




Utility of routinely collected electronic health records data to support effectiveness evaluations in inflammatory bowel disease: a pilot study of tofacitinib

Vivek Ashok Rudrapatna ^{1,2}, Benjamin Scott Glicksberg ^{1,3},
Atul Janardhan Butte ^{1,4}

To cite: Rudrapatna VA, Glicksberg BS, Butte AJ. Utility of routinely collected electronic health records data to support effectiveness evaluations in inflammatory bowel disease: a pilot study of tofacitinib. *BMJ Health Care Inform* 2021;**28**:e100337. doi:10.1136/bmjhci-2021-100337

► Additional supplemental material is published online only. To view, please visit the journal online (<http://dx.doi.org/10.1136/bmjhci-2021-100337>).

Received 05 February 2021
Revised 07 April 2021
Accepted 26 April 2021



© Author(s) (or their employer(s)) 2021. Re-use permitted under CC BY-NC. No commercial re-use. See rights and permissions. Published by BMJ.

¹Bakar Computational Health Sciences Institute, University of California, San Francisco, San Francisco, California, USA

²Division of Gastroenterology, Department of Medicine, University of California, San Francisco, CA, USA

³Mount Sinai Medical Center, New York, New York, USA

⁴Center for Data-Driven Insights and Innovation, University of California Health System, Oakland, California, USA

Correspondence to
Dr Atul Janardhan Butte;
atul.butte@ucsf.edu

ABSTRACT

Objectives Electronic health records (EHR) are receiving growing attention from regulators, biopharmaceuticals and payors as a potential source of real-world evidence. However, their suitability for the study of diseases with complex activity measures is unclear. We sought to evaluate the use of EHR data for estimating treatment effectiveness in inflammatory bowel disease (IBD), using tofacitinib as a use case.

Methods Records from the University of California, San Francisco (6/2012 to 4/2019) were queried to identify tofacitinib-treated IBD patients. Disease activity variables at baseline and follow-up were manually abstracted according to a preregistered protocol. The proportion of patients meeting the endpoints of recent randomised trials in ulcerative colitis (UC) and Crohn's disease (CD) was assessed.

Results 86 patients initiated tofacitinib. Baseline characteristics of the real-world and trial cohorts were similar, except for universal failure of tumour necrosis factor inhibitors in the former. 54% (UC) and 62% (CD) of patients had complete capture of disease activity at baseline (month -6 to 0), while only 32% (UC) and 69% (CD) of patients had complete follow-up data (month 2 to 8). Using data imputation, we estimated the proportion achieving the trial primary endpoints as being similar to the published estimates for both UC (16%, p value=0.5) and CD (38%, p-value=0.8).

Discussion/Conclusion This pilot study reproduced trial-based estimates of tofacitinib efficacy despite its use in a different cohort but revealed substantial missingness in routinely collected data. Future work is needed to strengthen EHR data and enable real-world evidence in complex diseases like IBD.

INTRODUCTION

Real-world evidence (RWE) refers to the use of observational data to support inference on diseases and treatments. This area has been growing for a variety of reasons,¹⁻⁴ including (1) rising costs and other challenges to the feasibility of randomised trials,⁵ (2) concerns

Summary

What is already known?

- Real-world data (RWD) are receiving increasing attention from regulators, payors and biopharmaceuticals as an emerging source of evidence on treatment effects.
- Although electronic health records (EHR) data are an important and granular source of RWD, their suitability for real-world evidence remains unknown in part due to their complexity.
- Tofacitinib was recently approved for the inflammatory bowel disease (IBD-subtype ulcerative colitis), but its effectiveness and safety in real-world cohorts who may not meet trial eligibility criteria is unclear.

What does this paper add?

- Although EHR data contain much of the data needed to assess treatment effectiveness in IBD, we found these data to be less accessible (primarily found within free text) and associated with significant missing values at baseline and follow-up.
- We propose an approach for estimating real-world treatment effects from these data using data abstraction protocols and methods for stochastic imputation of missing data.
- Although a majority of the patients treated at our centre did not meet the eligibility criteria corresponding to randomised trials of tofacitinib in IBD, its effectiveness appeared to be the same as that measured in the trials.

that trial cohorts may be unrepresentative of real-world patients^{6 7} and (3) the emergence of new datasets and methods for assessing treatment in routine clinical contexts.

Of the sources of real-world data (RWD) that are being explored for this purpose, electronic health records (EHR) are receiving particular attention. They have served as the primary ledger for clinical encounters over two decades and capture rich data on

exposures and outcomes. However, this optimism has been tempered by several challenges.¹ Beyond limitations common to observational settings (eg, confounding, mismeasurement), EHR data is commonly captured in free text rather than a tabular format. This creates a challenge for the study of diseases whose assessments may be captured in narratives (eg, patient-reported outcomes). Such data typically require the use of text processing, methods that can achieve high accuracy but remain laborious. However, the utility of pursuing these approaches remains unclear because the availability of the underlying data (ie, disease activity scores) in free text is understudied.

An example of a disease currently assessed by complex measures is inflammatory bowel disease (IBD). IBD is comprised of two subtypes, ulcerative colitis (UC) and Crohn's disease (CD). Treatment involves immunosuppression that is usually continued until treatment failure (eg, inadequate efficacy, adverse events). In trials, effectiveness is measured according to the Mayo Score and Crohn's Disease Activity Index (CDAI) for UC and CD respectively.

The first small molecule approved for IBD is tofacitinib. Tofacitinib induced clinical remission in 18.5% of the 476 participants with UC who were treated for 8 weeks (OCTAVE 1) and maintained remission in 34.3% of the 197 induction responders assigned to 52 weeks of continued treatment.⁸ Tofacitinib was also evaluated in phase 2b randomised controlled trials (RCTs) of CD.⁹ In these trials, 43% of the 86 patients randomised to the 10 mg arm achieved clinical remission following induction (week 8) and 39.5% of the 60 induction responders assigned to the 5 mg arm achieved response or remission at week 26. However, unlike for UC, tofacitinib did not show statistical superiority to placebo for CD and thus was never approved for that indication. Nonetheless, it has sometimes been used off-label for CD.

In this pilot study, we assessed the utility of EHR data for treatment effectiveness evaluations in a cohort of patients with IBD treated with tofacitinib. Our primary objective was to assess disease activity data at timepoints roughly corresponding to the trial endpoints. An exploratory objective was to estimate tofacitinib's effectiveness using EHR data and compare it with the trials. Other exploratory objectives included characterising differences in patient cohorts, time-to-treatment-failure and the reasons for treatment failure.

METHODS

This retrospective cohort study of patients with IBD treated with tofacitinib was performed according to the STROBE and RECORD guidelines (online supplemental file 1).

Patient identification

We directly queried an existing database derived from all EHR records at the University of California, San Francisco

(UCSF). This previously described database¹⁰ contains records from 6/2012 (instantiation of the Epic EHR) through 4/2019 (query date) and includes diagnoses, procedures, demographics and medications. Eligible records met these criteria: (1) age over 18 years, (2) a tofacitinib order and (3) a gastroenterologist-assigned IBD diagnosis code (ICD-10-CM K50*/K51*) (table 1). Records meeting the above informatics criteria were all manually reviewed to identify a cohort of all adult patients at UCSF who had (1) ever been prescribed tofacitinib for the treatment of IBD and (2) initiated treatment.

Study endpoints

The primary endpoint was the proportion of patients with complete measurements of the Mayo Score and CDAI at baseline and follow-up. For this pilot study, baseline was defined as month -6 to 0 relative to the start date of tofacitinib, and follow-up was defined as month 2 to 8. These time-windows were chosen to reflect typical patterns of data collection in clinical practice while also allowing for rough comparisons to the timepoints assessed in trials.

An exploratory endpoint was the proportion of patients meeting the endpoints as defined by the OCTAVE trials⁸ in UC and the CD trials by Panés *et al*⁹ (see 'Comparison to trial endpoints' below). Other exploratory endpoints included characterising differences in patient cohorts, time-to-treatment-failure, and the reasons for treatment failure.

Disease activity scores

The Mayo score is scored on a 0-12 scale corresponding to the sum of four equally weighted subscores. The CDAI ranges from 0 to over 600; it incorporates three patient-reported outcomes, comorbidities, weight, haematocrit and medication use. In the gastroenterology clinic at UCSF, elements of these scores are individually captured in clinical narratives as relevant to the provision of routine care; these are not captured as structured data (eg, 'smartforms').

Data quality, completeness, and handling of missing data

We assessed the quality of the data in detail prior to proceeding with downstream analysis. We annotated missing data and characterised its distribution (figures 1 and 2). The proportion of patients with complete capture of the Mayo score and CDAI at baseline and follow-up were computed (primary endpoint). We also computed the proportion of non-missing data elements taken as a whole.

We handled missing data using a model-based approach, which relies on the data meeting the missing at random assumption. This was deemed plausible because (1) the clinical decision to pursue additional testing is typically dictated by the results of other correlated data and the risks/benefits of additional studies, and (2) we collected a wide range of auxiliary variables that inform clinical decision making (see 'Covariate abstraction').

Table 1 Baseline demographics

	OCTAVE induction 1 (n=475)	Sample of UC cohort (n=28)
Male sex, n (%)	277 (58.2)	16 (57)
Age, years	41.3±14.1	43.2±14.4
Duration of disease, years		
Median	6.5	10.2
Range	0.3–42.5	2.2–51.4
Extent of disease, n/total n (%)		
Proctosigmoiditis	64/475 (13.7)	3/28 (10.7)
Left-sided colitis	158/475 (33.3)	6/28 (21.4)
Extensive colitis/pancolitis	252/475 (53.1)	19/28 (67.9)
Total Mayo score	9.0±1.4	8.5±1.8
Partial Mayo score	6.3±1.2	6±1.6
CRP, mg/L		
Median	4.4	5.8
Range	0.1–208.4	0.8–70.6
Glucocorticoid use at baseline*	214 (45.0)	17 (60.7)
Previous treatment with TNF inhibitor, n (%)	254 (53.4)	28 (100)
Previous treatment failure, n (%)		
TNF inhibitor	243 (51.1)	28 (100)
Glucocorticoid	350 (73.5)	24 (85.7)
Immunosuppressant	360 (75.6)	21 (75)
	Panés <i>et al</i> ⁹ (n=86)	Sample of CD cohort (n=13)
Female, n (%)	47 (54.7)	9 (69.2)
Age, years		
Mean (SD)	39.3 (13.7)	39.7 (19.5)
Weight, kg		
Mean (SD)	71.6 (18.8)	69.9 (16.3)
Race, n (%)		
White	72 (83.7)	9 (69.2)
Black	2 (2.3)	0 (0)
Asian	11 (12.8)	1 (7.7)
Others	1 (1.2)	0 (0)
Duration since CD diagnosis, years		
Mean (SD)	11.3 (9.7)	14.4 (8.2)
Extent of disease, n (%)		
L1 (Ileal)	7 (8.1)	1 (7.7)
L1/4 (Ileal + Upper GI)	2 (2.3)	2 (15.4)
L2 (Colonic)	5 (5.8)	0 (0)
L2/4 (Colonic + Upper GI)	16 (18.6)	1 (7.7)

Continued

Table 1 Continued

	OCTAVE induction 1 (n=475)	Sample of UC cohort (n=28)
L3 (Ileocolonic)	15 (17.4)	4 (30.8)
L3/4 (Ileocolonic)	39 (45.3)	5 (38.5)
Prior use of TNF inhibitor, n (%)	66 (76.7)	13 (100)
Use of corticosteroids at study entry, n (%)	28 (32.6)	7 (53.8)
Baseline CDAI score		
Mean (SD)	320 (61.66)	374 (183.73)
Baseline CRP, mg/L		
Median (min-max)	5.5 (0.2–126)	28.7 (3.5–107)

Within each pair of columns, the left column corresponds to the patient demographics of the tofacitinib-assigned arm in corresponding RCTs (eg, the OCTAVE trials reported by Sandborn *et al*, the CD trials reported by Panés *et al*). The right column reports the corresponding demographics of a sample of tofacitinib-treated patients at UCSF. Treatment failure is defined as an inadequate response to any treatment (eg, steroids, TNF inhibitor) as defined and documented by the treating clinician. CDAI, Crohn's Disease Activity Index; CRP, C-reactive protein; RCTs, randomised controlled trials; TNF, tumour necrosis factor; UCSF, University of California, San Francisco.

We performed multiple imputation by chained equations using random forest models (online supplemental file 1). These methods have a lower false discovery rate than last-observation-carried-forward,¹¹ a method commonly used in IBD trials.

Covariate abstraction

Patient records were reviewed via the clinician-facing interface, which contains all clinical data, including notes, patient-provider messaging, procedure reports and laboratory results (online supplemental eTable 1). The EHR contains all clinical data generated within UCSF as well as that shared from other health systems during clinical care.

All patients were assessed by the time-to-treatment-failure, defined as either a lack of efficacy or a significant adverse event recognised by both the clinician and the patient (figure 2). This variable was distinguished from treatment non-compliance defined as a patient-initiated discontinuation against medical advice. This was separately measured during abstraction and was found to be available for all patients (online supplemental file 1). Patients who had not failed treatment at the time of data abstraction were treated as having had non-informatively censored events. Treatment discontinuation due to loss of insurance coverage as well as relocation or other lost-to-follow-up events were rare and were treated as non-informatively censored.

A random sample of the patient records in this study was selected for abstraction of the remaining variables. This

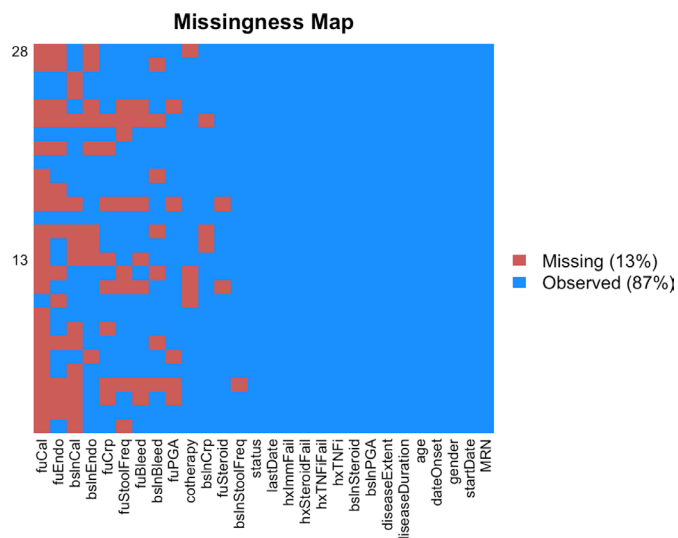


Figure 1 Distribution of missing data in the ulcerative colitis dataset. Variables are listed on the x-axis in order of decreasing missingness. Each row in the y-axis corresponds to a different patient. Variable abbreviations correspond to the following: MRN=medical record number; dateOnset=date of disease onset; diseaseDuration=length of disease; diseaseLocation=location of disease by Montreal classification; startDate=date of treatment initiation; lastDate=date of last known use of treatment; status=0 if still using tofacitinib at last date, 1 if no longer using tofacitinib at last date; cotherapy=use of other concomitant therapies (eg, mesalamine, curcumin, simple carbohydrate diet); bslnCrp=baseline C reactive protein; fuCrp=follow up C reactive protein; bslnCal=follow-up faecal calprotectin; fuCal=follow-up faecal calprotectin; bslnSteroid=baseline corticosteroid use; fuSteroid=follow-up corticosteroid use; bslnStoolFreq=baseline Mayo stool frequency subscore; fuStoolFreq=follow-up Mayo Stool Frequency subscore; bslnBleed=baseline Mayo Rectal Bleeding subscore; fuBleed=follow-up Mayo Rectal Bleeding subscore; bslnPGA=baseline Mayo Physician Global Assessment subscore; fuPGA=follow-up Mayo Physician Global Assessment subscore; bslnEndo=baseline Mayo Endoscopic subscore; fuEndo=follow-up Mayo Endoscopic subscore.

was done to strike a balance between estimating parameters with reasonable precision and the effort required for this manual review process (32 and 47 variables per record for UC and CD). The full list and definition of these variables is available in the protocol (online supplemental file 1).

CDAI elements incorporating an average daily rating over 7 days were calculated by extrapolating from a single day's mention within the time windows mentioned above. This decision was made based on accepted practices of the handling missing CDAI diary data in registrational trials (eg, UNITI-1 Statistical Analysis Plan section 5.2.1¹²) and the methods used to derive the CDAI.¹³

Comparison to trial outcomes

An exploratory endpoint of this study involved estimating the proportion of patients meeting the endpoint of the trials. As mentioned, a follow-up window of months 2–8

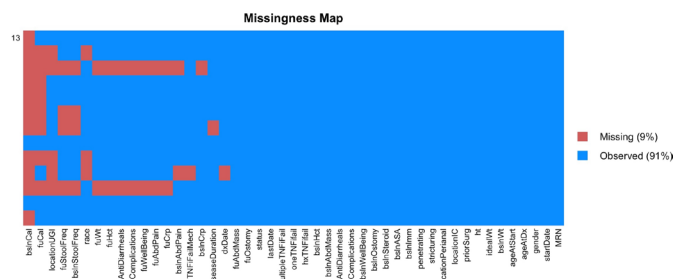


Figure 2 Distribution of missing data in the Crohn's disease dataset. Variables are listed on the x-axis in order of decreasing missingness. Each row in the y-axis corresponds to a different patient. Covariate abbreviations are as follows: 'bsln' and 'Fu' prefixes refer to variable at baseline or at follow-up; MRN=medical record number; dxDate=diagnosis date; startDate=date of treatment initiation; lastDate=date of last known use of treatment; status=0 if still using tofacitinib at last date, 1 if no longer using tofacitinib at last date; ageAtDx=age at diagnosis; ageAtStart=age at treatment start; Wt=weight; idealWt=ideal wt; ht=height; priorSurg=history of prior gastrointestinal surgery; locationLC=disease location in the lower gastrointestinal tract; locationPerianal=presence of disease in the perianal region; locationUGI=disease location in the upper gastrointestinal tract; Ostomy=presence of an ostomy; Imm=use of immunomodulators; ASA=use of aminosaliclates; steroid=use of corticosteroids; complications=complications CDAI subscore; wellbeing=wellbeing CDAI subscore; AbdPain=abdominal pain CDAI subscore; penetrating=penetrating disease behaviour; structuring=structuring disease behaviour; Hct=haematocrit; hxtNFIFail=history of TNF inhibitor failure; oneTNFiFail=history of only one prior TNF inhibitor failure; multipleTNFiFail=history of multiple TNF inhibitor failures; TNFiFailMech=classification of TNF inhibitor failure. CDAI, Crohn's Disease Activity Index; TNF, tumour necrosis factor.

after treatment was used to assess disease activity after initiating treatment. This window was chosen to account for the typical follow-up time in practice, but does not precisely match either the induction or maintenance endpoint times for either OCTAVE (weeks 8 and 52) or the corresponding CD trials⁹ (weeks 8 and 26).

Because our timepoint more closely matched that of maintenance than of induction, and because each trial only assessed remission among those achieving treatment response following induction (ie, others were assumed to be maintenance-phase non-responders), we recomputed the maintenance endpoint probability as the probability of induction patients being eligible for the maintenance trial by the probability of maintenance response among those enrolled. This probability was statistically compared with the endpoint probabilities in the UCSF-cohort.

