BMJ Health & Care Informatics

Web-based eHealth Clinical Decision Support System as a tool for the treat-to-target management of patients with systemic lupus erythematosus: *development and initial usability evaluation*

Agner Russo Parra Sanchez ,[1] Max G Grimberg,[2] Myrthe Hanssen,[2] Moon Aben,[2] Elianne Jairth,[2] Prishent Dhoeme,[2] Michel W P Tsang-A-Sjoe,[1] Alexandre Voskuyl,[1] Hendrik Jan Jansen,[2] Ronald van Vollenhoven[1]

To cite: Parra Sanchez AR, Grimberg MG, Hanssen M, *et al*. Web-based eHealth Clinical Decision Support System as a tool for the treat-to-target management of patients with systemic lupus erythematosus: *development and initial usability evaluation*. *BMJ Health Care Inform* 2023;**30**:e100811. doi:10.1136/bmjhci-2023-100811

Received 24 May 2023
Accepted 05 September 2023

[1]Rheumatology and Clinical Immunology, Amsterdam UMC Locatie VUmc, Amsterdam, Netherlands
[2]Medical Informatics, University of Amsterdam Faculty of Medicine, Amsterdam, Noord-Holland, Netherlands

Correspondence to
Dr Agner Russo Parra Sanchez; a.r.parrasanchez@amsterdamumc.nl

ABSTRACT

Background Treat-to-target (T2T) is a therapeutic strategy currently being studied for its application in systemic lupus erythematosus (SLE). Patients and rheumatologists have little support in making the best treatment decision in the context of a T2T strategy, thus, the use of information technology for systematically processing data and supporting information and knowledge may improve routine decision-making practices, helping to deliver value-based care.

Objective To design and develop an online Clinical Decision Support Systems (CDSS) tool “SLE-T2T”, and test its usability for the implementation of a T2T strategy in the management of patients with SLE.

Methods A prototype of a CDSS was conceived as a web-based application with the task of generating appropriate treatment advice based on entered patients’ data. Once developed, a System Usability Score (SUS) questionnaire was implemented to test whether the eHealth tool was user-friendly, comprehensible, easy-to-deliver and workflow-oriented. Data from the participants’ comments were synthesised, and the elements in need for improvement were identified.

Results The beta version web-based system was developed based on the interim usability and acceptance evaluation. 7 participants completed the SUS survey. The median SUS score of SLE-T2T was 79 (scale 0 to 100), categorising the application as ‘good’ and indicating the need for minor improvements to the design.

Conclusions SLE-T2T is the first eHealth tool to be designed for the management of SLE patients in a T2T context. The SUS score and unstructured feedback showed high acceptance of this digital instrument for its future use in a clinical trial.

WHAT IS ALREADY KNOWN ON THIS TOPIC

⇒ The treat-to-target (T2T) strategy is being studied as a therapeutic approach for managing patients with systemic lupus erythematosus (SLE).

⇒ Information technology offers the potential to improve decision-making in clinical practice, delivering value-based care.

WHAT THIS STUDY ADDS

⇒ The study presents the design and development of the ‘SLE-T2T’ a web-based Clinical Decision Support System (CDSS)—the first eHealth tool tailored for managing patients with SLE within a T2T framework.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

⇒ The ‘SLE-T2T’ CDSS tool could influence policy discussions on incorporating information technology in rheumatology care and promoting evidence-based and patient-centric approaches in managing SLE.

⇒ This study may pave the way for further research and validation, encouraging the adoption of digital tools like ‘SLE-T2T’ in routine clinical practice.

INTRODUCTION

Systemic lupus erythematosus (SLE) is a chronic, multisystemic and complex autoimmune disease, characterised by multiple manifestations and affecting predominantly woman of childbearing age.[1 2] Even when receiving the best possible care, SLE may be associated with damage accrual due to disease activity, comorbidities and the side effects of therapy (in particular, glucocorticoids), which negatively impacts patients’ health-related quality of life (HRQoL).[3] The treatment in SLE should, thus, aim for controlling the symptoms and disease activity while minimising the side effects and drug toxicity, ensuring survival, preventing organ damage and optimising HRQoL.[4] Formulating such a treatment plan for SLE is challenging due to the heterogeneity in its clinical presentation, disease course and prognosis. Clinicians from different medical specialisations may be involved in the management of patients with SLE and need to handle a vast amount of information to make clinical decisions that are difficult to capture in a single instrument.[5] It has been postulated that to achieve this, a treat-to-target (T2T) strategy would be beneficial. The essence of such a strategy can be summarised as setting a therapeutic target, intervening, assessing whether the target has been met, and adjusting therapy if it has not.[4 6 7] While endorsed by experts on SLE, the T2T strategy has not been formally proven effective and appears to be implemented only to a limited degree by practitioners.

