary dilation is provided only as needed during use of the autonomous AI system. For the current study, all patient participants received pharmacologic dilation with a single drop each of tropicamide 0.8% and phenylephrine 5%, repeated after 15 min if a pupil size of ≥ 4 mm was not achieved. This was done to facilitate indirect ophthalmoscopy by the specialist participants as required. The autonomous AI system guided the operator to acquire two color fundus images determined to be of adequate quality using an image quality assessment algorithm, one each centered on the fovea and the optic nerve, and directed the operator to retake any images of insufficient quality. This process

took approximately 10 min, after which the autonomous AI system reported one of the following within 60 s: "DED present, refer to specialist", "DED not present, test again in 12 months", or "insufficient image quality". The latter response occurred when the operator was unable to obtain images of adequate quality after three attempts.

Participants

This study included both physician participants and patient participants. Physician participants were retina specialists who gave written informed consent prior to enrollment. For specialist participants, the inclusion criteria were:

- Completed vitreoretinal fellowship training;
- Examined ≥ 20 patients per week with diabetes and no known DED over the prior three months;
- Performed laser retinal treatments or intravitreal injections on at least three DED patients per month over the same time period.

Exclusion criteria were:

- Using a clinical AI system integrated in their practice
- Inability to provide informed consent.

'AI-eligible patients' are clinic patients meeting the following criteria:

- Presenting to DECF for eye care;
- Age ≥ 22 years. While preregistration stated participants could be aged ≥ 18 years, the US FDA De Novo clearance for the autonomous AI limits eligibility to ≥ 22 years;
- Diagnosis of type 1 or type 2 diabetes prior to or on the day of recruitment;
- Best corrected visual acuity $\geq 6/18$ in the better-seeing eye;
- No prior diagnosis of DED;
- No history of any laser or incisional surgery of the retina or injections into either eye;
- No medical contraindication to fundus imaging with dilation of the pupil¹².

Exclusion criteria were:

- Inability to provide informed consent or understand the study;
- Persistent vision loss, blurred vision or floaters;
- Previously diagnosed with diabetic retinopathy or diabetic macular edema;
- History of laser treatment of the retina or injections into either eye or any history of retinal surgery;
- Contraindicated for imaging by fundus imaging systems.

Patient participants were AI-eligible patients who gave written informed consent prior to enrollment. All eligibility criteria remained unchanged over the duration of the trial.

Randomization, masking and concealment

B-PRODUCTIVE was a concealed cluster-randomized trial in which a block randomization scheme by clinic date was generated by the study statistician (JP) on a monthly basis, taking into account holidays and scheduled clinic closures. The random allocation of each cluster (clinic day) was concealed until clinic staff received an email with this information just before the start of that day's clinic, and they had no contact with the specialists during trial operations. Medical staff who determined access, specialists and patient participants remained masked to the random assignment of clinic days as control or intervention.

Intervention

After giving informed consent, patient participants provided demographic, income, educational and clinical data to study staff using an orally administered survey in Bangla, the local language.

Patients who were eligible but did not consent underwent the same clinical process without completing an autonomous AI diagnosis or survey. All patient participants, both intervention and control, completed the autonomous AI diagnostic process as described in the Autonomous AI implementation and workflow section above: the difference between intervention and control groups was that in the intervention group, the diagnostic AI output determined what happened to the patient next. In the control group, patient participants always went on to complete a specialist clinic visit after autonomous AI, irrespective of its output. In the intervention group, patient participants with an autonomous AI diagnostic report of “DED absent, return in 12 months” completed their care encounters without seeing a specialist and were recommended to make an appointment for a general eye exam in three months as a precautionary measure for the trial, minimizing the potential for disease progression (standard recall would be 12 months).

In the intervention group, patient participants with a diagnostic report of “DED present” or “image quality insufficient” completed their care encounters by seeing the specialist for further management. “Seeing the specialist” for not-consented, control group, and “DED present / insufficient” patient participants involved tonometry, anterior and posterior segment biomicroscopy, indirect ophthalmoscopy, and any further examinations and ancillary testing deemed appropriate by the specialist. After the patient participant completed the autonomous AI process, a survey with a 4-point Likert scale (“very satisfied,” “satisfied,” “dissatisfied,” “very dissatisfied”) was administered concerning the participant’s satisfaction with interactions with the healthcare team, time to receive examination results, and receiving their diagnosis from the autonomous AI system.