These binary endpoints were computed using the same definitions as those in the corresponding trials. For UC, this was the proportion with a total Mayo score ≤ 2 , no individual subscore greater than 1 and a rectal bleeding subscore of 0. For CD, this corresponded to the probability that a patient had either achieved a 100-point

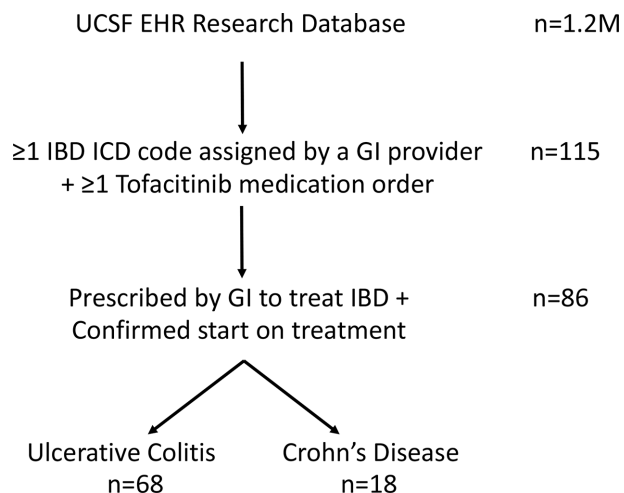


Figure 3 Cohort selection schematic. EHR, electronic health records; GI, gastrointestinal; IBD, inflammatory bowel disease; UCSF, University of California, San Francisco.

reduction in the CDAI from baseline or an absolute CDAI less than 150 at follow-up.

Statistics/computing

We computed point estimates and performed hypothesis testing using Wald test statistics with pooled standard errors.¹⁴ For analyses comparing the probability of remission in the real-world cohort with that of the RCTs, the prespecified null hypothesis was these two probabilities were equal. We estimated the time-to-treatment-failure survival distributions using the product-limit estimator. No competing events were observed. Code written in R was independently reviewed by a co-author. Data and analysis files were version-controlled using Docker.

RESULTS

Cohort identification

We identified 115 patient records following a query of our EHR database. Manual review confirmed that 86 patients—68 with UC and 18 with CD—had initiated tofacitinib specifically to treat IBD (figure 3). The other 29 patients were excluded during this process for multiple reasons, including failure to start treatment due to payor denial, the decision to forgo the ordered medical treatment in favour of surgery and treatment initiated by a non-gastroenterologist for another autoimmune condition. Non-compliance, defined as patient-initiated discontinuation of tofacitinib against medical advice, was rare (4%) in this cohort.

Data completeness

Out of 28 patients with UC randomly sampled for full assessment of the Mayo score and all other auxiliary variables at baseline and follow-up, 15 (54%) had a complete capture of the Mayo score at baseline and 9 (32%) had a complete capture at follow-up. The least available subscore was endoscopy (figure 1). With respect to the partial Mayo score, 21 (75%) and 17 (61%) were complete

at these timepoints. Out of 13 patients with CD sampled, 8 (62%) had complete capture of the CDAI at baseline and 9 (69%) had this available at follow-up.

Taken as a proportion of the total number of collected variables, 13% of the UC-related data and 9% of the CD-related were missing (figures 1 and 2). These missing data were handled by multiple imputation.

Cohort characterization

The baseline demographics of the subjects under study in the UCSF and RCT cohorts were similar (table 1). Notable differences include the universal failure of TNF inhibitors in the UCSF cohort, as well as a longer duration of disease in the patients with UC. Patient groups had similar baseline Mayo scores, C-reactive protein levels and prevalent corticosteroid use. Sixty-one per cent of the cohort had been using corticosteroids at baseline. Thirty-nine per cent of the cohort used at least one form of additional IBD treatment: these included mesalamine, curcumin and dietary changes.

Zero per cent of the patients with UC initiated on tofacitinib met the eligibility criteria of the corresponding phase 3 RCT.⁸ The reasons for this were multifactorial (online supplemental eTable 2) but include use of vedolizumab within the previous year, high-dose steroids at the time of treatment initiation and the possibility of requiring surgery during the treatment period.

We separately explored what proportion of patients met the specific RCT entry criteria defined by the Mayo score and CDAI for UC and CD, respectively. Ninety-three per cent (73–98) of the patients with UC had an eligible baseline Mayo score (6–12), whereas 50% (19–82) of the patients with CD had a baseline CDAI within the eligibility range of the corresponding RCT (220–450).

Effectiveness and safety

Time-to-treatment-failure analysis on the full cohort revealed similar survival distributions irrespective of IBD disease subtype (online supplemental eFigure 1). The overall probability of incident users continuing tofacitinib long-term was 68% (58%–80%). All failure events occurred within the first 7 months; among continued responders by month 6, the probability of sustained absence of treatment failure was 94%. Of note, the first use of the tofacitinib occurred in 2013, and the longest duration of effectiveness data relevant to treatment maintenance was 3.7 years.

We assessed the reasons for treatment failure (online supplemental eFigure 2). In the UC cohort, there were 17 treatment failure events: 12 with insufficient treatment efficacy, 4 with adverse events/intolerances and 1 due to patient preference. Of the 12 efficacy failures, 8 patients (67%) contained evidence of ongoing inflammation on the basis of biomarkers, imaging or lower endoscopy performed within the 2-month period prior to treatment failure. All patients who did not undergo objective confirmation of inflammation during this timeframe did have objective evidence of inflammation prior to treatment

Table 2 Potential approaches to strengthen routinely collected electronic health records data and better support real-world evidence studies

Problem	Example	Potential solutions
Complex and cumbersome disease activity scores limit practical use	The CDAI incorporates a comprehensive list of elements but only some apply to any given patient (eg, abdominal pain predominant, fistula predominant). Elements that are not relevant to a given patient are unlikely to be captured during routine clinical care	Develop and validate novel scores that accurately capture disease activity, are easy to administer and capture in real-world contexts and are relevant to different patient subgroups
Data capture by free text rather than structured data capture (eg, EHR smartforms)	<ul style="list-style-type: none"> ▶ Typing out clinical narratives is faster and more natural to clinicians than point-and-click interfaces ▶ These narratives are relatively inaccessible for RWE studies (requires natural language processing) and are subject to ambiguous documentation ▶ Unclear if current, documentation-oriented reimbursement schemes are compatible with smartform-entered data 	<ul style="list-style-type: none"> ▶ More streamlined and relevant scores as above ▶ Partnership between clinical, quality, operations, IT, user experience and research teams to optimise data capture ▶ Payors may be able to incentivise better data capture in support of outcome-based and risk-adjusted reimbursement
Patient-oriented and decision-oriented data capture rather than cohort-oriented data capture	<p>Patient 1 has a colonoscopy showing severe endoscopic disease. A precise characterisation and documentation of current patient symptoms is irrelevant to treatment decision making.</p> <p>Patient 2 has worsening symptoms and a rise in biochemical markers consistent with prior flares. The decision is made to change treatment without additional testing (eg, enterography, colonoscopy)</p> <p>Patients 1 and 2 individually have sufficient data to support personalised decision making, but collectively have inadequate data to support cross-cutting RWE studies of treatment outcomes</p>	<ul style="list-style-type: none"> ▶ Partnership between clinical, quality, operations, IT, user experience and research teams to optimise data capture ▶ Payors may be able to incentivise a shift towards and improved measurement of healthcare quality and disease-oriented population health management
Encounters are not well-timed relative to important clinical events	<p>Week -5: Patient is seen in clinic and agrees to switch therapy. Symptoms and disease activity captured in the EHR.</p> <p>Week 0: Patient fills prescription and begins treatment as an outpatient.</p> <p>Week 7: Patient returns for follow-up.</p> <p>Results: (1) No symptom capture at the time of treatment initiation, (2) Week 7 follow-up might not align with data capture of other patients</p>	<ul style="list-style-type: none"> ▶ Clinic-level harmonisation of practices concerning the timing of patient encounters and follow-up ▶ Use of interactive remote technologies (mobile apps, chatbots) to generically increase the frequency of data capture or time data capture ▶ Payors incentivise patients to participate in disease tracking (lower premiums), potentially in collaboration with pharmacies or infusion centres (optimally timed capture)
Encounter presence/absence correlated with clinical outcomes	<p>Patient 1 is feeling well 8 weeks after starting tofacitinib and is on a high-deductible plan. She does not want to take time off from work to go to clinic or pay the copay when she has no current clinical needs.</p> <p>Patient 2 is not feeling well 8 weeks after starting tofacitinib. He stops taking the medication and does not follow-up because he does not think the clinicians can help him.</p>	<ul style="list-style-type: none"> ▶ Use of interactive remote technologies (mobile apps, chatbots) to increase touchpoints with the clinic, improve trust and avoid the time and monetary expenses of a clinic visit ▶ Supplementing EHR data and supporting the function of the clinic with staff-initiated outreach (eg, an EHR/RWD-augmented registry)

CDAI, Crohn's Disease Activity Index; EHR, electronic health records; RWE, real-world evidence.

initiation. All but one patient with inadequate response completed a minimum of 7 weeks of treatment induction (11 weeks on average) prior to the adjudication of treatment failure.

In the CD cohort, there were five treatment failure events: one due to an adverse event (zoster) and one due to insufficient efficacy (all with concomitant objective evidence of ongoing inflammation). These patients completed 13.2 weeks of treatment on average.

Twenty-two per cent of all subjects participating in the induction phase of the UC RCT⁸ met the primary maintenance endpoint of week 52 clinical remission. We observed a similar response (16%) in the corresponding UCSF cohort (6%–37%, *p* value=0.5). Similarly, the proportion achieving the primary endpoint in the CD RCT⁹ (34%) was similar to the point estimate of the real-world cohort (38%, *p*-value=0.8).

We explored the extent to which steroid use may account for some of these results. In the UC cohort, 33% of patients had been using steroids at the time of follow-up. Among the patients who had been using steroids at baseline, 56% were steroid-free at the time of follow-up.

DISCUSSION

We assessed the completeness of routinely collected EHR data to support RWE studies of diseases with complex activity measures. Taking a use case of tofacitinib as used to treat IBD (both on-label and off-label), we found that the capture of the total Mayo score and the CDAI is currently modest at best, even at a tertiary-care medical centre.

On exploratory analyses, the real-world effectiveness of this drug appeared to be consistent with its published effectiveness from randomised trials despite its use in a substantially different cohort. We found that patients with IBD using tofacitinib appear to generally tolerate it well and that unlike biologics commonly used for IBD, secondary loss of response events for this small molecule was uncommon.

RWD has been receiving growing interest from a variety of parties including the FDA³ and EMA,¹⁵ biopharmaceuticals and payors. Despite this interest, it must be recognised that not all RWD are created equal. Unlike prospectively planned disease and treatment registries, the EHR data capture mechanism has historically been designed with other objectives in mind: healthcare coordination and delivery, revenue generation and medico-legal documentation among others.

Our pilot study highlights the substantial work that will be needed to close the quality gap between retrospective EHR data and prospective data and realise the promise of RWE. We outline the root causes of this quality gap as well as outline potential solutions in [table 2](#). Many of these solutions will ultimately require a close partnership between the many stakeholders in real-world clinical care: clinicians, patients, health IT, operations and especially

payors. Undoubtedly, this may require a significant investment in both time and money by these participants. However, we are of the opinion that the eventual rewards are worth the investment. These include the ability to better measure the quality of care, discover practice-changing evidence and enable continuous-improving learning health systems.

Strengths of this study include the use of a preregistered protocol and analysis plan, the use of rigorous methods for handling missing data,¹¹ as well as openly available code accompanied by deidentified raw EHR data in order to maximise the reproducibility and reusability of this work. The primary limitation of this work lies in its inability to draw inferences related to the real-world effectiveness of tofacitinib.

CONCLUSION

Routinely collected EHR data currently has uneven capture of the data needed to optimally assess IBD treatment effectiveness at baseline and follow-up. This work provides several insights into real-world practice, including typical patterns of data collection and the real-world effectiveness and safety of tofacitinib for IBD. It also offers an analytical approach to the analysis of missing real-world data. Future efforts are needed to improve inference from these data, such better data capture mechanisms and novel measures more suitable to routine care.

Twitter Vivek Ashok Rudrapatna @vivicality, Benjamin Scott Glicksberg @BenGlicksberg and Atul Janardhan Butte @atulbutte

Acknowledgements The authors thank the UCSF Academic Research Services and Clinical Data Research Consultation services for clinical informatics support. The authors would like to acknowledge Dana Ludwig for his help in deidentifying and interpreting the UCSF EHR.

Contributors VAR and AB conceived the project. VAR designed the chart review protocol, performed chart extraction, conducted statistical analyses and drafted this manuscript. BG performed code review and critically edited this manuscript. AB supervised the project and critically edited this manuscript.

Funding Research reported in this publication was supported by funding from the UCSF Bakar Computational Health Sciences Institute and the National Center for Advancing Translational Sciences of the National Institutes of Health under award UL1 TR001872. VAR was supported by the National Institute of Diabetes and Digestive and Kidney Disease of the National Institutes of Health grant under award T32 DK007007.

Disclaimer The contents are solely the responsibility of the authors and do not necessarily represent the official views of the NIH.

Competing interests Atul Butte is a co-founder and consultant to Personalis and NuMedii; consultant to Samsung, Mango Tree Corporation, and in the recent past, 10x Genomics, Helix, Pathway Genomics, and Verinata (Illumina); has served on paid advisory panels or boards for Geisinger Health, Regenstrief Institute, Gerson Lehman Group, AlphaSights, Covance, Novartis, Genentech, Merck, and Roche; is a shareholder in Personalis and NuMedii; is a minor shareholder in Apple, Facebook, Alphabet (Google), Microsoft, Amazon, Snap, Snowflake, 10x Genomics, Illumina, Nuna Health, Assay Depot (Scientist.com), Vet24seven, Regeneron, Sanofi, Royalty Pharma, Pfizer, BioNTech, AstraZeneca, Moderna, Biogen, Twist Bioscience, Pacific Biosciences, Editas Medicine, Invitae, and Sutro, and several other non-health related companies and mutual funds; and has received honoraria and travel reimbursement for invited talks from Johnson and Johnson, Roche, Genentech, Pfizer, Merck, Lilly, Takeda, Varian, Mars, Siemens, Optum, Abbott, Celgene, AstraZeneca, AbbVie, Westat, several investment and venture capital firms, and many academic institutions, medical or disease specific foundations and associations, and health systems. Atul Butte receives royalty payments through

Stanford University, for several patents and other disclosures licensed to NuMedii and Personalis. Atul Butte's research has been funded by NIH, Northrup Grumman (as the prime on an NIH contract), Genentech, Johnson and Johnson, FDA, Robert Wood Johnson Foundation, Leon Lowenstein Foundation, Intervallien Foundation, Priscilla Chan and Mark Zuckerberg, the Barbara and Gerson Bakar Foundation, and in the recent past, the March of Dimes, Juvenile Diabetes Research Foundation, California Governor's Office of Planning and Research, California Institute for Regenerative Medicine, L'Oreal, and Progenity.

Patient consent for publication Not required.

Ethics approval We obtained Institutional Review Board approval to obtain patient data and abstract all covariates.

Provenance and peer review Commissioned; externally peer reviewed.

Data availability statement Data are available in a public, open access repository. The analytic code in the form of a R markdown file as well as the accompanying data set needed to reproduce the analysis in this work are available in a Docker container to all investigators without restriction (<https://doi.org/10.7272/Q6PZ5715>). These individual participant data were de-identified to comply with the US Department of Health and Human Services 'Safe Harbor' guidance and applicable laws and regulations concerning privacy and/or security of personal information. The data dictionary is documented within the study protocol section of Supplemental Content.

Supplemental material This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been peer-reviewed. Any opinions or recommendations discussed are solely those of the author(s) and are not endorsed by BMJ. BMJ disclaims all liability and responsibility arising from any reliance placed on the content. Where the content includes any translated material, BMJ does not warrant the accuracy and reliability of the translations (including but not limited to local regulations, clinical guidelines, terminology, drug names and drug dosages), and is not responsible for any error and/or omissions arising from translation and adaptation or otherwise.

Open access This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: <http://creativecommons.org/licenses/by-nc/4.0/>.

ORCID iDs

Vivek Ashok Rudrapatna <http://orcid.org/0000-0003-1789-3004>

Benjamin Scott Glicksberg <http://orcid.org/0000-0003-4515-8090>

Atul Janardhan Butte <http://orcid.org/0000-0002-7433-2740>

REFERENCES

- 1 Rudrapatna VA, Butte AJ. Opportunities and challenges in using real-world data for health care. *J Clin Invest* 2020;130:565–74.
- 2 Chatterjee A, Chilukuri S, Fleming E. Real-World evidence: driving a new drug development paradigm in oncology, 2018. Available: https://www.mckinsey.com/~/media/mckinsey/industries/pharmaceuticals_and_medical_products/our_insights/real_world_evidence_driving_a_new_drug_development_paradigm_in_oncology/real-world-evidence-driving-a-new-drug-development-p [Accessed 2 Jul 2019].
- 3 Framework for FDA's Real-World Evidence Program, 2018. Available: www.fda.gov [Accessed 26 Jun 2019].
- 4 Medicines Agency E. An agency of the European Union Regulatory Perspective on Real World Evidence (RWE) in scientific advice EMA Human Scientific Committees' Working Parties with Patients' and Consumers' Organisations (PCWP) and Healthcare Professionals' Organisations (HCPWP).
- 5 Moore TJ, Zhang H, Anderson G, et al. Estimated costs of pivotal trials for novel therapeutic agents Approved by the US food and drug administration, 2015-2016. *JAMA Intern Med* 2018;178:1451–7.
- 6 Ha C, Ullman TA, Siegel CA, et al. Patients enrolled in randomized controlled trials do not represent the inflammatory bowel disease patient population. *Clin Gastroenterol Hepatol* 2012;10:1002–7.
- 7 Zoccali C, Blankestijn PJ, Bruchfeld A, et al. Children of a lesser God: exclusion of chronic kidney disease patients from clinical trials. *Nephrol Dial Transplant* 2019;34:1112–4.
- 8 Sandborn WJ, Su C, Sands BE, et al. Tofacitinib as induction and maintenance therapy for ulcerative colitis. *N Engl J Med* 2017;376:1723–36.
- 9 Panés J, Sandborn WJ, Schreiber S, et al. Tofacitinib for induction and maintenance therapy of Crohn's disease: results of two phase IIb randomised placebo-controlled trials. *Gut* 2017;66:1049–59.
- 10 Norgeot B, Glicksberg BS, Trupin L, et al. Assessment of a deep learning model based on electronic health record data to forecast clinical outcomes in patients with rheumatoid arthritis. *JAMA Netw Open* 2019;2:e190606.
- 11 Buuren S van. flexible imputation of missing data.
- 12 Feagan BG, Sandborn WJ, Gasink C, et al. Ustekinumab as induction and maintenance therapy for Crohn's disease. *N Engl J Med* 2016;375:1946–60.
- 13 Best WR, Beckett JM, Singleton JW, et al. Development of a Crohn's disease activity index. National cooperative Crohn's disease study. *Gastroenterology* 1976;70:439–44.
- 14 Rubin DB, Wiley J, York N. Multiple imputation for nonresponse in surveys, 1987. Available: <https://www.onlinelibrary.wiley.com/doi/pdf> [Accessed 20 Jun 2019].
- 15 Cave A, Cerreta F. Use of real world data in development programmes, 2017. Available: https://www.ema.europa.eu/en/documents/presentation/presentation-use-real-world-data-development-programmes-dr-alison-cave-dr-francesca-cerreta_en.pdf [Accessed 26 Jun 2019].

© 2021 Author(s) (or their employer(s)) 2021. Re-use permitted under CC BY-NC. No commercial re-use. See rights and permissions. Published by BMJ. This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: <http://creativecommons.org/licenses/by-nc/4.0/>. Notwithstanding the ProQuest Terms and Conditions, you may use this content in accordance with the terms of the License.