Electronic health (eHealth) and mobile health (mHealth) are becoming prominent components of healthcare and represent an innovative tool to support practitioners in clinical decision-making.[8] Computerised Clinical Decision Support Systems (CDSS) represents a type of eHealth tool that compiles great volume of available data and helps clinicians to sift through it effectively and reliably.[9] CDSS have also shown increasing adherence to clinical guidelines, which traditionally have been shown to be difficult to implement in practice, increasing confidence in making decisions and improving prescribing behaviour.[10]

Hence, we aimed to develop SLE-T2T, a CDSS web-based eHealth tool that could help physicians in their decision-making process, in the context of a T2T approach for patients with SLE. We also aimed to evaluate the feasibility and usability of the first prototype, determining whether the CDSS is user-friendly, comprehensible, easy-to-deliver and workflow-oriented.

METHODS

System design and development

The creation process of web-based applications is composed of three phases. For the first phase, the design, SLE-T2T was conceived with an 'user-centred design' and with a specific task: to generate appropriate treatment advices based on entered patients' data. A general sketch of the programme was made, and general consensus was achieved with regards to the desired functionalities. To develop the clinical decision support functionality, European League Against Rheumatism (EULAR) recommendations for the management of SLE and international evidence-based guidelines were reviewed.[11] A knowledge-based system was generated, capable of form rule statements from the data collected in the input, similar to first-order logic, knowledge-based systems capture the data inputted and create a rule according to the pre-established conditions in logical system.[12] For SLE-T2T, the rules made from literature and guidelines were organised in the form of 'IF/THEN' statements in a prespecified decision table. The input was categorised according to disease activity state into: (a) *remission, (b) mild disease activity and (c) moderate/severe disease activity*, measured by the clinical Systemic Lupus Erythematosus Disease Activity Index 2000 (cSLEDAI-2K) and Physician Global Assessment (PGA). Patient's medication was also taken into account, and categorised into: *(1) use of antimalarials (yes/no), (2) use or immunosuppressives (yes/no and duration) and (3) use of glucocorticoids (yes/no and dosage)*. The result from the input combining the diverse categories generates a rule, tied to a predesigned set of general recommendations, and shown in the system as an output, so the health professional can make a decision according to that result. Table 1 exemplifies one scenario of rule-based 'IF/THEN' statements in a portion of the prespecified decision table.

For the second phase, the development phase, in partnership with the Medical Informatics department of the University of Amsterdam, a beta version of SLE-T2T was developed using a free integrated development environment, and based on Javascript, HTML and CSS programming languages and framework, to be used in web browsers. There was an iterative process of development, with close cooperation between clinicians and developers of the application. After the development, the system was made available temporarily for the participants in the evaluation phase.

System evaluation

The third phase was the testing. During this phase, safety, validation and verification analyses were performed (*data not shown*) looking at Sommerville's dependable programming guidelines,[13] all aspects inherent to the development phase. The CDSS was also electronically tested to verify that recommendation results matched the prespecified decision tree. Once the beta version of SLE-T2T was ready, the system was tested in terms of usability, which refers to the effectiveness, efficiency and user satisfaction rating of a product in a specific environment by a specific user for a specific purpose.[14] A System Usability Score (SUS) survey [15 16] was chosen as the usability test tool, widely adopted in this type of products for usability evaluations given its simplicity and advantages: (1) short questionnaire, quick to answer; (2) versatile for the evaluation of websites, software, mobile devices and medical systems; (3) the final score is interpreted based on a well-established reference standard;[17] (4) is suitable even when applied to small samples (N<14) and (5) it has excellent reliability (0.85).[16 18] The SUS contains 10 questions based on the Likert five-point scale; questions 1, 3, 5, 7 and 9 are positive and questions 2, 4, 6, 8 and 10 are negative. The 10 questions are closely related and are employed for the comprehensive evaluation of a product. A higher SUS score indicates better product usability. Furthermore, the SUS was coupled with unstructured feedback about areas of improvement, collected from the participants using the 'think aloud' method.[17]

Participants' recruitment and data collection

The recruitment was based on a convenience sampling method, through invitations to researchers, clinicians

Table 1 Example of one scenario from the extended-entry decision table, where remission is the target and remission is not achieved