Study outcomes

The primary outcome was clinic productivity for diabetes patients (λ_d), measured as the number of completed care encounters per hour per specialist for control / non-AI ($\lambda_{d,C}$) and intervention / AI ($\lambda_{d,AI}$) days. $\lambda_{d,C}$ used the number of completed specialist encounters; $\lambda_{d,AI}$ used the number of eligible patients in the intervention group who completed an autonomous AI care encounter with a diagnostic output of “DED absent”, plus the number of encounters that involved the specialist exam. For the purposes of calculating the primary outcome, all diabetes patients who presented to the specialty clinic on study days were counted, including those who were not patient participants or did not receive the autonomous AI examination.

One of the secondary outcomes from this study was λ for all patients (patients both with and without diabetes) measured as the number of completed care encounters per hour per specialist by counting all patients presenting to the DECF specialty clinic on study days, including those without diabetes, for control (λ_C) and intervention days (λ_{AI}). *Complexity-adjusted specialist productivity* λ_{ca} was calculated for intervention and control arms by multiplying ($\lambda_{d,C}$) and ($\lambda_{d,AI}$) by the average patient complexity \bar{c} .

During each clinic day, the study personnel recorded the day of the week and the number of hours that each specialist participant spent in the clinic, starting with the first consultation in the morning and ending when the examination of the last patient of the day was completed, including any time spent ordering and reviewing diagnostic tests and scheduling future treatments. Any work breaks, time spent on performing procedures, and other duties performed outside of the clinic were excluded. Study personnel obtained the number of completed clinic visits from the DECF patient information system after each clinic day.

At baseline, specialist participants provided information on demographic characteristics, years in specialty practice and patient volume. They also completed a questionnaire at the end of the study, indicating their agreement (5-point Likert scale, “strongly agree” to “strongly disagree”) with the following statements regarding autonomous AI: (1) saves time in clinics, (2) allows time to be focused

on patients requiring specialist care, (3) increases the number of procedures and surgeries, and (4) improves DED screening.

Other secondary outcomes were (1) patient satisfaction; (2) number of DED treatments scheduled per day; and (3) complexity of patient participants. Patient and provider willingness to pay for AI was a preregistered outcome, but upon further review by the Bangladesh Medical Research Council, these data were removed based on their recommendation. The *complexity score* for each patient was calculated by a masked United Kingdom National Health Service grader using the International Grading system (a level 4 reference standard²⁴), adapted from Wilkinson et al. International Clinical Diabetic Retinopathy and Diabetic Macular Edema Severity Scales³¹ (no DED = 0 points, mild non-proliferative DED = 0 points, moderate or severe non-proliferative DED = 1 point, proliferative DED = 3 points and diabetic macular edema = 2 points.) The complexity score was summed across both eyes.

Power calculation

The null hypothesis was that the primary outcome parameter λ_d would not differ significantly between the study groups. The intra-cluster correlation coefficient (ICC) between patients within a particular cluster (clinic day) was estimated at 0.15, based on pilot data from the clinic. At 80% power, a two-sided alpha of 5%, a cluster size of eight patients per clinic day, and a control group estimated mean of 1.34 specialist clinic visits per hour (based on clinic data from January to March 2021), a sample size of 924 patients with completed clinically-appropriate retina care encounters (462 in each of the two study groups) was sufficient to detect a between-group difference of 0.34 completed care encounters per hour per specialist (equivalent to a 25% increase in productivity $\lambda_{d,AI}$, with autonomous AI).

Statistical methods

Study data were entered into Microsoft Excel 365 (Redmond, WA, USA) by the operators and the research coordinator in DECF. Data entry errors were corrected by the Orbis program manager in the US (NW), who remained masked to study group assignment.

Frequencies and percentages were used to describe patient participant characteristics for the two study groups. Age as a continuous variable was summarized with the mean and standard deviation. The number of treatments and complexity score were compared with the Wilcoxon rank sum test since they were not normally distributed. The primary outcome was normally distributed and compared between study groups using a two-sided Student’s t-test, and 95% confidence intervals around these estimates were calculated.

The robustness of the primary outcome was tested by utilizing linear regression modeling with generalized estimating equations that included clustering effects of clinic days. The adjustment for clustering of days since the beginning of the trial utilized an autoregressive first-order covariance structure since days closer together were expected to be more highly correlated. Residuals were assessed to confirm that a linear model fit the rate outcome. Associations between the outcome and potential confounders of patient age, sex, education, income, complexity score, clinic day of the week, and autonomous AI output were assessed. A sensitivity analysis with multivariable modeling included patient age and sex, and variables with p -values < 0.10 in the univariate analysis. All statistical analyses were performed using SAS version 9.4 (Cary, North Carolina).