UK learning about digital health and COVID-19

Helene Feger,¹ Bernard Crump,¹ Philip Scott ²

To cite: Feger H, Crump B, Scott P. UK learning about digital health and COVID-19. *BMJ Health Care Inform* 2021;**28**:e100376. doi:10.1136/bmjhci-2021-100376

Received 01 April 2021
Accepted 10 May 2021

In a few short weeks, the COVID-19 pandemic radically changed the way health and care services are delivered. The rapid acceleration of digital transformation has been one of the most dramatic changes. Professionals and members of the public have welcomed the benefits that virtual working has brought about, though serious reservations have been expressed by some people and implementation challenges remain a key concern. A report by the Royal College of General Practitioners¹ shows changes to clinical practice, with nearly three-quarters of patients consulting their general practitioner remotely via computer or phone during the pandemic compared with nearly three-quarters attending in person in 2019.

Digital is here to stay, but it is vital that we learn lessons from the experience of front-line clinicians, care professionals and patients to address the challenges and opportunities that transformation presents. The Professional Record Standards Body (PRSB)² is a UK-wide member organisation set up by the Department of Health and Social Care to set standards for the information shared in health and care records. PRSB asked nearly 100 members³ and partners including the royal colleges, social care system leaders, health and care providers, patient groups, regulators and others for their views and experience of digital health and care during the pandemic—both positive and negative. The themes that emerged were discussed with representatives from 64 of the stakeholder organisations with whom the PRSB works.

In February 2021, a round table discussion was arranged by PRSB with some 20 senior leaders from the National Health Service (NHS), social care, regulators, royal colleges and other professional and patient bodies reinforced the importance and relevance of the findings to the future delivery of care. Participants agreed more coordinated action is needed to understand the safety, regulatory and workforce implications

of digital transformation brought about by the pandemic as well as the impact on the accessibility of care for people using services. The findings and recommendations centred around the following themes:

BUILDING ON THE MOMENTUM FOR CHANGE

Harnessing the enthusiasm of patients and professionals for digital transformation should be seized and practical solutions should be adopted, rather than seeking technical perfection. However, challenges must be addressed. The PRSB calls on professional bodies and patient groups to consider a targeted review of the safety implications of remote consultation, including assessing the impact on clinical risk management and continued patient access to face-to-face consultations. It should identify and address gaps in existing guidance (eg, a policy on providing recordings of consultations to patients); address access issues for the digitally excluded and any potential liabilities arising from the shift to virtual consultations and sharing recordings.

INTEGRATED CARE

The pandemic has highlighted the pressing need to integrate health and social care, particularly in England. In collaboration with NHS Digital's Social Care Programme⁴ and 16 local pathfinders, PRSB has developed standards which have the potential to have a major impact on those using social care services. An implementation plan for the new standards is needed in order to build support for digital transformation in areas where it is most needed and to realise the promise of integrated care.

SELF-MANAGEMENT AND REMOTE MONITORING

PRSB members foresee a step change in digital remote monitoring and self-care tools including apps and other digital health



© Author(s) (or their employer(s)) 2021. Re-use permitted under CC BY-NC. No commercial re-use. See rights and permissions. Published by BMJ.

¹Professional Record Standards Body, London, UK

²Centre for Healthcare Modelling & Informatics, University of Portsmouth Faculty of Technology, Portsmouth, UK

Correspondence to

Helene Feger;
Helene.Feger@theprsb.org



technologies. But the tools deployed need to command public trust as well as professional confidence. The current regulation of digital health technologies is not sufficiently robust or responsive and needs to be strengthened urgently by the NHS, working with professional bodies, patient groups and medical device regulators to ensure their safe use.

REDUCING THE BURDEN OF DATA COLLECTIONS

Clinicians and care professionals welcomed suspension of some national data collections for secondary uses during the early phase of the pandemic, when the Data Coordination Board was put on hold. System leaders have seized the initiative to improve the number, quality and timeliness of data collections so they are fit for purpose, avoid duplication and reduce burden on over-stretched clinicians. PRSB and its member professional bodies are working with NHSX on its plans to streamline data collections and ensure alignment between data collected for direct care and other uses such as commissioning and research.

RESETTING SERVICES

The pandemic has also highlighted the need to redesign the front door to urgent and emergency services in order to avoid overcrowding, reduce infection risks and improve safety for patients and staff. NHS 111,⁵ England's rapid medical advice line for non-emergencies, is a key feature of the new model of care but better information based on standardised flows from 111 that interoperate with acute and emergency services as well as primary and community care are needed. This will need to be underpinned by new standards for information exchange and NHSX, NHS Digital, PRSB and others have begun this work.

SHARED DECISION-MAKING AND END-OF-LIFE CARE

The pressures to clear waiting lists and prioritise patients while dealing with COVID-19 highlight the need for a consistent approach to shared decision-making. The pandemic may change the view of the balance of risks for some patients and they need to be enabled to engage in

these decisions. PRSB believes that a national standard for shared decision-making with a meaningful implementation programme is an urgent priority for the NHS. Equally, PRSB supports NHS England's Palliative and End-of-Life Care programme⁶ in its efforts to align different approaches to the recording of end-of-life care wishes with the Electronic Palliative Care Coordination Systems. Given the impact of the pandemic on end-of-life care, this work is urgently needed and digital solutions should give certainty about the provenance and curation of end-of-life information.

The speed and nature of digital transformation arising from the pandemic is a testament to everyone working in health and social care. However, given the concerns raised, issues around the safety and effectiveness of implementation must now be addressed if we are to maximise the benefits while minimising the risk that digital technologies to patients and health and care workers.

Contributors HF outlined the article with contributions from BC and PS. HF wrote the first draft, which was revised by BC and PS.

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests None declared.

Patient consent for publication Not required.

Provenance and peer review Commissioned; externally peer reviewed.

Open access This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: <http://creativecommons.org/licenses/by-nc/4.0/>.

ORCID iD



Philip Scott <http://orcid.org/0000-0002-6289-4260>

REFERENCES

- 1 RCGP. General practice in the post Covid world, 2020. Available: <https://www.rcgp.org.uk/-/media/Files/News/2020/general-practice-post-covid-rcgp.ashx?la=en> [Accessed Feb 2021].
- 2 PRSB. Available: <https://theprsb.org/> [Accessed Feb 2021].
- 3 PRSB. Our members. Available: <https://theprsb.org/aboutus/membership/> [Accessed Feb 2021].
- 4 NHS. Social care programme. Available: <https://digital.nhs.uk/services/social-care-programme> [Accessed Feb 2021].
- 5 NHS. Get medical help. Available: <https://111.nhs.uk> [Accessed Feb 2021].
- 6 NHS England. Available: <https://www.england.nhs.uk/eolc/> [Accessed Feb 2021].

© 2021 Author(s) (or their employer(s)) 2021. Re-use permitted under CC BY-NC. No commercial re-use. See rights and permissions. Published by BMJ. This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: <http://creativecommons.org/licenses/by-nc/4.0/>. Notwithstanding the ProQuest Terms and Conditions, you may use this content in accordance with the terms of the License.

Smartphone-based remote monitoring of vision in macular disease enables early detection of worsening pathology and need for intravitreal therapy

Meriam Islam,¹ Stafford Sansome,¹ Radha Das,¹ Marko Lukic,¹ Kelvin Yi Chong Teo,^{2,3} Gavin Tan,² Konstantinos Balaskas,¹ Peter B M Thomas ,¹ Lucas M Bachmann ,⁴ Andrew M Schimel,⁵ Dawn A Sim¹

To cite: Islam M, Sansome S, Das R, *et al.* Smartphone-based remote monitoring of vision in macular disease enables early detection of worsening pathology and need for intravitreal therapy. *BMJ Health Care Inform* 2021;**28**:e100310. doi:10.1136/bmjhci-2020-100310

Received 16 December 2020
Revised 07 March 2021
Accepted 13 April 2021



© Author(s) (or their employer(s)) 2021. Re-use permitted under CC BY-NC. No commercial re-use. See rights and permissions. Published by BMJ.

¹NiHR Biomedical Research Centre for Ophthalmology, Moorfields Eye Hospital NHS Foundation Trust and UCL Institute of Ophthalmology, Moorfields Eye Hospital NHS Foundation Trust, London, UK

²Department of Ophthalmology, Singapore National Eye Centre, Singapore

³Department of Ophthalmology, NUS Medical School, Singapore

⁴Department of Clinical Epidemiology, University of Zurich, Zurich, Switzerland

⁵Department of Ophthalmology, Centre for Excellence in Eye Care, Miami, Florida, USA

Correspondence to

Dawn A Sim; dawnsim@nhs.net

ABSTRACT

Background/aims To assess the outcomes of home monitoring of distortion caused by macular diseases using a smartphone-based application (app), and to examine them with hospital-based assessments of visual acuity (VA), optical coherence tomography-derived central macular thickness (CMT) and the requirement of intravitreal injection therapy.

Design Observational study with retrospective analysis of data.

Methods Participants were trained in the correct use of the app (Alleye, Oculocare, Zurich, Switzerland) in person or by using video and telephone consultations. Automated threshold-based alerts were communicated based on a traffic light system. A 'threshold alarm' was defined as three consecutive 'red' scores, and turned into a 'persistent alarm' if present for greater than a 7-day period. Changes of VA and CMT, and the requirement for intravitreal therapy after an alarm were examined.

Results 245 patients performing a total of 11 592 tests (mean 46.9 tests per user) were included and 85 eyes (164 alarms) examined. Mean drop in VA from baseline was -4.23 letters (95% CI: -6.24 to -2.22 ; $p < 0.001$) and mean increase in CMT was $29.5 \mu\text{m}$ (95% CI: -0.08 to 59.13 ; $p = 0.051$). Sixty-six eyes (78.5%) producing alarms either had a drop in VA, increase in CMT or both and 60.0% received an injection. Eyes with persistent alarms had a greater loss of VA, -4.79 letters (95% CI: -6.73 to -2.85 ; $p < 0.001$) or greater increase in CMT, $+87.8 \mu\text{m}$ (95% CI: 5.2 to 170.4 ; $p = 0.038$).

Conclusion Smartphone-based self-tests for macular disease may serve as reliable indicators for the worsening of pathology and the need for treatment.

INTRODUCTION

In the ophthalmic realm, self-testing of vision enhances patient health empowerment with the added benefit of enabling the efficient use of hospital resources and improving access to treatment.

In recent years, several digital home vision tests have become available to patients with

Summary

What is already known?

- ▶ Avoidable harm may ensue from delays to ophthalmic care.
- ▶ Self-testing and the remote monitoring of vision is gaining traction.
- ▶ The Alleye app is accurate in detecting age-related macular degeneration.

What does this paper add?

- ▶ The development of an Alleye alarm and their frequency of generation in patients with all causes of macula pathology provide a mechanism of alerting clinicians to worsening disease.
- ▶ These alarms should prompt clinicians to consider expediting an in person review for patients with such pathology.

macula disease.¹ Self-testing was first established in patients with age-related macular degeneration (AMD) using preferential hyperacuity perimetry on a standalone device (ForeseeHome, Notal Vision, United States), with a randomised trial demonstrating earlier detection of disease progression and lower reduction in visual acuity (VA) compared with standard care.² Currently, there are two smartphone-based vision tests available for the remote monitoring of metamorphopsia in patients with macular pathology that are Food and Drug Administration 510(k) cleared and CE (European Conformity)—marked Class I device approved; myVisiontrack and Alleye (Oculocare, Switzerland). myVisiontrack uses a shape discrimination task examining 3° of the central visual field and the Alleye test a dot-alignment task covering 12° of the central visual field.^{3–5}

In October 2019, the medical retina team at Moorfields Eye Hospital in South London

first implemented the use of the Alleye app in patients receiving intravitreal therapy for macular disease who were undergoing extension of their treatment interval. Following the arrival of the global pandemic and the associated national lockdown in the UK from 23 March 2020, the challenge of providing a retinal therapy service in the COVID-19 era was created. At Moorfields Eye Hospital, 'forward triage' was used.⁶ Patients triaged to postponed appointments were invited to participate in remote monitoring of their vision.

In this paper, we studied the relationship between the frequency of Alleye alarms and the central macular thickness (CMT) and VA at subsequent follow-up, in addition to the need for intravitreal injection therapy following the easing of the lockdown restrictions.

MATERIALS AND METHODS

Patient enrolment

This study looked at patients recruited at two time periods: prior to the lockdown, from October 2019 to February 2020 and after the commencement of the lockdown in March 2020. Prior to the lockdown, patients with diabetic macular disease attending the medical retina injection clinics (in a face to face environment) were offered the option to use the Alleye application for home-monitoring of their vision and if they agreed, trained to use the app in person. Patients were advised that if an alarm was triggered, they would be contacted by the clinic team by telephone and given the option of bringing their appointment forward (shortening interval between clinic visits). Patients who did not own or have access to a smartphone were loaned a new iPod Touch (6th generation, Apple, Cupertino, California, USA) with the Alleye app pre-installed for their use. Each patient was given an anonymised identifier allowing their scores to be monitored remotely, and this remained underway until the commencement of the first UK lockdown in March 2020.

Home monitoring protocol

Following the Lockdown, patients who were triaged to 'medium' or 'low' risk and had their appointments deferred received a telephone call from the clinical team and were offered the opportunity for self-testing of their vision.⁶ Those who were willing to start the home monitoring were offered telephone or video training to use the app. All patients were asked to contact the clinical team by telephone if there were concerns, and were provided with an email to contact the clinical team directly with queries.

Patients' scores were monitored weekly. A 'threshold alarm' was defined as three consecutive 'red' scores, defined as a 'persistent alarm' if present for a >7-day period. Patients who met this criterion of a persistent alarm were called by telephone by a clinician and asked if their vision was worse/better/unchanged. If worse, they were asked if they felt they needed an injection or that

their vision was worse. If they answered in the affirmative, their appointment was expedited, within 1–2 weeks.

If once called the patient felt that their vision was unchanged or better, then they were asked to continue testing their vision using the app, and would be called again the following week if their scores remained in the 'red' zone. This was also the case for patients who did not answer the telephone, until they were seen face to face in the clinic. Any patients who discontinued testing following a threshold alarm being generated were also contacted by telephone.

Clinical characteristics

Baseline VA and the optical coherence tomography (OCT)-derived CMT in addition to medical retina diagnosis and laterality were recorded from the electronic medical record. Demographics were recorded in the way of age, gender and ethnicity. The first follow-up visit provided with a VA and OCT scan post threshold alarm was recorded for each patient that has had a face to face review since threshold alarm generation.

Statistical analysis

We summarised continuous variables with means and SD and dichotomous variables with percentages. We calculated changes of VA (Early Treatment of Diabetic Retinopathy Study letters) and CMT from baseline and tested differences statistically using one-sample t-tests, considering a p value <0.05 statistically significant. We plotted the occurrence of worsening of VA and CMT against the need for an injection. To assess selection patterns of enrolment before and during the lockdown, we compared clinical characteristics at baseline statistically. Analyses were performed using the Stata V.16.1 statistics software package (StataCorp)

RESULTS

During the months of April and May 2020, 605 patients triaged as 'low' and 'medium' risk were contacted by telephone. Of these, 222 patients met the minimum VA criteria of 6/24 to participate and wished to be involved, being booked for subsequent appointments of training in the use of the application at a later date. Of these patients, 90 were onboarded via video consultation using the Attend Anywhere platform (Attend Anywhere, Victoria, Australia), and 60 were onboarded via telephone. The remainder of patients involved were recruited via technicians in virtual clinics and clinicians reviewing patients in a face to face environment, who were then contacted for a telephone training appointment at a later time. Details of the selection process are shown in the flowchart of [figure 1](#).

A total of 212 patients tested with Alleye over the first UK lockdown and performed 9938 test results, in addition to 33 patients with diabetic macular oedema, who had been using Alleye prior to the lockdown, generating 1654 test results. The baseline clinical characteristics did

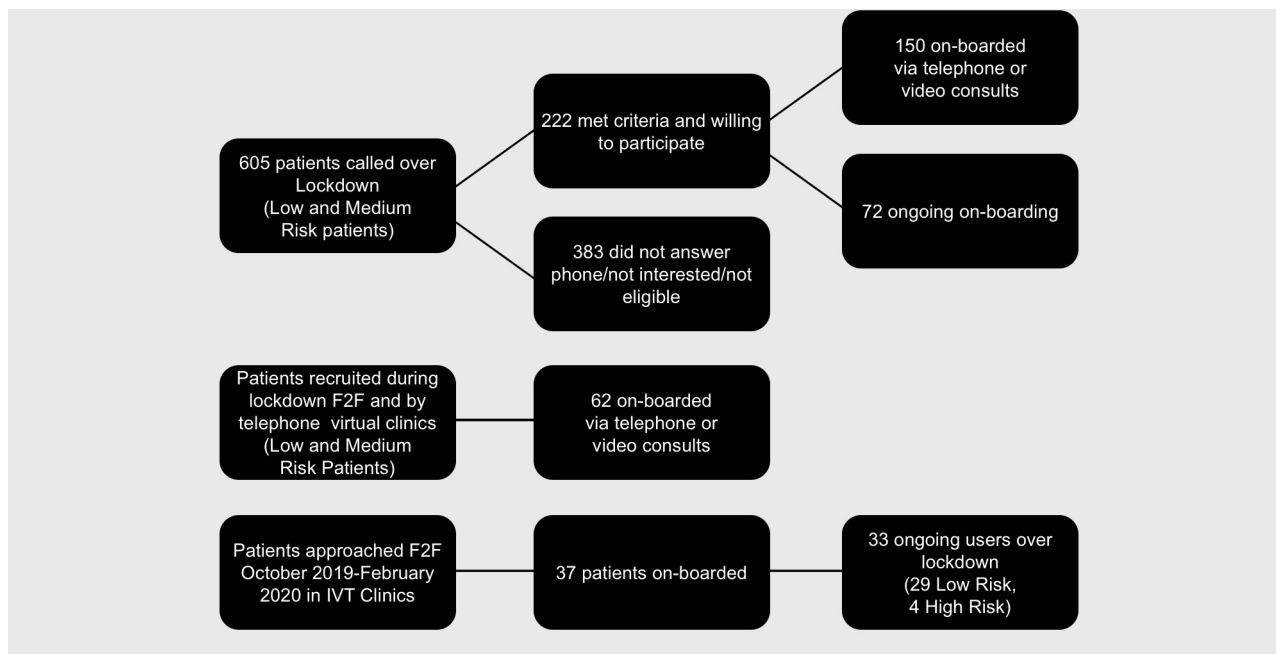


Figure 1 Flowchart outlining patient recruitment for home monitoring of vision using the Alleye app. Patients were recruited either by telephone calls to those stratified as ‘low’ and ‘medium’ risk during the COVID-19 pandemic or in person when attending virtual clinics and face to face (F2F) clinic appointments. Following recruitment, patients were subsequently onboarded with the smartphone application via telephone or video consultation.

not differ between the two cohorts that were onboarded before and during the lockdown. The average number of tests performed per user was 46.9 tests. A threshold alarm was defined as three consecutive Alleye scores that are red on separate days, therefore signifying potential deterioration in the score. A persistent alarm was defined as a threshold alarm that continues to remain for 7 days. Over the period from 23 March 2020 to 11 August 2020, 98 eyes of 65 patients each developed a minimum of one ‘Alleye alarm’.

The total number of threshold alarms produced was 164. Of the 98 eyes producing alarms, follow-up data of 85 eyes were available. Of these 164, 30 persistent alarms were generated. The mean age of patients with alarms was 65.1 years (SD=11.1; range 41–90 years) and 56.2% were female. On clinical review, 28 eyes were considered worse, followed by 38 considered as stable, 12 eyes improved. Classification of seven eyes was missing. The summary of patients’ characteristics is available in [table 1](#).

The mean change in VA from baseline to follow-up post threshold alarm generation was -4.23 letters (95% CI: -6.24 to -2.22 ; $p < 0.001$). The mean change in CMT was $+29.5 \mu\text{m}$ (95% CI: -0.08 to 59.13 ; $p = 0.051$). The mean number of alarms generated per patient was 1.67 (range 1–9). Based on the clinical assessment and the patients’ willingness to be treated, 51 eyes (60.0%) received an intravitreal injection, 29 eyes received no injection, 3 eyes had a contra-indication for treatment and 2 eyes were not treated because the patients declined injection. Sixty-six eyes (78.5%) producing alarms either had a drop in VA, increase in CMT or both ([figure 2](#)).

The 30 persistent alarms were generated by 24 eyes (22 patients). VA dropped by -4.79 letters (95% CI: -6.73 to -2.85 ; $p < 0.001$) and CMT increased by $+87.8 \mu\text{m}$ (95% CI: 5.2 to 170.4 ; $p = 0.038$) in this group of patients. Injections were provided in 20/24 eyes (83.3%). Compared with the group with threshold alarms, the likelihood for an injection was significantly higher ($p = 0.034$). The patient flow outcomes for these patients are demonstrated in [figure 3](#).

Eyes considered to have worsened since the last clinical review had a higher probability of receiving an injection (89.3%) than stable (42.1%) or improved (66.7%) eyes. Similarly, eyes showing a worsening in the follow-up visit had a higher drop in VA (-8.4 letters (SD 8.8)) than stable (-2.3 letters (SD 9.6)) or improved (-3.0 (SD 8.4)) eyes. In addition, CMT increases were highest in those eyes classified as worse ($+116.9 \mu\text{m}$ (SD 144.8)), while stable ($-17.4 \mu\text{m}$ (SD 51.2)) and improved ($-18.1 \mu\text{m}$ (SD 206.3)) eyes had a slightly lower CMT than before the lockdown.

DISCUSSION

Main findings

Due to the exceptional circumstances during the lockdown, we were able to study the consequences of delayed treatment and the relationship between changes of the Alleye signal and clinical parameters in the absence of any treatment. We demonstrated a close relationship between the frequency of the Alleye alarms and the subsequent deterioration of macular disease coupled with the need for intravitreal injection therapy. Four out of five patients

Table 1 Patient characteristics of eyes triggering ‘threshold alarms’ (n=98)

Characteristics	Mean	SD
Patients' age	65.1	11.1
Female gender (%)	56.2	
VA (ETDRS letters) baseline	75.8	9.7
Central macular thickness (CMT) (µm) baseline	264.9	85.3
Mean changes in VA (ETDRS letters) from baseline	-4.2	9.3
Mean changes in CMT (µm) from baseline	29.5	134.7
Ethnicity	n (eyes)	
English/Welsh/Scottish/Northern Irish/British	27	
Mixed/multiple ethnic groups	2	
Any other white background	6	
Indian	15	
Pakistani	12	
Any other Asian background	1	
African	4	
Caribbean	9	
Any other Black/African/Caribbean background	2	
Not stated	20	
Diagnoses	n (eyes)	
DMO	57	
RVO	14	
3-AMD	9	
Diabetic maculopathy	5	
Diabetic retinopathy no maculopathy	7	
Sickle cell retinopathy	1	
Atherosclerosis	2	
Myopic CNV	3	
Frequency of alarms	n (eyes)	
1	70	
2	14	
3	5	
4	3	
5	3	
7	1	
8	1	
9	1	

AMD, age-related macular degeneration; CMT, central macular thickness; CNV, choroidal neovascularisation; DMO, diabetic macular oedema; ETDRS, Early Treatment of Diabetic Retinopathy Study; RVO, retinal vein occlusion; VA, visual acuity.

presenting with alarms showed signs of clinical progression and two out of three patients required immediate intravitreal therapy. [Figure 3](#) flowchart demonstrates the outcomes of those patients generating persistent alarms.

Results in context of the existing literature

Macular disease can affect individuals of all ages and is particularly debilitating as it affects central vision. Nearly 1.5 million people in the UK have macular disease.⁷ Self-monitoring of vision for macular disease is an area gaining significant traction, fuelled by clinical needs, service requirements and patient interest.¹ The WHO has identified ‘self-care’ as a key health topic and defines it as the ability of individuals, families and communities to promote health, prevent disease, maintain health and cope with illness and disability with or without the support of a health worker.⁸ The WHO further advocates evidence-based digital technologies and mHealth approaches which can be accessed fully or partially outside of formal health services. Within the UK, there is an increasing backlog of routine appointments as a direct consequence of the COVID-19 pandemic.⁹ This is due to the postponement of clinical reviews following clinical triage resulting from the need to reduce face to face contact during the peak of the pandemic.¹⁰ The literature so far discussed the immediate challenges for the management of glaucoma.^{11 12} However, even prior to this, ophthalmology as a specialty was well on its way to buckling under the strain applied by our National Health Service.¹³ It has been known for several years that ophthalmology is the busiest outpatient specialty in the UK, with an estimated 30%–40% of increased demand over the next two decades.¹⁴ A combination of workforce shortages, clinic space and increasing healthcare requirements for an ageing population has resulted in health-service initiated delays, demonstrated by the RCOphth and British Ophthalmological Surveillance Unit to be resulting in permanent and severe visual loss for our patients.¹⁵

It is therefore of benefit to patients and doctors to avoid unnecessary in person hospital appointments that have direct impacts on waiting lists for others. It is also worth sparing a thought for the stable patient, attending hospital for a clinical review that is not required. This is a waste of resources for both the patient and hospital, and a source of potential unnecessary anxiety generated for an individual, which must not be underestimated. Conversely, there will be many patients in the community waiting for their first hospital appointment with long delays, with potential avoidable harm incurred.¹⁶ Prior to this study, it had been demonstrated that Alleye was highly accurate in detecting wet AMD.^{4 5} This is the first study of its kind observing all patients with known macula pathology.

Strengths and limitations

This study was conducted using a pragmatic approach, which we recognise as both a strength using a real-world approach, and as a potential limitation. Due to the lockdown, the onboarding of patients using video and telephone consultations was necessary. Therefore, it cannot be entirely ensured that the information and instructions provided to the patients was always understood to the desired degree. In addition, regular patient contact was not feasible and an inherent assumption was made

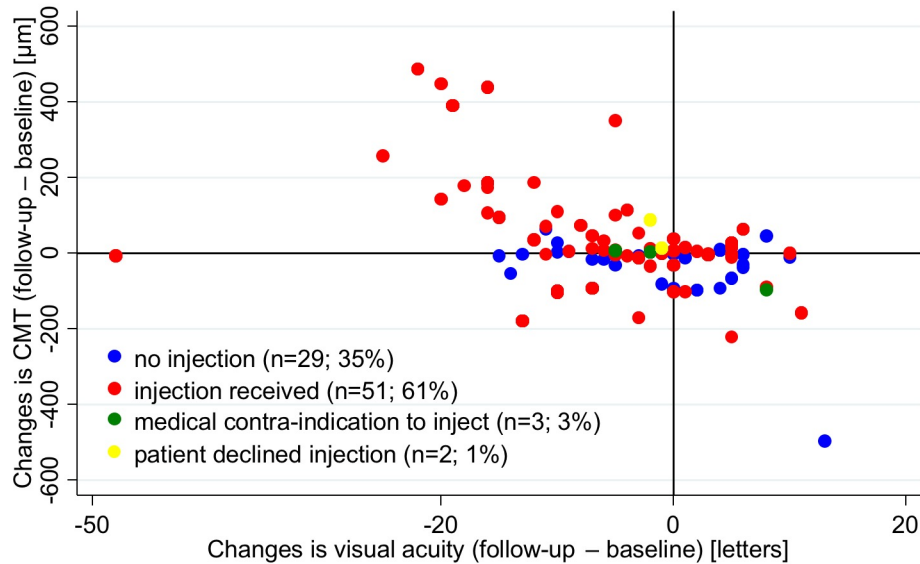


Figure 2 Distribution of changes of visual acuity and central macular thickness (CMT) for patients at the next follow-up visit after triggering ‘threshold alarms’ (n=84) and whether intravitreal injections were received in these patients or not.

that patients would adhere to our testing recommendations. Finally, due to the backlog of patients requiring hospital visits following the easing of the lockdown restrictions, a few of the patients with Alleye alarms had not yet returned for a clinical follow-up before writing up this paper, precluding the collection of the results of all threshold alarms. Using video and telephone consultations, we maintained what we felt was an appropriate level of contact and were able to successfully triage patients requiring a hospital visit. Inevitably, some delays may have occurred due to natural concerns from patients who were reluctant to attend hospital due to the risk of catching COVID-19, and these patients may have returned for later reviews to the hospital despite Alleye threshold alarms, however, no patients voiced a deterioration of vision at the same time as a direct reluctance to attend. If advised, all

patients attended for earlier review. Delays in returning to the hospital may have contributed to disease progression resulting in a positive correlation between occurrence or frequency of Alleye alarms and the likelihood of the evidence of clinical signs of progression in the follow-up visit.

Implications for research

These results provide promise for the future directions of research for remote monitoring of vision in the context of macular disease.^{6 17–19} Future studies should focus on the false negative rate of Alleye, that is, those patients returning at follow-up post lockdown whose clinical parameters have deteriorated, but did not trigger a threshold Alleye alarm. Moreover, our observation that patients with persistent alarms had a higher likelihood

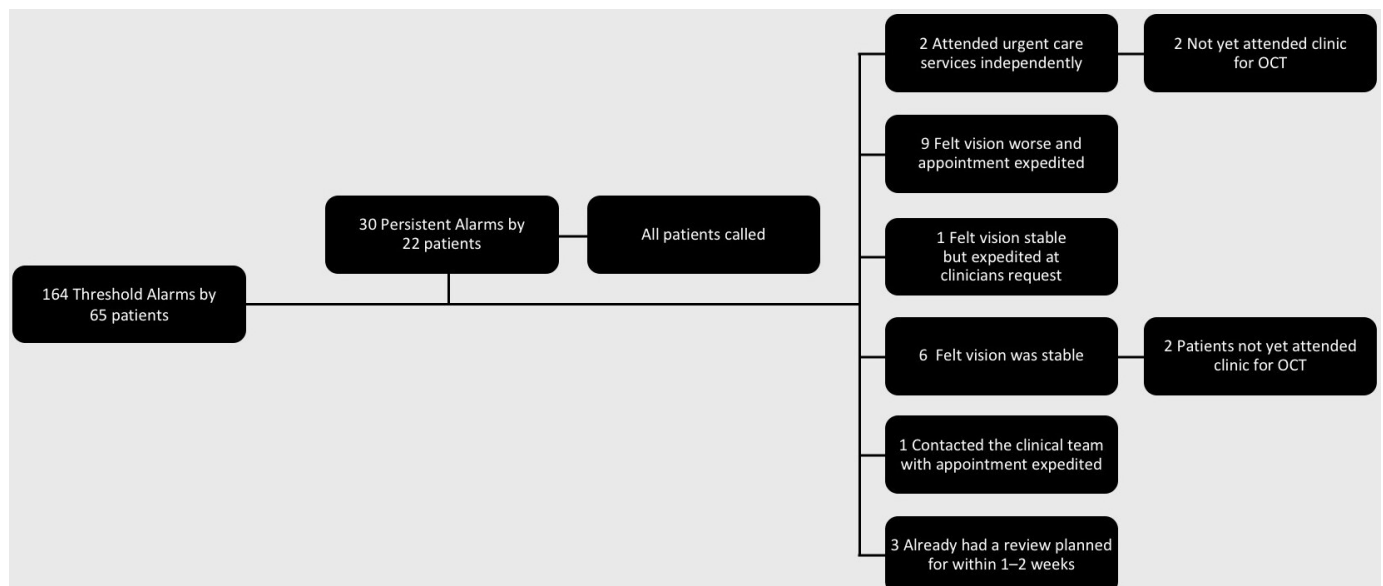


Figure 3 Flow of patient of outcomes after triggering a ‘persistent alarms’. OCT, optical coherence tomography.

for progression and subsequently required intravitreal injections could be further explored. Our data suggest that changing the policy of alarm management could have an impact on the specificity of the alarm signal. If confirmed, healthcare providers could easily adapt the management based on Alleye alarms according to their local requirements and service capacities. While the sensitivity of Alleye in the screening setting has been assessed before,^{4,5} sensitivity and positive predictive values in the monitoring situation still need confirmation. Another stream of research should assess the impact of home monitoring in conjunction with telemedicine to increase the efficiency of care delivery and the monetary impact of home monitoring on health service cost. While there is some indication from other clinical fields that remote management of chronic diseases is cost-beneficial,^{20,21} robust evidence in the field of ophthalmology is limited. Finally, with view to the drug pipeline extended release applications, the value of home monitoring to support patient management between reviews should be assessed, as the prolongation of intervals between face to face visits bears the risk of missing deterioration.

Implications for practice

Telemedicine in ophthalmology has been steadily gaining traction over the last few years. Increasingly, time is far more binary, categorised into the 'Pre-COVID-19' and the 'Post COVID-19' era.¹⁰ People often talk of a 'new normal'. It is important to recognise that an effective, efficient and safe telemedicine programme cannot be implemented overnight. It is imperative that for technological solutions to have longevity that they harness systems and procedures that are already in place.²² For self-testing and home monitoring to be incorporated into routine clinical practice, it requires a network of ophthalmic technicians, nurses and clinicians to support its use. If incorporated into routine care, particularly in an era where face to face review is the exception and not the rule, we could move towards our ultimate aim of ensuring the right patient is seen at the right time. We are still at the precipice of decentralised care for retinal diseases. In addition to the monitoring of functional limitations by Alleye, morphological examination methods using mobile OCT devices will soon be available.²³ This provides exciting potential for the exploration of the interaction between function and morphology to strengthen the informative value of home measurements.

CONCLUSION

Our work illustrates that Alleye alarms and their frequency act as reliable indicators for worsening of macula pathology, and should prompt clinicians to consider expediting such patients' face to face clinical reviews. We are at the beginning of a new era of medical care, in which valid clinical data from patients staying at home will be available in real time and can be included in the decision-making process of the need for on-site clinical

visits. The ecosystem of digital developments is only just beginning. We envision a future in which the interaction of integrated digital solutions will have a major impact on the way healthcare is organised for chronic eye diseases.

Twitter Peter B M Thomas @pbmthomas, Lucas M Bachmann @medignition and Dawn A Sim @dawnasim

Contributors DAS, MI and LMB provided substantial contributions to the conception and design of the work. MI, DAS, SS, RD, ML conducted the acquisition of the data. DAS, MI and LMB were involved in the drafting the initial manuscript and revising it critically for important intellectual content.

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests LMB is a founding member of Oculocare Medical, which develops innovative products in eye care, such as the self-monitoring test described in this paper.

Patient consent for publication Not required.

Ethics approval The research described adhered to the tenets of the Declaration of Helsinki and this study was registered under the Digital Clinical Laboratory and Audit department of Moorfields Eye Hospital NHS Foundation Trust as part of a service improvement project. All patients included as part of routine clinical care gave their informed consent by accepting the user agreement within the app during sign up, which allowed the use of their anonymised data for this analysis. Data for this analysis were obtained from a review of the electronic medical records from these patients.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement All data relevant to the study are included in the article or uploaded as supplemental information.

Open access This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: <http://creativecommons.org/licenses/by-nc/4.0/>.

ORCID iDs

Peter B M Thomas <http://orcid.org/0000-0003-1681-7659>

Lucas M Bachmann <http://orcid.org/0000-0002-9868-154X>

REFERENCES

- 1 Faes L, Bachmann LM, Sim DA. Home monitoring as a useful extension of modern tele-ophthalmology. *Eye* 2020;34:1950-3.
- 2 Chew EY, Clemons TE, Bressler SB, et al. Randomized trial of the ForeseeHome monitoring device for early detection of neovascular age-related macular degeneration. The HHome Monitoring of the Eye (HOME) study design - HOME Study report number 1. *Contemp Clin Trials* 2014;37:294-300.
- 3 Kaiser PK, Wang Y-Z, He Y-G, et al. Feasibility of a novel remote daily monitoring system for age-related macular degeneration using mobile handheld devices: results of a pilot study. *Retina* 2013;33:1863-70.
- 4 Schmid MK, Faes L, Bachmann LM, et al. Accuracy of a self-monitoring test for identification and monitoring of age-related macular degeneration: a diagnostic case-control study. *Open Ophthalmol J* 2018;12:19-28.
- 5 Schmid MK, Thiel MA, Lienhard K, et al. Reliability and diagnostic performance of a novel mobile app for hyperacuity self-monitoring in patients with age-related macular degeneration. *Eye* 2019;33:1584-9.
- 6 Hollander JE, Carr BG. Virtually perfect? Telemedicine for Covid-19. *N Engl J Med* 2020;382:1679-81.
- 7 Types of macular disease, 2015. Available: <https://www.macularsociety.org/types-macular-conditions> [Accessed 16 Nov 2020].
- 8 Self-care interventions for health, 2020. Available: https://www.who.int/health-topics/self-care#tab=tab_1 [Accessed 16 Nov 2020].
- 9 Mehrotra A, Chernew M, Linetsky D. The rebound in visits has occurred across all specialties. The Commonwealth Fund, 2020.

- Available: <https://www.commonwealthfund.org/chart/2020/rebound-visits-has-occurred-across-all-specialties>
- 10 Group HEiRPW. Direct and indirect impacts of COVID-19 on health and wellbeing, 2020. Available: <https://www.ljmu.ac.uk/~media/phi-reports/2020-07-direct-and-indirect-impacts-of-covid19-on-health-and-wellbeingpdf> [Accessed 18 Nov 2020].
 - 11 Che Hamzah J, Daka Q, Azuara-Blanco A. Home monitoring for glaucoma. *Eye* 2020;34:155–60.
 - 12 Jayaram H, Strouthidis NG, Gazzard G. The COVID-19 pandemic will redefine the future delivery of glaucoma care. *Eye* 2020;34:1203–5.
 - 13 Digital NHS. Hospital outpatient activity, 2017. Available: <https://digital.nhs.uk/data-and-information/publications/statistical/hospital-outpatient-activity/2017-18> [Accessed 11 Oct 2020].
 - 14 The Royal College of Ophthalmologists Report. The way forward, 2017. Available: <https://www.rcophth.ac.uk/standards-publications-research/the-way-forward/> [Accessed 22 Oct 2020].
 - 15 Foot B, MacEwen C. Surveillance of sight loss due to delay in ophthalmic treatment or review: frequency, cause and outcome. *Eye* 2017;31:771–5.
 - 16 Ting DSJ, Krause S, Said DG, *et al.* Psychosocial impact of COVID-19 pandemic lockdown on people living with eye diseases in the UK. *Eye* 2020. doi:10.1038/s41433-020-01130-4. [Epub ahead of print: 10 Aug 2020].
 - 17 Ko MW, Busis NA. Tele-Neuro-Ophthalmology: vision for 20/20 and beyond. *J Neuroophthalmol* 2020;40:378–84.
 - 18 Koh AHC, Koh LRS, Sheu S-J, *et al.* What COVID-19 has taught us: lessons from around the globe. *Graefes Arch Clin Exp Ophthalmol* 2020;258:2091–4.
 - 19 Temesgen ZM, DeSimone DC, Mahmood M, *et al.* Health care after the COVID-19 pandemic and the influence of telemedicine. *Mayo Clin Proc* 2020;95:S66–8.
 - 20 Hummel JP, Leipold RJ, Amorosi SL, *et al.* Outcomes and costs of remote patient monitoring among patients with implanted cardiac defibrillators: an economic model based on the predict RM database. *J Cardiovasc Electrophysiol* 2019;30:1066–77.
 - 21 Monahan M, Jowett S, Nickless A, *et al.* Cost-Effectiveness of Telemonitoring and self-monitoring of blood pressure for antihypertensive titration in primary care (TASMINH4). *Hypertension* 2019;73:1231–9.
 - 22 Faes L, Rosenblatt A, Schwartz R, *et al.* Overcoming barriers of retinal care delivery during a pandemic-attitudes and drivers for the implementation of digital health: a global expert survey. *Br J Ophthalmol* 2020. doi:10.1136/bjophthalmol-2020-316882. [Epub ahead of print: 16 Oct 2020].
 - 23 Maloca P, Hasler PW, Barthelmes D, *et al.* Safety and feasibility of a novel sparse optical coherence tomography device for patient-delivered retina home monitoring. *Transl Vis Sci Technol* 2018;7:8.