DATA AVAILABILITY

Data is available upon reasonable request to the corresponding author.

CODE AVAILABILITY

The underlying code for this study is not publicly available for proprietary reasons.

Received: 8 June 2023; Accepted: 20 September 2023;
Published online: 04 October 2023

REFERENCES

1. Lavizzo-Mourey, R. J., Besser, R. E. & Williams, D. R. Understanding and mitigating health inequities—past, current, and future directions. *N. Engl. J. Med.* **384**, 1681–1684 (2021).
2. Guilbert, J. J. The World Health Report 2006: working together for health. *Educ. Health* **19**, 385–387 (2006).
3. Zurn, P., Dal Poz, M. R., Stilwell, B. & Adams, O. Imbalance in the health workforce. *Hum. Resour. Health* **2**, 13 (2004).
4. Monaco, R. M. & Phelps, J. H. Health care prices, the federal budget, and economic growth. *Health Aff.* **14**, 248–259 (1995).
5. Chen, L. et al. Human resources for health: overcoming the crisis. *Lancet* **364**, 1984–1990 (2004).
6. O'Brien, P. & Gostin, L. Health worker shortages and global justice. *Health Worker Shortages and Global Justice, Millbank Memorial Fund.* (2011). <https://www.milbank.org/publications/health-worker-shortages-and-global-justice/>.
7. Hussey, P. S. et al. A systematic review of health care efficiency measures. *Health Serv. Res.* **44**, 784–805 (2009).
8. Backman, J. & Gainsbrugh, M. Productivity and living standards. *Ind. Labor Relat. Rev.* **2**, 163–194 (1949).
9. Baumol, W. J. *The Cost Disease: Why Computers Get Cheaper and Health Care Doesn't* (Yale University Press, 2012).
10. US Bureau of Labor Statistics. *Major Sector Productivity and Costs*, Vol. 2017 (US Department of Labor, Bureau of Labor Statistics, 2017).
11. Kamarainen, V. J., Peltokorpi, A., Torkki, P. & Tallbacka, K. Measuring healthcare productivity—from unit to system level. *Int. J. Health Care Qual. Assur.* **29**, 288–299 (2016).
12. Frakt, A. B. Making health care more productive. *JAMA* **322**, 2274–2275 (2019).
13. Abramoff, M. D., Lavin, P. T., Birch, M., Shah, N. & Folk, J. C. Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. *Nat. Digit. Med.* **1**, 39 (2018).
14. U.S. Food & Drug Administration (FDA). *FDA Permits Marketing of Artificial Intelligence-based Device to Detect Certain Diabetes-related Eye Problems.* <https://www.fda.gov/newsevents/newsroom/pressannouncements/ucm604357.htm> (FDA, 2018).
15. Abramoff, M. D. et al. A reimbursement framework for artificial intelligence in healthcare. *NPJ Digit. Med.* **5**, 72 (2022).
16. Centers for Medicare & Medicaid Services. *Proposal to Establish Values for Remote Retinal Imaging (CPT code 92229) (Pages 56ff).* Medicare Program; CY 2022 Payment Policies under the Physician Fee Schedule (2021).
17. Lorenzoni, L., Marino, A., Morgan, D. & James, C. *Health Spending Projections to 2030: New Results Based on a Revised OECD Methodology.* OECD Working Papers no. 110 (2019).
18. Bevan, G. & Brown, L. D. The political economy of rationing health care in England and the US: the 'accidental logics' of political settlements. *Health Econ. Policy Law* **9**, 273–294 (2014).
19. Blumenthal, D., Stremikis, K. & Cutler, D. Health care spending—a giant slain or sleeping? *N. Engl. J. Med.* **369**, 2551–2557 (2013).
20. Cutler, D. M. Where are the health care entrepreneurs? The failure of organizational innovation in health care. *Innov. Policy Econ.* **11**, 1–28 (2011).
21. Oliner, S. D., Sichel, D. E. & Stiroh, K. J. *Explaining a Productive Decade.* FEDS Working Paper No. 2007-63 (2007).
22. Bhargava, H. K. & Mishra, A. *Electronic Medical Records and Physician Productivity: Evidence from Panel Data Analysis* (Social Science Research Network, 2011).
23. Char, D. S., Abramoff, M. D. & Feudtner, C. Identifying ethical considerations for machine learning healthcare applications. *Am. J. Bioeth.* **20**, 7–17 (2020).
24. Abramoff, M. D. et al. Foundational considerations for artificial intelligence using ophthalmic images. *Ophthalmology* **129**, e14–e32 (2022).
25. Abramoff, M. D. et al. Considerations for addressing bias in artificial intelligence for health equity. *NPJ Digit. Med.* **6**, 170 (2023).
26. Hansen, M. B. et al. Results of automated retinal image analysis for detection of diabetic retinopathy from the Nakuru Study, Kenya. *PLoS ONE* **10**, e0139148 (2015).
27. Channa, R., Wolf, R. M., Abramoff, M. D. & Lehmann, H. P. Effectiveness of artificial intelligence screening in preventing vision loss from diabetes: a policy model. *NPJ Digit. Med.* **6**, 53 (2023).
28. Sun, H. et al. IDF Diabetes Atlas: global, regional and country-level diabetes prevalence estimates for 2021 and projections for 2045. *Diabetes Res. Clin. Pr.* **183**, 109119 (2022).
29. Concato, J. & Corrigan-Curay, J. Real-world evidence—where are we now? *N. Engl. J. Med.* **386**, 1680–1682 (2022).
30. Hassin, R. *Rational Queuing* 1st edn (Chapman and Hall/CRC, 2016).
31. Dai, T. & Tayur, S. Healthcare operations management: a snapshot of emerging research. *Manuf. Serv. Oper. Manag.* **22**, 869–887 (2019).
32. Fundus photographic risk factors for progression of diabetic retinopathy. ETDRS report number 12. Early Treatment Diabetic Retinopathy Study Research Group. *Ophthalmology* **98**, 823–833 (1991).
33. Abramoff, M. D. et al. Approach for a clinically useful comprehensive classification of vascular and neural aspects of diabetic retinal disease. *Invest. Ophthalmol. Vis. Sci.* **59**, 519–527 (2018).
34. American Academy of Ophthalmology Retina/Vitreous Panel & Hoskins Center for Quality Eye Care. *Preferred Practice Patterns: Diabetic Retinopathy* (ed. American Academy of Ophthalmology Retina Panel) (American Academy of Ophthalmology, 2016).