© 2021 Author(s) (or their employer(s)) 2021. Re-use permitted under CC BY-NC. No commercial re-use. See rights and permissions. Published by BMJ. This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: <http://creativecommons.org/licenses/by-nc/4.0/>. Notwithstanding the ProQuest Terms and Conditions, you may use this content in accordance with the terms of the License.

Digital biomarkers for the prediction of mental health in aviation personnel

Laura Müller,¹ Diederik De Rooy ^{1,2}

To cite: Müller L, De Rooy D. Digital biomarkers for the prediction of mental health in aviation personnel. *BMJ Health Care Inform* 2021;**28**:e100335. doi:10.1136/bmjhci-2021-100335

Received 05 February 2021
Accepted 26 April 2021

Researchers have proposed to use information from digital sources such as smartphones and wearable technology to objectify patient mental health characteristics. Using big data analysis methods, patterns can be detected. This is called digital phenotyping. In this communication, we will discuss the use of digital phenotyping for professionals in aviation. We choose this very specific area of medicine because there have been several aviation crashes in the last years that were due to a mental problem of a pilot.¹ The mental health of flight crews remains one of the biggest challenges for improving aviation safety. Digital biomarkers may be highly promising here, while at the same time easily misused, with enormous consequences. Of course, most of our findings may also be applicable outside this area.

Digital phenotyping can include smartphone sensors, keyboard interaction and voice and speech features, but it can also go as far as social media posts, internet searches and Bluetooth recognition of other mobile devices.² The value of digital phenotyping is that it offers a multidimensional and measurable method to objectively gather patient data. Whereas a clinical interview by a psychiatrist only provides static information in an artificial setting, smartphone data are collected throughout the day in the patient's daily setting.³ Furthermore, clinical interviews are dependent on the therapist's interpretations, in contrast to smartphone data which are unbiased and quick to assemble.⁴

The most important challenges of digital phenotyping and ways to overcome these are summarised in [table 1](#). An important risk is that the privacy of the patient may be violated or that the amount of assembled information might be perceived as being intrusive. As a possible solution, it has been proposed to collect only 'content-free' data, such as human-computer interaction.³ This would mean, for example, that the manner in which someone types is analysed without gathering

the content of what is being typed, with the purpose that no personal information can be extracted from it. Still, combinations of data and context information can provide valuable personal knowledge. Therefore, the data use should be strictly regulated.

In our opinion, digital phenotyping is most promising for monitoring those with already identified mental conditions, on a voluntary basis. It should not be used for random screening for mental disorders. This would be a too large infringement of privacy. Especially when people are under pressure by their employer, they might be persuaded to agree to giving access to their information, while not actually consenting to the privacy risks. Also, in an unselected population, the risk of false-positive results increases, which means that someone is wrongly identified as being at risk for or having a mental disorder. On the contrary, digital phenotyping might be useful to monitor pilots who are being treated or have recovered from mental health problems and who consciously agree to the use of digital phenotyping. Then, it might help clinicians to better predict recovery or early relapse, and could perhaps even shorten the period a pilot is being grounded for mental health problems. It should not be used to replace clinical examinations, but only to provide additional information to clinicians, to improve the quality of clinical examinations. When combined with normal clinical care, the risk of false-positive and false-negative results diminishes, as well as the risk of 'gaming', meaning that patients will do things only to let the algorithm show better results.

A systematic review about the use of digital phenotyping for patients with affective disorders shows that there were 27 feasibility studies investigating digital biomarkers. The studies reported the association between mood status and phone usage (9 studies), physical activity (8), location (8), voice features (8), light exposure (3) and heart-rate



© Author(s) (or their employer(s)) 2021. Re-use permitted under CC BY-NC. No commercial re-use. See rights and permissions. Published by BMJ.

¹Psychiatry, Leiden University Medical Center, Leiden, The Netherlands

²Transparant Mental Healthcare, Leiden, The Netherlands

Correspondence to
Dr Diederik De Rooy;
derooy@outlook.com

Table 1 Using digital phenotyping for mental health in aviation professionals

Challenges regarding digital phenotyping	Possible solutions
Reliability: how well do digital biomarkers associate with mental health?	<ul style="list-style-type: none"> ▶ More RCTs comparing digital phenotyping with clinician's prediction ▶ Comparison of different parameters ▶ More research into developing algorithms ▶ Investigating the use of machine learning ▶ Research in healthy individuals
Clinical utility: does it help to improve symptoms and clinical evaluation?	<ul style="list-style-type: none"> ▶ Testing benefits of quick detection of onset/relapse ▶ Research into role of monitoring in treatment or during follow-up ▶ Defining clinical outcomes based on symptoms in future studies
Privacy: how much personal information will be gathered?	<ul style="list-style-type: none"> ▶ Data are regarded as medical data to which medical confidentiality laws are applicable ▶ Protection of data by dedicated regulation ▶ Restricted amount of and 'content-free' data modalities ▶ Retractable informed consent
Regulation: who is accountable for proper use and protection of data?	<ul style="list-style-type: none"> ▶ Only approved apps: guidance for clinicians ▶ Only use by healthcare professionals ▶ Healthcare professional is responsible for choosing reliable commercial apps
Application: what should it be used for?	<ul style="list-style-type: none"> ▶ Monitoring, but not screening ▶ Not mandatory

RCT, randomised controlled trial.

variability (2). Twenty studies (also) included subjective self-assessments. The quality of the studies was limited, as many did not have a control group, used small sample sizes and had a short follow-up period. The data analysis that was applied differed regarding the number of analysed parameters and what algorithm was used and consequently, efficacy results were inconsistent.⁵ Only one randomised controlled trial was reported in the systematic review. This study showed that in bipolar patients an intervention consisting of daily self-assessments did not yield significant change in either depressive or manic symptoms, compared with the control group.⁶

Therefore, digital phenotyping is an interesting innovation, which developments should be watched closely, but critically, as there should be considerably more research into the association between digital measurements and mental health as well as the clinical utility. It has potential to be beneficial in mental disorders in aerospace medicine, but there are important technical and ethical challenges, regarding effectiveness, privacy and regulation.

Contributors Both authors designed the study, analysed the articles and wrote and approved the final version of the manuscript.

Competing interests None declared.

Patient consent for publication Not required.

Provenance and peer review Not commissioned; externally peer reviewed.

Open access This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: <http://creativecommons.org/licenses/by-nc/4.0/>.

ORCID iD

Diederik De Rooy <http://orcid.org/0000-0001-8866-8416>

REFERENCES

- Mulder S, De Rooy DPC. Improving pilot mental health: negative life-events, peer-support and just culture. *Aerosp Med Hum Perform* 2018;1:41–51.
- Bush NE, Armstrong CM, Hoyt TV. Smartphone apps for psychological health: a brief state of the science review. *Psychol Serv* 2019;16:188–95.
- Insel TR. Digital phenotyping: a global tool for psychiatry. *World Psychiatry* 2018;17:276–7.
- Naarding P, Marijnissen RM, Westerhof GJ. Digitale psychiatrie [Digital psychiatry]. *Tijdschr Psychiatr* 2019;61:335–42.
- Dogan E, Sander C, Wagner X, et al. Smartphone-based monitoring of objective and subjective data in affective disorders: where are we and where are we going? Systematic review. *J Med Internet Res* 2017;19:e262.
- Faurholt-Jepsen M, Frost M, Ritz C, et al. Daily electronic self-monitoring in bipolar disorder using smartphones – the MONARCA I trial: a randomized, placebo-controlled, single-blind, parallel group trial. *Psychol Med* 2015;45:2691–704.

© 2021 Author(s) (or their employer(s)) 2021. Re-use permitted under CC BY-NC. No commercial re-use. See rights and permissions. Published by BMJ. This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: <http://creativecommons.org/licenses/by-nc/4.0/>. Notwithstanding the ProQuest Terms and Conditions, you may use this content in accordance with the terms of the License.

Development and validation of a machine learning model to predict mortality risk in patients with COVID-19

Anna Stachel ¹, Kwesi Daniel,¹ Dan Ding,¹ Fritz Francois,² Michael Phillips,³ Jennifer Lighter⁴

To cite: Stachel A, Daniel K, Ding D, *et al.* Development and validation of a machine learning model to predict mortality risk in patients with COVID-19.

BMJ Health Care Inform 2021;**28**:e100235. doi:10.1136/bmjhci-2020-100235

► Additional supplemental material is published online only. To view, please visit the journal online (<http://dx.doi.org/10.1136/bmjhci-2020-100235>).

Received 03 September 2020
Revised 18 December 2020
Accepted 13 January 2021



© Author(s) (or their employer(s)) 2021. Re-use permitted under CC BY-NC. No commercial re-use. See rights and permissions. Published by BMJ.

¹Department of Infection Prevention and Control, NYU Langone Health, New York, NY, USA

²Department of Medicine, NYU Grossman School of Medicine, New York, NY, USA

³Department of Medicine, Division of Infectious Diseases, NYU Grossman School of Medicine, New York, NY, USA

⁴Department of Pediatrics, Division of Pediatric Infectious Diseases, NYU Grossman School of Medicine, New York, NY, USA

Correspondence to

Dr Anna Stachel;
anna.stachel@nyulangone.org

ABSTRACT

New York City quickly became an epicentre of the COVID-19 pandemic. An ability to triage patients was needed due to a sudden and massive increase in patients during the COVID-19 pandemic as healthcare providers incurred an exponential increase in workload, which created a strain on the staff and limited resources. Further, methods to better understand and characterise the predictors of morbidity and mortality was needed.

Methods We developed a prediction model to predict patients at risk for mortality using only laboratory, vital and demographic information readily available in the electronic health record on more than 3395 hospital admissions with COVID-19. Multiple methods were applied, and final model was selected based on performance. A variable importance algorithm was used for interpretability, and understanding of performance and predictors was applied to the best model. We built a model with an area under the receiver operating characteristic curve of 83–97 to identify predictors and patients with high risk of mortality due to COVID-19. Oximetry, respirations, blood urea nitrogen, lymphocyte per cent, calcium, troponin and neutrophil percentage were important features, and key ranges were identified that contributed to a 50% increase in patients' mortality prediction score. With an increasing negative predictive value starting 0.90 after the second day of admission suggests we might be able to more confidently identify likely survivors

Discussion This study serves as a use case of a machine learning methods with visualisations to aide clinicians with a better understanding of the model and predictors of mortality.

Conclusion As we continue to understand COVID-19, computer assisted algorithms might be able to improve the care of patients.

BACKGROUND

New York City quickly became an epicentre of the COVID-19 pandemic in the USA.¹ As of 28 April, we identified 7352 cases across our three major medical campuses, of which 3995 were admitted. Due to a sudden and massive increase in patients during COVID-19 pandemic, healthcare providers

incurred an exponential increase in workload that created a strain on the staff and limited resources. While mortality prediction models have been developed in patients with septic shock, heart failure and in the intensive care unit, literature does not show a model tailored for patients with COVID-19 in the USA.^{2–4} As COVID-19 is not well characterised, we developed a prediction model using machine learning techniques to identify predictors and patients with high risk of mortality. A prediction model can be used to risk adjust hospitals and unit care, incorporated into an AI notification tool and used in additional studies where a mortality risk score is needed.^{5–7} Hospitals can develop straightforward models with high accuracy to identify predictors that characterise a disease in their patient population.

This study adds another prediction model methodology to the literature using primarily objective data readily available electronic health record (EHR) information to classify COVID-19 patient's risk of mortality. This study aims to: (1) develop models to predict daily risk of mortality in hospitalised patients by applying modern machine learning techniques using discrete information found in the EHR and (2) understand and visualise predictors associated with mortality in patients with COVID-19 using variable importance techniques. This study also provides an example how hospitals can leverage their own EHR data to build customised prediction models.

METHODS

Design

Adhering to 'Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis' (TRIPOD) model

evaluation, this retrospective cohort study mined structured patient data from the EHR at NYU Langone Health and applied machine learning methods to predict the risk of mortality in patients admitted to the hospital with COVID-19.⁸ NYU Langone Health is an academic medical centre located in New York City with over 2000 licenced beds during COVID-19. This study includes patients from three of the medical/surgical campuses comprising of approximately 1700 beds. Data for this study are derived from our Enterprise Data Warehouse—data aggregated from clinical and EPIC clarity tables. The outcome was death during admission in inpatients with COVID-19 confirmed by PCR within prior 60 days of visit. Three datasets with different samplings of the patient population were used to develop three separate models, and a final performance evaluation of the three models was conducted on daily patient mortality predictions to determine the most versatile model (table 1). All three final 'models included patient demographic information at admission in addition to the following information: (1) laboratory and vital results on the first calendar day of admission—'admission'; (2) last available laboratory and vital results during the admission—'last-value' and (3) laboratory and vital results selected on a random day during the admission—'time-vary'.

The training cohorts for all the three datasets included adult inpatients with admitted and subsequently discharged either alive or dead during 3 March 2020–28 April 2020 (n=3395). Patients not admitted or under 18 years of age were excluded. We used the time-holdout method and split hospital admissions into a training dataset (3 March–12 April: n=2054) and an internal validation dataset (13–16 April, n=477) for internal validation including model tuning and model selection. For the external test set (17–28 April, n=864), we used future subsequent discharges to test (estimating accuracy of the selected, fully-specified models) and monitor the performance of the models over time. Dividing the data temporally (rather than randomly via cross-validation) for external validation better simulates more realistic results as models trained from historical data will perform similarly in the risk stratification of future patients.

Feature engineering

A full cohort dataset comprising of 971 patient-level and admission-level features were derived from 83 variables from the EHR (table 1). Features that rely on human decisions such as treatment, or interpretation and documentation such as symptoms and image reviews were excluded to limit the introduction of bias in the model. The cumulative mean, median, min and max of all patients results were calculated along with iterations of the absolute differences among these results at the end of each day until discharge. This engineering allows for laboratory and vital results to be put in the context of the patient rather than the population. For example, low blood pressure might be normal for one patient but unusual for

another, and the change in these results during a hospital admission might be indicative of disease progression.

Continuous variables were categorised/binning into five groups based on median cutoffs (<first quintile, second quintile, third quintile, fourth quintile and ≥fifth quintile). Variables with missing information were grouped into a sixth bin.

Binning was performed in the training dataset, and those thresholds were applied in the validation and test datasets. The number of features was reduced to decrease computational memory and avoid overfitting of the training model. Features that appeared in less than 20% in the training dataset were excluded.

Machine learning algorithms and tuning parameters

We applied machine learning algorithms to predict mortality on the constructed features. The following commonly used algorithms in healthcare research were used to create prediction models and assessed for performance: logistic regression (LR), decision tree (DT), gradient boosting decision trees (GB), support vector machine (SVM) and neural network (NN).⁹ To deal with overfitting in model selection, algorithms were tuned with the internal validation set using default and associated hyperparameters listed in supplementary material (online supplemental table 1).

Missing data

For LR, SVM and NN, missing values were imputed on datasets using median values from observations found in the training set in order to avoid dropping incomplete cases and improve model training. For binary or categorical variables, the median was rounded to the nearest integer. For DT and GB, missing values were treated as separate values and used in the calculation of the worth of a splitting rule. This consequently produces a splitting rule that assigns the missing values to the branch that maximises the worth of the split. This can be a desirable option as existence of a missing values such as lab test can be predictive of mortality.

Model performance

We used the area under the receiver operating characteristic curve (AUC), as well as accuracy, sensitivity, specificity, positive predictive value and negative predictive value using a prediction estimate threshold of 50% to evaluate the ability to discriminate survivors from non-survivors. Each algorithm on the three sampled datasets (admission, last-value, time-vary) and their associated validation and test sets were applied. We visually evaluated the calibration by examining the models' calibration curves aligned with the diagonal line that represented perfect calibration.^{10 11} Similarly, we created graphs grouping prediction by deciles on the x-axis and the proportion of observed mortality on the y-axis to assess calibration at select time points during a patient stay.¹² These graphs of prediction estimates stratified by deciles are more intuitive for clinicians compared with the traditional

Table 1 Features extracted for three training datasets: features on first calendar day of admission, last available value and selected 1 day at random from patient's stay

Dataset sample	Feature engineering	Variable
Data from admission	Quintile binning on training set for continuous variables	Demographic and hospital characteristics: previous positive COVID-19 PCR test during an outpatient or inpatient visit within 60 days, race, age, sex, body mass index (BMI) and days in hospital (current day minus admission date).
Data from first calendar day at admission, last available value, and 1 day selected at random from patient's stay	Quintile binning on training set variables: current value, first value, minimum value, maximum value, mean value, median value, difference in current value from mean, difference in current value from median, difference in first value from mean, difference in first value from median, difference in max value from mean, difference in max value from median, difference in minimum value from mean and difference in minimum value from median	Laboratory values: albumin, alkaline phosphatase (ALPKPHOS), alanine aminotransferase (ALT), anion gap (ANIONGAP), activated partial thromboplastin time (APTT), aspartate aminotransferase (AST), atypical lymphocytes per cent (ATYLYMREL), bands per cent (BANDSPCT), conjugated bilirubin (BILIDB), bilirubin direct (BILIDIRECT), bilirubin total, natriuretic peptide B (BNPEPTIDE), blood urea nitrogen (BUN), calcium, CKTOTAL, chloride, carbon dioxide (CO2), creatinine, C reactive protein (CRP), d-dimer, glomerular filtration rate – African American (EGGRAA), glomerular filtration rate – non-African American (EGFRNONAA), erythrocyte sedimentation rate (ESR), ferritin, fibrinogen, fraction of inspired oxygen arterial blood gas (FIO2ABG), glucose, HCT, haemoglobin, haemoglobin (HA1C), immunoglobulin A (IGA), immunoglobulin G (IGG), glomerular basement membrane (IGBM), absolute immature granulocytes (IMMGRANABS), per cent immature granulocytes (IMMGRANPCT), interleukin-1 beta (INTERL1B), interleukin 6 (INTRLKN6), potassium (K), potassium plasma (KPLA), lactate arterial blood gas (LACTATEABG), lactate venous blood gas (LACTATEVBG), lactate dehydrogenase (LDH), lipase, lymphocyte absolute calculated (LYMPABSCAL), lymphocyte per cent (LYMPHPCT), lymphocyte absolute (LYMPHSABS), magnesium (MG), sodium (NA), NEUTABSCAL, neutrophil absolute (NEUTSABS), neutrophils per cent (NEUTSPCT), carbon dioxide in arterial blood (PCO2ART), carbon dioxide in venous blood (PCO2VEN), pH of arterial blood (PHART), phosphorous, pH of venous blood (PHVBG), platelet, P02ABG, P02VB, procalcitonin (PROCAL), total protein (PROTTOTAL), prothrombin time (PT), platelet poor plasma (PTT), red blood cell (RBC), troponin (TROPONINI), troponin point of care (TRPNONPOC) and white blood cell count (WBC).
Data from first calendar day at admission, last available value, and 1 day selected at random from patient's stay	Quintile binning on training set: current value, first value, minimum value, maximum value, mean value, median value, difference in current value from mean, difference in current value from median, difference in first value from mean, difference in first value from median, difference in max value from mean, difference in max value from median, difference in minimum value from mean and difference in minimum value from median	Vitals: systolic blood pressure, diastolic blood pressure, pulse pressure, oximetry, respiratory rate, pulse and temperature.

calibration plots used by data science engineers. All model performance measures were reported on external future holdout test set to evaluate most conservatively. We

selected the algorithms and hyperparameters based on the best discrimination using AUC on the associated test sets for each of the three dataset types. The calibration of

the model with the best discrimination was reviewed to ensure it was generally well calibrated. Based on the aforementioned performance metrics, three models derived from each dataset (admission, last-value and time-vary) were selected.