ACKNOWLEDGEMENTS

We are grateful for the assistance provided by Mike Jadoo and Corby Garner, Office of Productivity, Bureau of Labor Statistics, US Department of Labor, Washington DC, and Ravi Parikh, MD, MPH, New York University School of Medicine, New York, New York. The project was supported by Orbis International and the Global Vision Foundation. Digital Diagnostics, Inc. provided the autonomous AI systems and training and technical support for operators, staff and specialists at no cost but did not fund any of the investigators or staff.

AUTHOR CONTRIBUTIONS

M.D.A., N.W., N.C., D.H.C., T.D., J.L.P., L.H. and M.S.H.T. conceptualized the manuscript, designed the work, acquired and analyzed the data, and drafted the manuscript. All authors (M.D.A., N.W., J.L.P., E.R., M.A., L.H., M.Y.H., M.S.H.T., D.W., T.D., B.D.W., D.H.C., N.C. and K.I.) critically reviewed the manuscript, made edits to reach the final submitted and approved version, made substantial contributions to the conception or design of the work or the acquisition, analysis or interpretation of the data, drafted the work or revised it critically for important intellectual content, gave final approval of the completed version, and are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

COMPETING INTERESTS

M.D.A. reports the following conflicts of interest: Investor, Director, Consultant, Digital Diagnostics Inc., Coralville, Iowa, USA; patents and patent applications assigned to the University of Iowa and Digital Diagnostics that are relevant to the subject matter of this manuscript; Chair, Healthcare AI Coalition, Washington DC; member, American Academy of Ophthalmology (AAO) AI Committee; member, AI Workgroup Digital Medicine Payment Advisory Group (DMPAG); member, Collaborative Community for Ophthalmic Imaging (CCOI), Washington DC; Chair, Foundational Principles of AI CCOI Workgroup. D.W. is an employee of Digital Diagnostics. The other authors declare no competing interests.

ADDITIONAL INFORMATION

Correspondence and requests for materials should be addressed to Michael D. Abramoff.

Reprints and permission information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2023