The performance of these final three models were further assessed on ability to discriminate during the duration of patient's entire stay. The AUCs from each day of the patient's stay were plotted to evaluate the models' ability to discriminate over time: 7 days after admission and 7 days prior to discharge. Using estimates from admission and to discharge allows for clearer understanding of accuracy as sample sizes inevitably vary due to early discharge and differences in length of stay. For example, all patients in the test set were in the hospital for 1 day (n=864); however, on day 2, some were discharged (n=859). Similarly, all patients were discharged on their last day of the stay (n=864); however, less patients were in the hospital 2 days prior to their discharge cohort (n=859) as some patients only had a 1-day stay. The model with the highest and largest proportion of AUCs during the time period was selected as the final model. This was determined using the test dataset (17–28 April 2020) of daily values of a patient's stay.

Variable importance

There are algorithms available to facilitate the understanding and trust in machine learning prediction models.¹³ We used the variable importance measure to explore and understand the 'black box' model of the final selected model. Variable importance displays the importance of each variable as measured by its contribution to the change in the residual sum of squared errors value. The scores reflect the contribution each feature makes in classifying or predicting the target outcome, with the contribution stemming from both the feature's role as a primary splitter and its role as a surrogate to any of the primary splitters. The feature with the highest sum of improvements is scored at 100, and subsequent features will have decreasing lower scores. A feature with an importance score of zero indicates it was not used as either a primary or a surrogate splitter, therefore not needed for predictions. Finally, to better understand how each feature impacted the overall prediction and facilitate better visualisation for clinicians, a heat map was created. This was done by creating dummy variables, a mean prediction score calculated for each level of the important features and plotted via a heat map.

All extraction, analysis and visualisation were conducted using SAS base V.9.4 and SAS enterprise miner V.14.3 (SAS Institute, Cary, North Carolina, USA) and Python V.3.8.2 (Seaborn 0.10.0, Pandas 1.0.3, Matplotlib).

RESULTS

Model selection and performance

Of the 3395 discharged patients, 452 (22%), 116 (24%) and 208 (24%) died in the training, validation and test

sets, respectively. The distribution of these features were similar across all three datasets. We used discrimination to assess the model with the best ability to rank patients by risk of mortality. To determine the model with the best discrimination, we used the model with the highest AUC value in their respective test set (table 2). The gradient boosting algorithm had the highest of AUC of 0.83 (95% CI 0.80 to 0.86), 0.93 (95% CI 0.91 to 0.95) and 0.93 (95% CI 0.91 to 0.95) for the admission, last-value and time-vary model, respectively. Table 3 and figure 1 demonstrate as all models approach the time of discharge their ability to discriminate mortality increases. For example, table 2 shows the models' AUC was higher for 7 days after admission (AD 7) versus on admission (AD 1). Similarly, the models' AUC was the higher the day before discharge (DD 2) versus 7 days before discharge (DD 7). All models showed the more data provided to the algorithm, the better the model predicted. However, the model based on admission data had the least improvement in discrimination over time (table 3 and figure 1). Also, of note, all models performed better with the imputed dataset as the imputed data provided inferred missing lab/vital results with the assumption results from the day prior would be similar. After review of the three models on the daily and daily-imputed set, we determined the GB time-vary model with the imputed dataset performed best for our needs as it sustained a higher AUC over time and had better calibration. The hyperparameters for this model in addition to the default included: 150 iterations, 0.1 shrinkage, 70% train proportion, maximum branch 2, maximum depth 5, minimum categorical size 5, missing values use in search, leaf fraction 0.01, number of surrogate rules 0 and subtree assessment using average square error. To assess overfitting, the test was compared with the validation set, and a 0.018 difference in the AUC on admission was found showing the model continued to predict well on external data. The final model's algorithm with an example dataset is available at Zenodo.¹⁴

Figure 2 shows the calibration plots during GB time-vary model in the test set from different time periods of patients stay: on admission, 7 days after admission and 3 days prior to discharge. The model is generally well calibrated although with a slight propensity to overpredict at these various points in time during a patient's stay. Figure 3 depicts a more intuitive presentation of the calibration of the model, via the proportion of observed mortality stratified by prediction risk deciles from the GB time-vary model in the test sets. It also shows the calibration of the model during different time periods: on admission, 7 days after admission and 3 days prior to discharge. The model performed better as the prediction approached discharge. Predictions after 7 days of admission and 3 days before discharge show 98% and 100% of patients in the highest decile of predicted risk died, respectively, and 0% of patients in the lowest decile died for all the time periods. These calibrations by decile offer a more intuitive illustration of the performance of the model for clinicians.

Table 2 AUC, accuracy (acc), sensitivity (sens), specificity (spec), NPV and PPV of LR, DT, GB, SVM and NN models on admission benchmark, last-value and time-varying models in test sets

Model	N	Admission						Last-value						Time-vary					
		AUC	Acc	Sens	Spec	PPV	NPV	AUC	Acc	Sens	Spec	PPV	NPV	AUC	Acc	Sens	Spec	PPV	NPV
		LR	864	0.79	0.80	0.34	0.94	0.66	0.82	0.98	0.23	0.97	0.90	0.97	0.88	0.83	0.66	0.89	0.65
DT	864	0.69	0.76	0.47	0.85	0.49	0.83	0.93	0.21	0.96	0.85	0.94	0.81	0.78	0.61	0.83	0.53	0.87	
GB	864	0.83	0.82	0.53	0.91	0.64	0.86	0.99	0.24	0.97	0.90	0.98	0.93	0.88	0.81	0.90	0.72	0.94	
SVM	864	0.77	0.74	0.56	0.80	0.47	0.85	0.99	0.23	0.94	0.83	0.97	0.85	0.80	0.68	0.84	0.57	0.89	
NN	864	0.82	0.81	0.49	0.92	0.65	0.85	0.97	0.24	0.95	0.87	0.87	0.90	0.84	0.77	0.86	0.63	0.92	

AUC, area under the receiver operating characteristic curve; DT, decision tree; GB, gradient boosting decision trees; LR, logistic regression; NN, neural network; NPV, negative predictive value; PPV, positive predictive value; SVM, support vector machine.

Variable importance

The prediction scores of the model ranged from 0% to 100% with 142 features important to the model. We explored the important feature results using variable importance.¹⁵ Figure 4 shows a heat map of the 30 features most associated with the mortality and the overall per cent association on patient's information on discharge day from the validation and test set combined. It lists the important features along with a calculation of the average change in prediction score of patients for each of level of the feature. Briefly, variable importance varies between zero and one, with higher values indicating features associated more strongly with predictions. Some important features identified by the model included the difference between two values which is results in weight the changes in a patient's value rather than population. The features of pulse oximetry, respirations, systolic blood pressure, blood urea nitrogen, white blood cell, age, length of stay and lymphocyte per cent had a relative importance of at least 2%. On average, patients had at least a 50% increase in their prediction score if they had any of the following characteristics compared with not having it: respirations ranging 22–44, blood urea nitrogen >31, oximetry value <91%, lymphocyte per cent ranging <7, temperature >99.6, calcium ranging 4.1–8.1, mean respirations ranging 23.4–37.2, troponin value 0.09–69.4 and neutrophil percentage ranging 84–99. Conversely, patients with a median-min difference in oximetry value of 0–0.5, respiratory mean-min difference of 0–0.97 or lymphocyte per cent of 24–93 had at least a 20% decrease in their prediction for mortality.

DISCUSSION

This study describes the development of a machine learning model to predict mortality of patients who present and are admitted to the hospital with a confirmed COVID-19 by PCR and provide an accurate daily risk estimate during the patient's stay. The aim of this study was to explore and compare three methods to build a model that could accurately predict risk of death on admission and at each day during the stay of the patient.

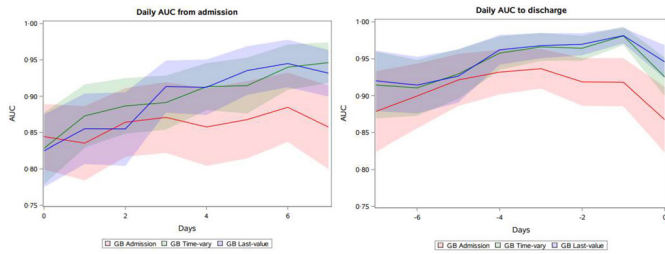
A strength of the current study was the use over 3000 discharges in a US population. We plan to apply this model to data exported out of clarity each day and provide clinician with daily prediction estimates. The model can be found at Zenodo, along with a sample table that is created prior to applying the model. Unlike other models, it does not require manual calculation of a score, a welcome improvement for the busy clinician. Because the model has high accuracy and is well calibrated, it can be used in other studies as an objective estimation of disease severity. The objective nature of the model is important as it limits biases from documentation issues of overwhelmed clinicians and differences in treatments and provides transparent objective data to characterise severity of a novel disease. Additionally, novel feature engineering methodologies were included such as changes in laboratory/vital



Table 3 Daily AUC, accuracy (acc), sensitivity (sens), specificity (spec), NPV and PPV comparison of GB admission, last-value and time-vary models on patient's daily prediction (a) AUC, accuracy, sens, spec, NPV and PPV of daily prediction test set

Day	N	Admission					Last-value					Time-vary							
		AUC	Acc	Sens	Spec	PPV	NPV	AUC	Acc	Sens	Spec	PPV	NPV	AUC	Acc	Sens	Spec	PPV	NPV
Days After Admission 1	864	0.84	0.82	0.53	0.91	0.64	0.86	0.82	0.79	0.40	0.92	0.61	0.83	0.83	0.77	0.58	0.82	0.51	0.86
Days After Admission 2	859	0.84	0.79	0.60	0.86	0.57	0.87	0.86	0.82	0.49	0.93	0.70	0.85	0.87	0.80	0.60	0.86	0.59	0.87
Days After Admission 3	822	0.86	0.79	0.69	0.83	0.57	0.89	0.85	0.81	0.46	0.93	0.68	0.84	0.89	0.83	0.64	0.90	0.67	0.88
Days After Admission 4	759	0.87	0.78	0.77	0.79	0.56	0.91	0.91	0.84	0.57	0.94	0.76	0.86	0.89	0.84	0.73	0.88	0.69	0.90
Days After Admission 5	680	0.86	0.75	0.77	0.74	0.53	0.89	0.91	0.83	0.57	0.93	0.77	0.85	0.91	0.84	0.76	0.87	0.69	0.90
Days After Admission 6	603	0.87	0.75	0.80	0.73	0.55	0.90	0.94	0.83	0.59	0.92	0.77	0.85	0.91	0.83	0.73	0.87	0.70	0.89
Days After Admission 7	539	0.88	0.74	0.83	0.70	0.56	0.90	0.94	0.82	0.62	0.91	0.76	0.84	0.94	0.86	0.81	0.89	0.77	0.91
Days Before Discharge 1	864	0.87	0.77	0.83	0.76	0.52	0.93	0.95	0.94	0.85	0.96	0.88	0.95	0.93	0.95	0.93	0.95	0.86	0.98
Days Before Discharge 2	859	0.92	0.81	0.93	0.78	0.57	0.97	0.98	0.94	0.81	0.98	0.91	0.94	0.98	0.93	0.90	0.94	0.84	0.97
Days Before Discharge 3	822	0.92	0.82	0.91	0.79	0.58	0.96	0.97	0.91	0.74	0.97	0.88	0.92	0.96	0.92	0.87	0.93	0.81	0.96
Days Before Discharge 4	759	0.94	0.80	0.90	0.77	0.58	0.96	0.97	0.89	0.71	0.96	0.86	0.90	0.97	0.88	0.86	0.89	0.74	0.95
Days Before Discharge 5	680	0.93	0.77	0.86	0.73	0.55	0.93	0.96	0.88	0.72	0.94	0.82	0.89	0.96	0.86	0.83	0.87	0.71	0.93
Days Before Discharge 6	603	0.92	0.77	0.85	0.73	0.57	0.92	0.93	0.84	0.67	0.92	0.77	0.87	0.93	0.85	0.78	0.88	0.73	0.91
Days Before Discharge 7	539	0.90	0.76	0.82	0.73	0.58	0.90	0.91	0.82	0.64	0.91	0.76	0.84	0.91	0.81	0.75	0.83	0.68	0.88

AUC, area under the receiver operating characteristic curve; GB, gradient boosting decision trees; NPV, negative predictive value; PPV, positive predictive value.



A AUCs each day from admission of patients' stay **B** AUCs each day to discharge of patients' stay

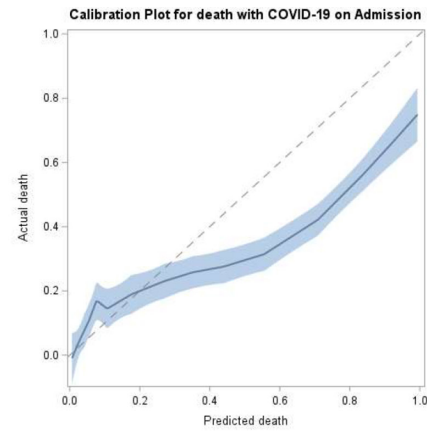
Figure 1 Daily AUCs from the three final models (admission, last-value and time-vary) and their performance over time (7 days after admission and prior to discharge) on the test set and 'imputed' test set (N=864). (A) Compares the AUCs each day from admission of patients' stay. (B) Compares the AUCs each day to discharge of patients' stay. AUC, area under the receiver operating characteristic curve.

results within the context of an individual patient, rather than in the population only, which helped to improve the model's predictions over the course of a patient's stay.

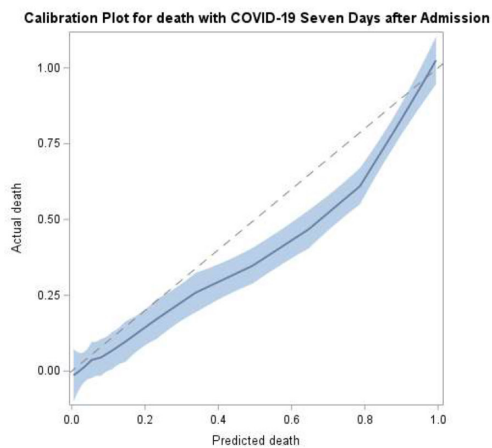
Prior studies suggest AI has been slowly gaining traction in healthcare due to the perception that machine learning models are 'black boxes' or not interpretable by the user.¹⁶ The methods demonstrated in this study are more approachable and easily understood by the clinician. This study presented calibration via deciles that is more intuitive for the non-data scientist. Also, a heat map was created to present results from the variable importance algorithm—the distribution of prediction estimates across the binned variables. Users might hesitate to rely on AI for decisions without knowing the risk factors driving the model, despite the computer making accurate recommendations. By providing user's information about the model such as variable importance, the association of each feature level with the outcome provides additional insights serve to facilitate trust that is needed to increase the adoption of AI in the healthcare industry.

This model highlights individualised current and prior laboratory and vital results to determine patient-specific mortality risk. Important determinants of risk are further evaluated to illustrate the changes in prediction among patient populations. The interpretability of the model in this study serves to provide insights to intensivists, researchers and administrators of predictors for survivability from a disease with unpredictable or little known outcomes.

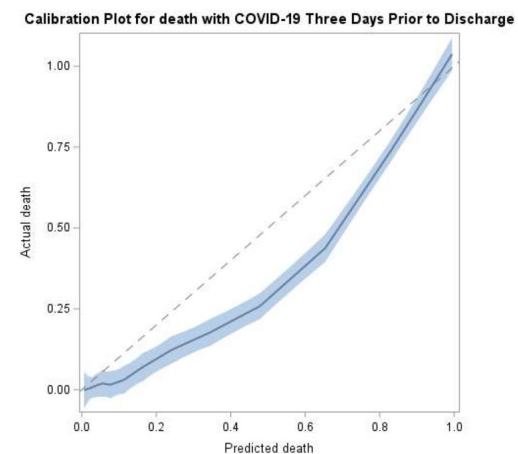
This retrospective study applied machine learning algorithms to structured patient data from the EHR of a large urban academic health system to create a risk prediction model to predict mortality during admission in patients with confirmed COVID-19. With an AUC of 0.83 at admission, and 0.97 3 days prior to discharge on imputed data, the model discriminates well and is well calibrated. Additionally, the final model's AUC was consistent on both the time held out internal validation and external test sets, which gives more confidence the model will continue to perform well on future data. Because we continue to have large amounts of discharges daily, potential changes



A Prediction deciles on admission



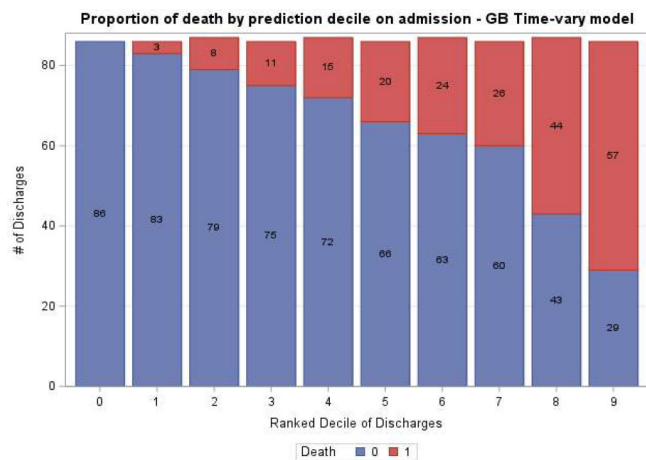
B Prediction deciles seven days after admission



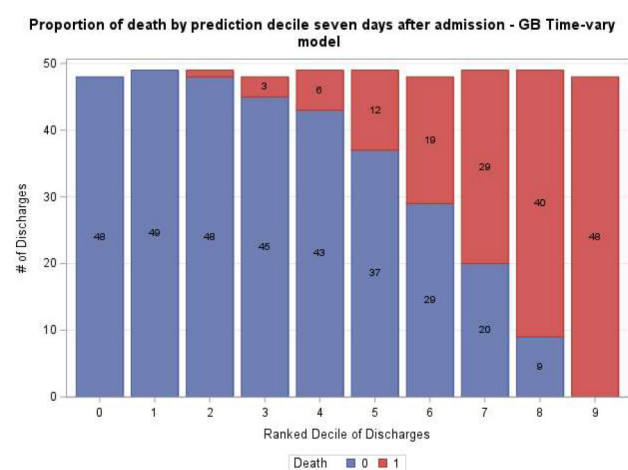
C Prediction deciles three days prior to discharge

Figure 2 Calibration plots using time-vary model on test set (A) on admission, (B) 7 days after admission and (C) 3 days before discharge (N=864). The plots show a slight propensity for the model to over predict during various points of patients' stays.

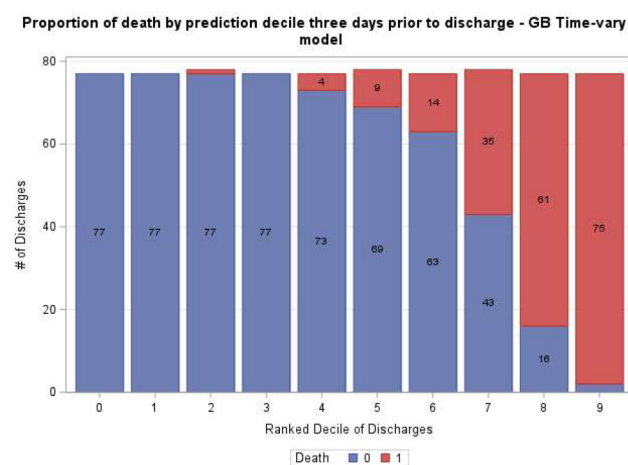
in populations and modification of treatment protocols, we plan to continue to monitor performance and retrain model when discrimination falls below 0.8. Ideally, the



A Prediction deciles on admission



B Prediction deciles seven days after admission



C Prediction deciles three days prior to discharge

Figure 3 Proportion of actual mortality by predicted mortality score decile ranking in imputed test set. (A) On admission, (B) 7 days after admission and (C) 3 days before discharge (N=864). The model shows an increase in actual mortality among decile groups with higher predicted mortality.

monitoring of the AUC score should be automated and alert the data scientist when the value falls below a predefined threshold. Hospitals should consider developing their own mortality prediction models based on their specific cohorts, as patient populations may differ across facilities therefore affecting validation results.¹⁷

Finally, and perhaps most importantly, implementation plays a critical role in supporting in the adoption of AI as healthcare systems face increasingly dynamic and resource-constrained conditions.¹⁸⁻¹⁹ While a plethora of literature exists addressing data acquisition, development and validation of models, the application of AI in a real-world healthcare setting has not been substantially addressed.²⁰⁻²² Often, prediction model results are used to risk adjust and benchmark rates of an outcome.²³⁻²⁵ In addition to using prediction estimates as part of a tool, we suggest models be used as tool in the process of understanding and studying a disease.

Limitations/next steps

The usual limitations associated with an EHR might affect our model. While this model relies on mostly objective data, some inherent bias might be introduced in terms of demographic and laboratory/vital collection and documentation. For example, certain laboratory tests might be ordered on sicker patients or certain types of clinicians might use similar ordering practices that would bias the model. Therefore, the model might be relying on the subjective nature of a clinician rather than purely objective data. On a similar note, patients that might have died after discharge would bias the model. As suggested earlier, results from the model may not be generalisable to other institutions or patient populations; therefore, hospitals should develop tailored models for their own patient population, especially for a disease that is not yet well understood. Because of this, the 'external validation' dataset in this study does not meet the TRIPOD definition as it is using a sample from the same patient population although future population. Furthermore, models need to be continually monitored and retrained when performance degrades. Lastly, this model intends to allocate resources, ensure basic and routine care is completed and quantify the health of a patient.

The prediction estimates can be used to create reports adjusting mortality rates by physician, ward or hospital facility. The estimates can also be used to identify high performers to gain insights on potential successful aspects of their care and treatment. The model can be further enhanced by predicting patients who are most likely to unexpectedly expire to gain more insights on how predictors compare with current model. The estimates can also be used for other studies where an objective metric for disease severity is needed. Finally, prediction estimates can be incorporated into an AI tool that can allow clinicians facing a new illness with an uncertain course to identify and prioritise patients who might benefit from targeted, experimental therapy.

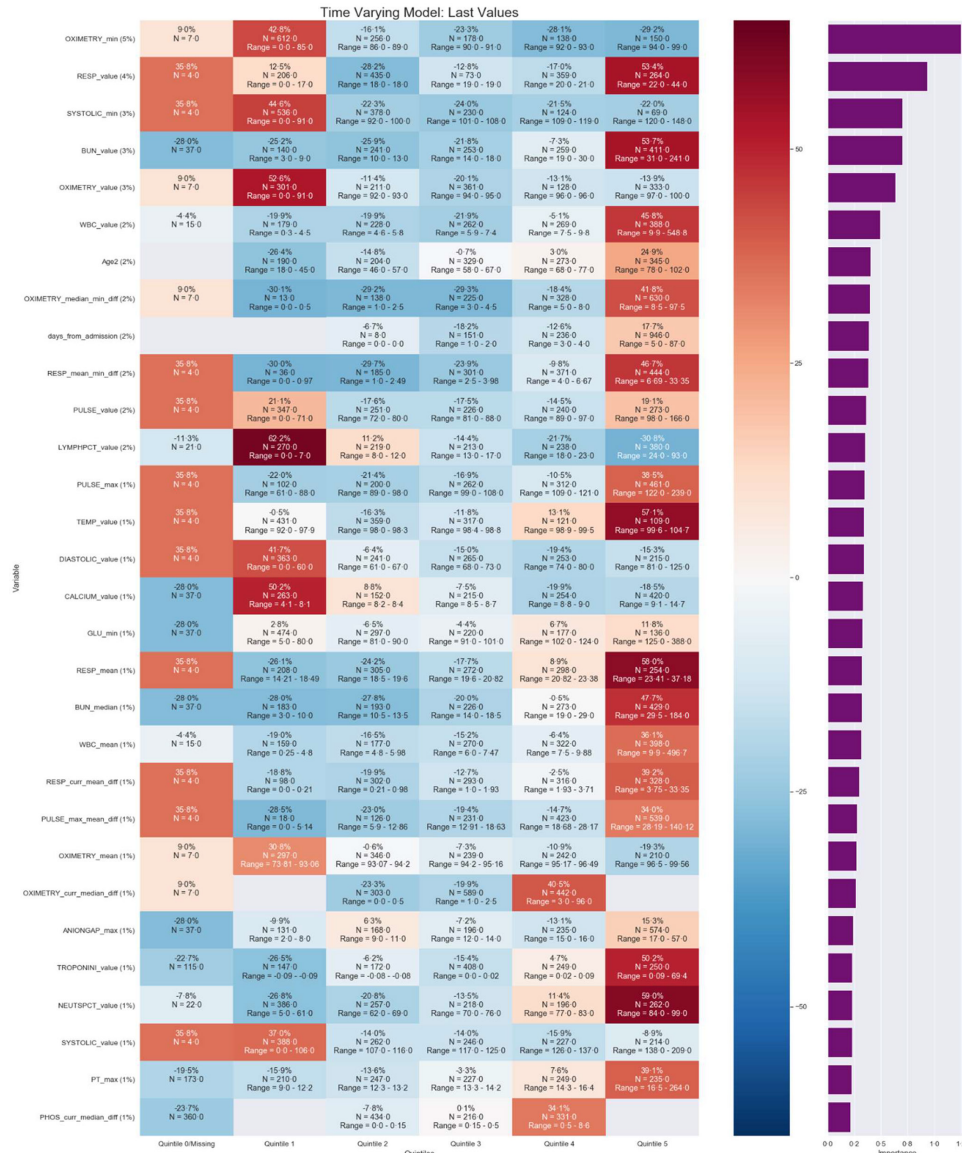


Figure 4 Ranking of most important 30 of 142 features of the final selected model based on per cent relative importance of the last lab value available in the test set. The purple graph on rightmost column of figure displays the variable importance value. The map also lists the average influence of a feature's level on a patient's overall prediction score with darker red boxes and darker blue boxes indicating an increase and decrease in the prediction, respectively (N = 864). Full map in supplemental material.

CONCLUSION

Hospitals can develop customised prediction models as the amount of EHR data increases, computing power and speeds are faster and machine learning algorithms are broadly accessible. During times of high demand and large uncertainty around a disease, prediction models can help to identify underlying patterns of predictors of disease and be deployed. This study shows how to build a prediction model whereby the predictions improve during the patient's course of stay. Results from a highly accurate model can serve as an objective measure of disease severity where manual review of every cases is not feasible. Similar to other industries, machine learning should be integrated into research and healthcare workflows to better understand and study a disease as well as be

incorporated into tools to assist in care, allocate resources and aid in discharge decisions to hopefully save lives.

Acknowledgements The authors would like thank to Dr Levi Waldron for his general guidance and mentorship on methodology.

Contributors All of the authors were involved in the conception and design of the study, the acquisition of data, the analysis or interpretation of data, and drafting and critical revision of the manuscript for important intellectual content. All approved the final version submitted for publication. AS had full access to the data in the study and takes responsibility for the integrity of the data and accuracy of the data analysis. This study was approved by the NYU Grossman School of Medicine Institutional Review Board (No i20-00671_MOD02), which granted both a waiver of informed consent and a waiver of the Health Information Portability and Privacy Act.

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests None declared.

Patient consent for publication Not required.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement No data are available. The raw data are not available; however, the model is available at Zenodo (in citation of manuscript).

Supplemental material This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been peer-reviewed. Any opinions or recommendations discussed are solely those of the author(s) and are not endorsed by BMJ. BMJ disclaims all liability and responsibility arising from any reliance placed on the content. Where the content includes any translated material, BMJ does not warrant the accuracy and reliability of the translations (including but not limited to local regulations, clinical guidelines, terminology, drug names and drug dosages), and is not responsible for any error and/or omissions arising from translation and adaptation or otherwise.

Open access This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: <http://creativecommons.org/licenses/by-nc/4.0/>.

ORCID iD

Anna Stachel <http://orcid.org/0000-0002-7362-9390>

REFERENCES

- 1 NYC DOHMH. COVID-19: data, 2020. Available: <https://www1.nyc.gov/site/doh/covid/covid-19-data.page>
- 2 Lagu T, Pekow PS, Stefan MS, *et al*. Derivation and validation of an in-hospital mortality prediction model suitable for profiling Hospital performance in heart failure. *J Am Heart Assoc* 2018;7:e005256.
- 3 Johnson AEW, Mark RG. Real-Time mortality prediction in the intensive care unit. *AMIA Annu Symp Proc* 2017;2017:994–1003.
- 4 Schwarzkopf D, Fleischmann-Struzek C, Rüdell H, *et al*. A risk-model for hospital mortality among patients with severe sepsis or septic shock based on German national administrative claims data. *PLoS One* 2018;13:e0194371.
- 5 Raoult D, Zumla A, Locatelli F, *et al*. Coronavirus infections: epidemiological, clinical and immunological features and hypotheses. *Cell Stress* 2020;4:66–75.
- 6 Adam J, Adamová D, Aggarwal MM, *et al*. Anomalous Evolution of the Near-Side Jet Peak Shape in Pb-Pb Collisions at $\sqrt{s_{NN}}=2.76$ TeV. *Phys Rev Lett* 2017;119:102301.
- 7 Vincent J-L, Taccone FS. Understanding pathways to death in patients with COVID-19. *Lancet Respir Med* 2020;8:430–2.
- 8 Moons KGM, Altman DG, Reitsma JB, *et al*. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): explanation and elaboration. *Ann Intern Med* 2015;162:W1–73.
- 9 Min X, Yu B, Wang F. Predictive modeling of the hospital readmission risk from patients' claims data using machine learning: a case study on COPD. *Sci Rep* 2019;9:2362.
- 10 Austin PC, Steyerberg EW. Graphical assessment of internal and external calibration of logistic regression models by using Loess smoothers. *Stat Med* 2014;33:517–35.
- 11 Steyerberg EW, Vickers AJ, Cook NR, *et al*. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology* 2010;21:128–38.
- 12 Reips JM, Schuemie MJ, Suchard MA, *et al*. Design and implementation of a standardized framework to generate and evaluate patient-level prediction models using observational healthcare data. *J Am Med Inform Assoc* 2018;25:969–75.
- 13 Lundberg SM, Lee SI. A unified approach to interpreting model predictions. *Adv Neur In* 2017;30.
- 14 Stachel A. Development and validation of a machine learning prediction model as part of an AI notification tool to predict mortality risk in patients with COVID-19 2020. doi:10.5281/zenodo.3893846
- 15 Friedman JH PB. *Predictive learning via rule ensembles*, 2005.
- 16 McGovern A, Balfour A, Beene M, *et al*. Storm Evader: using an iPad to teach kids about Meteorology and technology. *Bull Am Meteorol Soc* 2015;96:397–404.
- 17 Siontis GCM, Tzoulaki I, Castaldi PJ, *et al*. External validation of new risk prediction models is infrequent and reveals worse prognostic discrimination. *J Clin Epidemiol* 2015;68:25–34.
- 18 Shelton RC, Lee M, Brotzman LE, *et al*. What is dissemination and implementation science?: an introduction and opportunities to advance behavioral medicine and public health globally. *Int J Behav Med* 2020;27:3–20.
- 19 Bauer MS, Damschroder L, Hagedorn H, *et al*. An introduction to implementation science for the non-specialist. *BMC Psychol* 2015;3:32.
- 20 Amarasingham R, Patzer RE, Huesch M, *et al*. Implementing electronic health care predictive analytics: considerations and challenges. *Health Aff* 2014;33:1148–54.
- 21 CMS. The skilled nursing facility value-based purchasing program (snf VBP), 2018 [Accessed 10 Oct 2019].
- 22 Kansagara D, Englander H, Salanitro A, *et al*. Risk prediction models for hospital readmission: a systematic review. *JAMA* 2011;306:1688–98.
- 23 Lefering R, Huber-Wagner S, Nienaber U, *et al*. Update of the trauma risk adjustment model of the TraumaRegister DGU™: the revised injury severity classification, version II. *Crit Care* 2014;18:476.
- 24 Shaw RE, Anderson HV, Brindis RG, *et al*. Development of a risk adjustment mortality model using the American College of Cardiology-National cardiovascular data registry (ACC-NCDR) experience: 1998-2000. *J Am Coll Cardiol* 2002;39:1104–12.
- 25 Pine M, Jordan HS, Elixhauser A, *et al*. Enhancement of claims data to improve risk adjustment of hospital mortality. *JAMA* 2007;297:71–6.

© 2021 Author(s) (or their employer(s)) 2021. Re-use permitted under CC BY-NC. No commercial re-use. See rights and permissions. Published by BMJ. This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: <http://creativecommons.org/licenses/by-nc/4.0/>. Notwithstanding the ProQuest Terms and Conditions, you may use this content in accordance with the terms of the License.

Comparison and impact of COVID-19 for patients with cancer: a survival analysis of fatality rate controlling for age, sex and cancer type

Haiquan Li ¹, Edwin Baldwin,¹ Xiang Zhang,¹ Colleen Kenost,² Wenting Luo,¹ Elizabeth A Calhoun,³ Lingling An,¹ Charles L Bennett,⁴ Yves A Lussier ²

To cite: Li H, Baldwin E, Zhang X, *et al*. Comparison and impact of COVID-19 for patients with cancer: a survival analysis of fatality rate controlling for age, sex and cancer type. *BMJ Health Care Inform* 2021;**28**:e100341. doi:10.1136/bmjhci-2021-100341

► Additional supplemental material is published online only. To view, please visit the journal online (<http://dx.doi.org/10.1136/bmjhci-2021-100341>).

HL and EB contributed equally.

Received 21 February 2021
Revised 08 April 2021
Accepted 20 April 2021



© Author(s) (or their employer(s)) 2021. Re-use permitted under CC BY-NC. No commercial re-use. See rights and permissions. Published by BMJ.

For numbered affiliations see end of article.

Correspondence to

Dr Haiquan Li;
haiquan@arizona.edu

Dr Yves A Lussier;
Yves.Lussier@utah.edu

Dr Charles L Bennett;
bennettc@cop.sc.edu

ABSTRACT

Objectives Prior research has reported an increased risk of fatality for patients with cancer, but most studies investigated the risk by comparing cancer to non-cancer patients among COVID-19 infections, where cancer might have contributed to the increased risk. This study is to understand COVID-19's imposed HR of fatality while controlling for covariates, such as age, sex, metastasis status and cancer type.

Methods We conducted survival analyses of 4606 cancer patients with COVID-19 test results from 16 March to 11 October 2020 in UK Biobank and estimated the overall HR of fatality with and without COVID-19 infection. We also examined the HRs of 13 specific cancer types with at least 100 patients using a stratified analysis.

Results COVID-19 resulted in an overall HR of 7.76 (95% CI 5.78 to 10.40, $p < 10^{-10}$) by following 4606 patients with cancer for 21 days after the tests. The HR varied among cancer type, with over a 10-fold increase in fatality rate (false discovery rate ≤ 0.02) for melanoma, haematological malignancies, uterine cancer and kidney cancer.

Although COVID-19 imposed a higher risk for localised versus distant metastasis cancers, those of distant metastases yielded higher overall fatality rates due to their multiplicative effects.

Discussion The results confirmed prior reports for the increased risk of fatality for patients with COVID-19 plus hematological malignancies and demonstrated similar findings of COVID-19 on melanoma, uterine, and kidney cancers.

Conclusion The results highlight the heightened risk that COVID-19 imposes on localised and haematological cancer patients and the necessity to vaccinate uninfected patients with cancer promptly, particularly for the cancer types most influenced by COVID-19. Results also suggest the importance of timely care for patients with localised cancer, whether they are infected by COVID-19 or not.

INTRODUCTION

In localised cancers or haematological malignancies, timely cancer diagnosis and treatment is critical for increasing a patient's survivability. Otherwise, localised cancers may progress into distant (distant organ systems) metastasis,^{1 2} while distant metastases may

become uncontrollable, both of which result in more fatalities.³⁻⁵ However, with COVID-19 evidently impacting cancer care, diagnosis and treatment delays are inevitable due to the unavailability of medical resources, potential exposure risks of COVID-19 in medical facilities and complications of treatment (eg, chemotherapies worsening the fatality rate)^{6 7} attributed to the weaker immune systems of patients with cancer. Therefore, attention to the timeliness of therapy for patients with cancer is encouraged to minimise the risk of fatality.³ Still, the extent of risk that delays in cancer therapies add for persons with COVID-19 is not known^{8 9} and is likely to vary depending on cancer type,¹⁰ stage, grade and treatment.^{9 11} Therefore, estimating the risk COVID-19 imposes on each type of cancer is critical. Although prior research has been conducted, most studied the HR or OR by comparing patients with cancer to non-cancer patients among patients with COVID-19, which did not reflect the impact of COVID-19 on specific cancer types, could be confounded by the therapy types¹²⁻¹⁴ and is limited by the small sample size available for specific cancer types. Exceptions included Passamonti *et al*'s¹⁵ study, which reported 536 haematological cancer patients with COVID-19 infection at 66 hospitals in Italy. Those patients with severe or critical COVID-19 infections had an HR of 4.08 for mortality when taking mild severity as reference in Italy.¹⁵ However, the association between severity and fatality is evident, and the study was not designed to study the added risk of fatality from COVID-19 infection for the haematological malignancies. Here, we report a study comparing fatality rates among persons with a wide range of cancer diagnoses with and without COVID-19 while controlling for age, sex and type of cancer. Age and sex are essential biological

variables underlying the fatality of COVID-19^{16–19} and thus should be controlled for. We further studied the added COVID-19 risk to fatality for specific cancer types to aid oncologists in making optimal treatment decisions from various risk factors, several of which may be contradictory, such as delay of care and infection risk.

METHODS

We conducted a retrospective survival study using UK Biobank (UKB)²⁰ under the UKB COVID-19 policy.²¹ Started in 2006, UKB is a government funded biobank with longitudinal COVID-19 test results, death registries, cancer registries and inpatient records for approximately 500 000 patients.²² COVID-19 tests started on 16 March 2020 for symptomatic patients, during which testing capacity was limited and results were provided by Public Health England.²³ There were roughly 67 000 living cancer subjects at the beginning of COVID-19 testing. Inclusion criteria for the study were: (1) subjects of British ancestry with a history of hospitalisation in UKB (updated to the end of September 2020) and (2) subjects conducted a COVID-19 test no later than 11 October 2020, for which we could obtain a 21-day follow-up in the death registry from National Health Service (NHS) Digital and NHS Central Register, UK. The inclusion criteria resulted in 6528 cancer patients. We then excluded 893 patients with cancer whose cancer diagnoses were 10 or more years ago, without another primary cancer or recurrence in the record since they are unlikely in remission and more closely resemble non-cancer patients. We further excluded five cancer patients with inconsistent self-report of sex and 1024 ‘non-melanoma skin cancer patients’ reported with truncated International

Classification of Diseases, Tenth Revision (ICD-10) codes in the UKB, which conflates non-lethal basal cell carcinomas with lethal forms of non-melanoma skin cancers. The final cancer cohort comprised 4606 patients, where 288 (6.3%) were positive for COVID-19 (table 1). We also built a randomised non-cancer cohort of 4606 patients for comparative studies, which matched the COVID-19 status, sex, age (per 5-year bin) and specific laboratory testing facility of the corresponding patients with cancer. We exclude seven non-cancer patients with COVID-19 tests conducted after death during the sampling.

We analysed the case fatality rate (CFR) of patients with cancer and their cancer types using a multivariate Cox proportional hazard model.²⁴ COVID-19 status was obtained from the patient’s first test result (negative controls remained so throughout). Control covariates included ethnicity, age, sex, and cancer status. Age was treated as a continuous variable, and cancer status was categorised as localised versus distant metastatic based on UKB ICD-10 codes from inpatient records where distant metastasis status was characterised by ICD-10 codes of C78 (metastatic to respiratory and digestive organs), C79 (metastatic to other and unspecified sites) and C80 (metastasis without specific sites or multiple sites), whereas localised metastatic status were patients without any of the three codes, possibly including C77 (spread to lymph nodes) as well. R’s *survival* package²⁵ was used to diagnose the model and plot Kaplan-Meier curves.

RESULTS

Quartiles of time from COVID-19 diagnosis to death for patients with cancer (64 total) are 5, 10 and 14 days, respectively. The CFR within 21 days of diagnosis for COVID-19

Table 1 Demographic and clinical characteristics of 4606 COVID-19 tested cancer subjects in the UK Biobank

	COVID-19 positive associated			COVID-19 negative associated		
	Fatalities	(%)	Survivors	Fatalities	(%)	Survivors
Race						
British ancestry	64	(22.2)	224	153	(3.5)	4165
Age (years)						
50–59	2	(7.1)	26	8	(2.9)	271
60–69	6	(9.7)	56	26	(3.0)	850
70–79	42	(26.4)	117	94	(3.5)	2562
80–84	14	(35.9)	25	25	(4.9)	482
Sex						
Male	40	(23.8)	128	90	(3.8)	2285
Female	24	(20.0)	96	63	(3.2)	1880
History of cancer (years)						
>10	18	(27.7)	47	35	(3.4)	1008
5<years ≤10	15	(18.5)	66	24	(2.3)	1005
1<years ≤5	3	(13.0)	20	8	(4.3)	178
≤1	28	(23.5)	91	86	(4.2)	1974

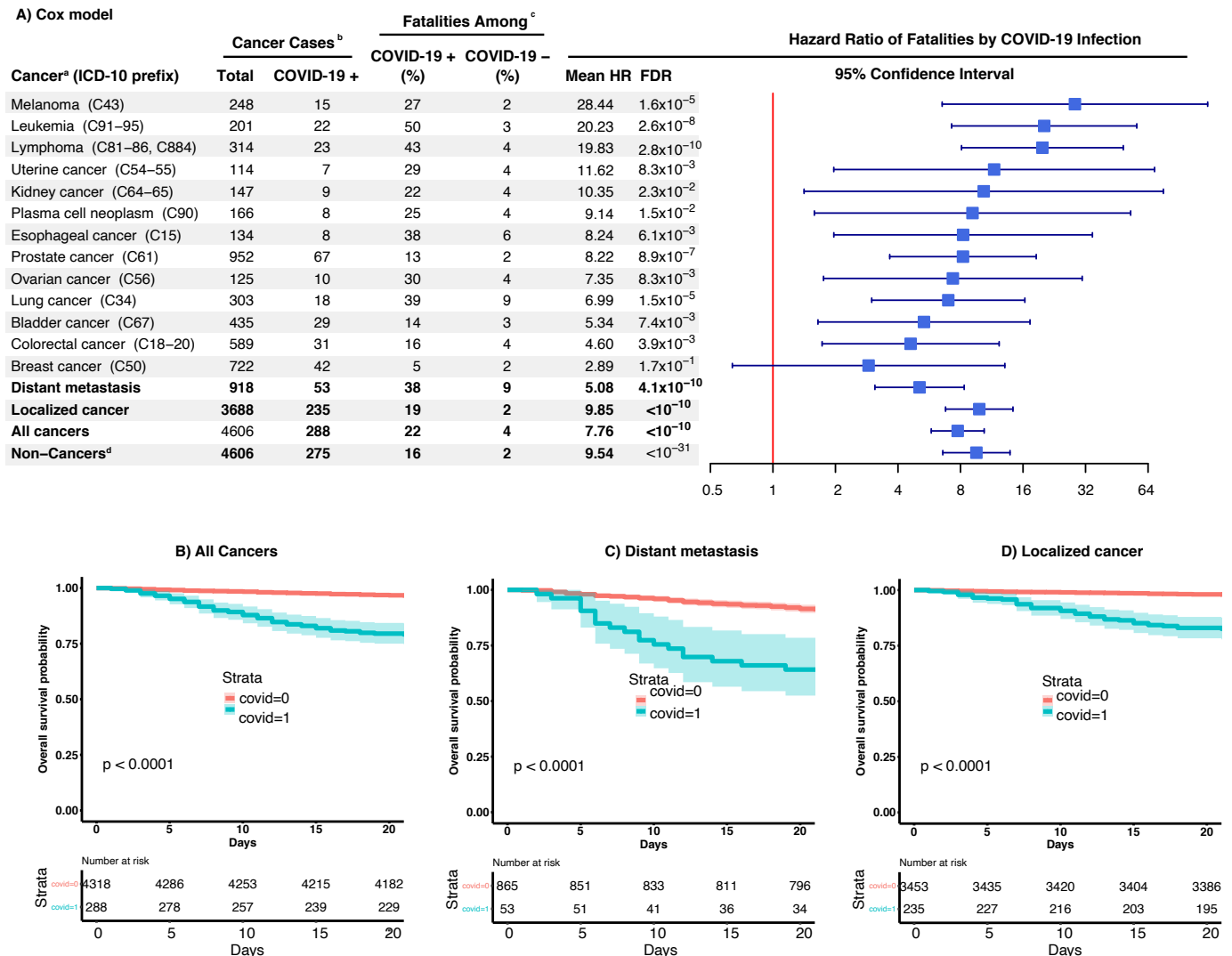


Figure 1 (A) HRs of COVID-19 associated death among patients with cancer and types of cancers. (a) Results were from stratified analyses using the patients stratifying conditions listed in the column named ‘cancer’. All analyses used Cox model with covariates of sex, age, COVID-19 infection status, and cancer status (localised or distant metastasis) except for stratified analyses of localised cancer, distant metastasis, and non-cancers. Cancer subtypes were analysed when they were composed of 100 or more total cases. (b) Four patients with cancer with inconsistent self-reporting and gene sex were excluded from the study. (c) Fatality event was assessed for the 21 days following the first COVID-19 positive testing or the first COVID-19 negative testing (negative controls remained so throughout) and was available for all cancer subjects under study. (d) Sixty non-cancer patients did not match the testing facility with a patient with cancer due to lacking subjects of matching all factors (eg, age). (B-D) Kaplan-Meier curves for COVID-19 positive vs. negative cancer patients: (B) all cancer patients, (C) patients with distant metastasis, and (D) patients with only localized cancer.

positive patients with cancer was sixfold higher than that of COVID-19 negative ones (22.2% (64 out of 288) versus 3.5% (153 out of 4318)) (table 1). Distant metastasis patients demonstrated a higher fatality rate, where CFR within 21 days of a positive COVID-19 diagnosis was nearly double that of patients with localised cancer (37.7% versus 18.7%) (figure 1). No matter the spread of cancer, the CFR of COVID-19 within 21 days (22.2%) was higher than that of non-cancer patients when positive (15.6%), from the cohort using matched covariates such as test result, sex, age and testing venue (a proxy of the hospitalisation system). A multivariate Cox model was built to study individual risk factors and showed an HR of 7.76

(95% CI 5.78 to 10.40, $p < 10^{-10}$) for COVID-19 positive patients with cancer compared with COVID-19 negative patients with cancer after controlling for ethnicity, sex, age and metastatic versus localised cancer confounders (figure 1). The model suggested increased fatality rates by COVID-19 infections particularly for patients with melanoma, lymphoma, leukaemia, uterine or kidney cancer. The HRs of COVID-19 were 10-fold higher for these five cancer diagnoses using stratified analyses with matched cancer types (figure 1). As expected, distant metastasis was a risk factor for fatality, with an HR of 3.92 (95% CI 2.99 to 5.13, $p < 10^{-10}$) compared with localised cancers. Age remained an important factor, with an HR of 1.04 per

year (95% CI 1.02 to 1.07; $p=3.1 \times 10^{-4}$), implying an over 10-fold higher fatality risk for the 60-year difference that may exist between the seniors and youths. Sex was not deemed significant. No factor was found to significantly deviate from the model assumption ($p>0.25$). Logistic regression led to similar but slightly inflated results because the logistic model does not require constant risk over time, whereas the Cox model does.

DISCUSSION

Our study demonstrated that COVID-19 adds 10-fold more risk to 21-day fatality rates for patients with melanoma, lymphoma, leukaemia, uterine and kidney cancer with a positive COVID-19 infection versus no COVID-19 infection. Our results for lymphoma and leukaemia patients confirm reports from Italy, while the findings of increased 21-day mortality rates in COVID-19 infections among melanoma, uterine and kidney cancer patients have not been reported previously. The findings do support prior kidney injury reports among patients with COVID-19 as well.²⁶ Fortunately, our study suggests that COVID-19 does not impose a larger risk to distant metastasised cancers as compared with localised cancers (eg, lymphoma and leukaemia) in general. However, the overall fatality rate in distant metastases was still about twice that of localised cancers due to the multiplicative effect of HRs in the model. It should be noted that fatality rates were dependent on cancer type, and COVID-19 did impose larger risk to distant metastasis of some types of cancer, such as melanoma (HR 49.37, 95% CI 3.70–658.85), prostate cancer (HR 22.11, 95% CI 6.15 to 79.54) and ovarian cancer (HR 13.04, 95% CI 2.61 to 65.14), based on a stratified analysis using only metastasis patients (online supplemental figure 1). In all cases, higher rates of fatality among patients with COVID-19 of older age were consistent with literature.^{16 17 27} Our results focused on fatality are complementary to those few studies focused on oncological procedures unimpacted by COVID-19. Indeed, investigations of the National Cancer Data Base on prostate radiotherapy²⁸ and breast cancer surgeries²⁹ report unchanged overall survival during COVID-19.

The strengths of this study include the large UKB cancer cohort size for patients with COVID-19 and its reliable death registry. Limitations include the unavailability of complete cancer stage and grade, plus a relatively small sample size for some specific cancer types. Furthermore, the study was unable to include other pre-existing conditions that may have been associated with the fatality, and the conclusions may be limited to symptomatic patients and hospitalised patients due to the inclusion criteria.

Our findings reinforce the clinical importance of timely treatment of COVID-19 among older cancer patients with localised cancers.¹ Timely care is especially important for those with haematological malignancies, melanoma, uterine or kidney cancer due to the notable additional risk of fatality from COVID-19 infection plus the added risk of metastasis due to the delay of therapies, which

leads to an even higher likelihood of fatality because of the multiplicative effects of risk. The findings also support specific guidelines³⁰ emphasising the importance of timely care for COVID-19 infected patients with cancer and strongly support a change in COVID-19 vaccine strategy with haematological malignancies in particular because the benefit of vaccination far outweighs its risk of side effects.^{31 32}

Author affiliations

¹Department of Biosystems Engineering, The University of Arizona, Tucson, Arizona, USA

²Department of Biomedical Informatics, The University of Utah School of Medicine, Salt Lake City, Utah, USA

³Department of Population Health, University of Kansas Medical Center, Kansas City, Kansas, USA

⁴Department of Clinical Pharmacy and Outcomes Sciences, University of South Carolina, Columbia, South Carolina, USA

Twitter Haiquan Li @haiquanlab and Yves A Lussier @LussierY

Acknowledgements This research has been conducted using the UK Biobank Resource under Application Number 28979.

Contributors All authors contributed to the conception, analysis, and interpretation of the results. YL, CB, HL, and EC conceived the study. HL led the analysis. CB and YL directed the oncology discussion. EB undertook the data processing and wrote the first draft. XZ and LA performed the model diagnosis. WL and CK designed the figures. All authors revised the manuscript.

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests None declared.

Patient consent for publication Not required.

Provenance and peer review Commissioned; internally peer reviewed.

Supplemental material This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been peer-reviewed. Any opinions or recommendations discussed are solely those of the author(s) and are not endorsed by BMJ. BMJ disclaims all liability and responsibility arising from any reliance placed on the content. Where the content includes any translated material, BMJ does not warrant the accuracy and reliability of the translations (including but not limited to local regulations, clinical guidelines, terminology, drug names and drug dosages), and is not responsible for any error and/or omissions arising from translation and adaptation or otherwise.

Open access This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: <http://creativecommons.org/licenses/by-nc/4.0/>.

ORCID iDs

Haiquan Li <http://orcid.org/0000-0002-8049-0278>

Yves A Lussier <http://orcid.org/0000-0001-9854-1005>

REFERENCES

- Dirie J, Mahesan T, Hart E, *et al*. Delivering safe and timely cancer care during COVID-19: lessons and successes from the transition period. *BJU Int* 2021. doi:10.1111/bju.15343. [Epub ahead of print: 21 Jan 2021].
- Jones D, Neal RD, Duffy SRG, *et al*. Impact of the COVID-19 pandemic on the symptomatic diagnosis of cancer: the view from primary care. *Lancet Oncol* 2020;21:748–50.
- Maringe C, Spicer J, Morris M, *et al*. The impact of the COVID-19 pandemic on cancer deaths due to delays in diagnosis in England, UK: a national, population-based, modelling study. *Lancet Oncol* 2020;21:1023–34.

- 4 Sud A, Torr B, Jones ME, *et al.* Effect of delays in the 2-week-wait cancer referral pathway during the COVID-19 pandemic on cancer survival in the UK: a modelling study. *Lancet Oncol* 2020;21:1035–44.
- 5 Robilotti EV, Babady NE, Mead PA, *et al.* Determinants of COVID-19 disease severity in patients with cancer. *Nat Med* 2020;26:1218–23.
- 6 Zhang L, Zhu F, Xie L, *et al.* Clinical characteristics of COVID-19-infected cancer patients: a retrospective case study in three hospitals within Wuhan, China. *Ann Oncol* 2020;31:894–901.
- 7 Yang K, Sheng Y, Huang C, *et al.* Clinical characteristics, outcomes, and risk factors for mortality in patients with cancer and COVID-19 in Hubei, China: a multicentre, retrospective, cohort study. *Lancet Oncol* 2020;21:904–13.
- 8 Lee LY, Cazier JB, Starkey T. COVID-19 mortality in patients with cancer on chemotherapy or other anticancer treatments: a prospective cohort study. *The Lancet* 2020.
- 9 Jee J, Foote MB, Lumish M, *et al.* Chemotherapy and COVID-19 outcomes in patients with cancer. *J Clin Oncol* 2020;38:3538–46.
- 10 He W, Chen L, Chen L. COVID-19 in persons with haematological cancers. *Leukemia* 2020;1–9.
- 11 Yu J, Ouyang W, Chua MLK, *et al.* SARS-CoV-2 transmission in patients with cancer at a tertiary care hospital in Wuhan, China. *JAMA Oncology* 2020;6:1108.
- 12 Liang W, Guan W, Chen R, *et al.* Cancer patients in SARS-CoV-2 infection: a nationwide analysis in China. *Lancet Oncol* 2020;21:335–7.
- 13 Williamson EJ, Walker AJ, Bhaskaran K, *et al.* Factors associated with COVID-19-related death using OpenSAFELY. *Nature* 2020;584:430–6.
- 14 Desai A, Sachdeva S, Parekh T, *et al.* COVID-19 and cancer: lessons from a pooled meta-analysis. *JCO Glob Oncol* 2020;6:557–9.
- 15 Passamonti F, Cattaneo C, Arcaini L, *et al.* Clinical characteristics and risk factors associated with COVID-19 severity in patients with haematological malignancies in Italy: a retrospective, multicentre, cohort study. *Lancet Haematol* 2020;7:e737–45.
- 16 Natale F, Ghio D, Tarchi D. COVID-19 cases and case fatality rate by age. *European Commission* 2020;52:154–64.
- 17 Ghisolfi S, Almás I, Sandefur JC, *et al.* Predicted COVID-19 fatality rates based on age, sex, comorbidities and health system capacity. *BMJ Glob Health* 2020;5:e003094.
- 18 Dehingia N, Raj A. Sex differences in COVID-19 case fatality: do we know enough? *Lancet Glob Health* 2021;9:e14–15.
- 19 Mayo Clinic Proceedings. *Sex differences in case fatality rate of COVID-19: insights from a multinational registry.* Elsevier, 2020.
- 20 Sudlow C, Gallacher J, Allen N, *et al.* Uk Biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med* 2015;12:e1001779.
- 21 Khanji MY, Aung N, Chahal CAA, *et al.* COVID-19 and the UK Biobank-Opportunities and challenges for research and collaboration with other large population studies. *Front Cardiovasc Med* 2020;7:156.
- 22 Collins R. What makes UK Biobank special? *Lancet* 2012;379:1173–4.
- 23 Chadeau-Hyam M, Bodinier B, Elliott J, *et al.* Risk factors for positive and negative COVID-19 tests: a cautious and in-depth analysis of UK Biobank data. *Int J Epidemiol* 2020;49:1454–67.
- 24 Therneau TM, Grambsch PM. *The COX model. modeling survival data: extending the COX model.* Springer, 2000: 39–77.
- 25 Therneau TM, Lumley T. Package 'survival'. *R Top Doc* 2015;128:28–33.
- 26 Zaki N, Alashwal H, Ibrahim S. Association of hypertension, diabetes, stroke, cancer, kidney disease, and high-cholesterol with COVID-19 disease severity and fatality: a systematic review. *Diabetes Metab Syndr* 2020;14:1133–42.
- 27 Hoffmann C, Wolf E. Older age groups and country-specific case fatality rates of COVID-19 in Europe, USA and Canada. *Infection* 2021;49:111–6.
- 28 Dee EC, Mahal BA, Arega MA, *et al.* Relative timing of radiotherapy and androgen deprivation for prostate cancer and implications for treatment during the COVID-19 pandemic. *JAMA Oncol* 2020;6:1630–2.
- 29 Minami CA, Kantor O, Weiss A, *et al.* Association between time to operation and pathologic stage in ductal carcinoma in situ and early-stage hormone receptor-positive breast cancer. *J Am Coll Surg* 2020;231:434–47.
- 30 Burki TK. Cancer guidelines during the COVID-19 pandemic. *Lancet Oncol* 2020;21:629–30.
- 31 Ribas A, Sengupta R, Locke T, *et al.* Priority COVID-19 vaccination for patients with cancer while vaccine supply is limited. *Cancer Discov* 2021;11:233–6.
- 32 Desai A, Gainor JF, Hegde A. COVID-19 vaccine guidance for patients with cancer participating in oncology clinical trials. *Nature Reviews Clinical Oncology* 2021.

© 2021 Author(s) (or their employer(s)) 2021. Re-use permitted under CC BY-NC. No commercial re-use. See rights and permissions. Published by BMJ. This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: <http://creativecommons.org/licenses/by-nc/4.0/>. Notwithstanding the ProQuest Terms and Conditions, you may use this content in accordance with the terms of the License.