Rules					
Conditions (IF)		R1	R2	R3	R4
	SLEDAI-2K (*Applicable when LLDAS as target*)	-	-	-	-
	cSLEDAI-2K	≥1	≥1	≥1	0
	PGA	>0.5	>0.5	>0.5	≤0.5
	Antimalarials	Y	Y	N	Y
	Immunosuppressives	Y	Y	N	N
	Glucocorticoids (*prednisolone dose*)	≤ 5 mg/day	5–7.5 mg/day	≤ 5 mg/day	>7.5 mg/day
		⬇	⬇	⬇	⬇
Actions (THEN)	Consider adjusting the treatment to achieve the target.	X	X	X	
	Consider flare and adjusting the treatment if a SLEDAI score greater than or equal to 3 points and a greater than or equal to 1-point increase in PGA is observed from previous visit.		X		
	Maintain antimalarial dose, or consider increasing it, if the maximum dose has not been reached and if tolerated.	X	X		X
	Consider initiating antimalarials, unless contraindicated. Note: *HCQ is recommended for all patients with SLE to decrease the risk of flares. HCQ is also associated with other beneficial effects, such as thrombosis risk in anti-phospholipid syndrome, fetal outcome in pregnancy, fasting glucose and lipid profile*.			X	
	Consider increasing the dose of immunosuppressant, if maximum dose has not been reached; or consider switching to a different drug, including biologics.	X			
	Consider early initiation of immunosuppressive agents (including biologics) for better disease control and to limit glucocorticoid toxicity.		X	X	X
	Consider (temporary) increase of glucocorticoids for fast control. Consider pulse or high-dose steroids for organ-threatening disease activity.	X			
	Maintain the dose of GC or consider (temporary) increase of glucocorticoids for fast control. Consider pulse or high-dose steroids for organ threatening disease activity				
	Consider increasing the dose of glucocorticoids if the patient's condition so required, otherwise maintain the dose of GC, or decrease if possible, and add other treatment options		X	X	
	Other considerations: Continue non-pharmacological interventions: Enhance UV light protection. If indicated, keep vaccinations up to date ► Implement lifestyle changes to reduce CV cardiovascular risk factors (no smoking, body weight, blood pressure, lipids, fasting glucose, exercise). ► Consider topical agents for cutaneous manifestations	X	X	X	X
	Follow-up SLE disease activity in 3 months	X	X	X	
	Follow-up of SLE disease activity in 6 months				X

Remission is defined according to the 2021 DORIS definition:[24] Clinical SLEDAI=0, PGA <0.5 (0–3), Irrespective of serology, and the patient may be on antimalarial, low-dose glucocorticoids (prednisolone ≤5 mg/day) and/or stable immunosuppressives including biologics.
Other categories included: mild disease activity (SLEDAI=1 to 5 and PGA ≥0.5 to ≤1), moderate disease activity (SLEDAI=6 to 10 and PGA >1 to ≤2), high disease activity (SLEDAI=11 to 19 and PGA >2 to ≤3) and severe disease activity (SLEDAI=≥20).
GC, Glucocorticoids; HCQ, hydroxychloroquine; LLDAS, Lupus Low Disease Activity State; PGA, Physician Global Assessment; SLE, systemic lupus erythematosus; SLEDAI, Systemic Lupus Erythematosus Disease Activity Index 2000; UV, Ultraviolet.

and related healthcare personnel to participate in the evaluation of the SLE-T2T website as independent users (not related to the development of the website). Once the participants agreed to take part in the evaluation, consent was obtained and they were invited to a 20–30 min video call to navigate through the page selecting the appropriate options according to a hypothetical clinical case, while describing aloud their overall perception as users; this was followed by the completion of the SUS survey about their experiences with the website. The questionnaire was sent via email and completed via personal computers and mobile terminals.

Statistical analysis

For descriptive statistical analysis, basic information about the participants was collected, including gender, age, education and profession, followed by the calculation of the SUS scores for each of the participants, and the mean SUS score, as described by the author,[15 16] using SPSS V.25. Qualitative data were collected through unstructured feedback, was analysed by first, creating an individual list of problems identified by each participant, to then group the duplicate problems between individuals and categorise them in terms of *system strengths, anticipated barriers and design recommendations.*

System refinement

The SUS score and/or unstructured feedback from the participants in the evaluation phase will enable to identify the necessary elements in need for improvement in the beta version of the CDSS, based on these, a set of criteria for software revision will be defined and the software version will be modified accordingly to reach a final version for later implementation in a pilot study.

RESULTS

System overview

SLE-T2T web-based system was developed. The processing of the system takes place on the user's computer, and, since no data is stored, the architecture of this decision support system is essentially composed of: (1) an input scheme consisting in the diverse set of index and scores existing for the measurement of SLE disease activity (cSLEDAI-2K, SLEDAI-2K, PGA score) as well as the used medication; (2) a rule-based interface that collects and processes patients' data and (3) an output dashboard with the generated set of recommendations tailored for the patients' clinical state and aiming to reach a preestablished target of treatment, based on the T2T strategy. Figures 1 and 2 depict a comprehensive view of the system architecture.

System Usability Scores

A total of seven participants completed the SUS questionnaire for this research. The participants included rheumatologist specialised in the management of patients with SLE and clinical researchers in the field of rheumatology. The mean usability rating given by the participants

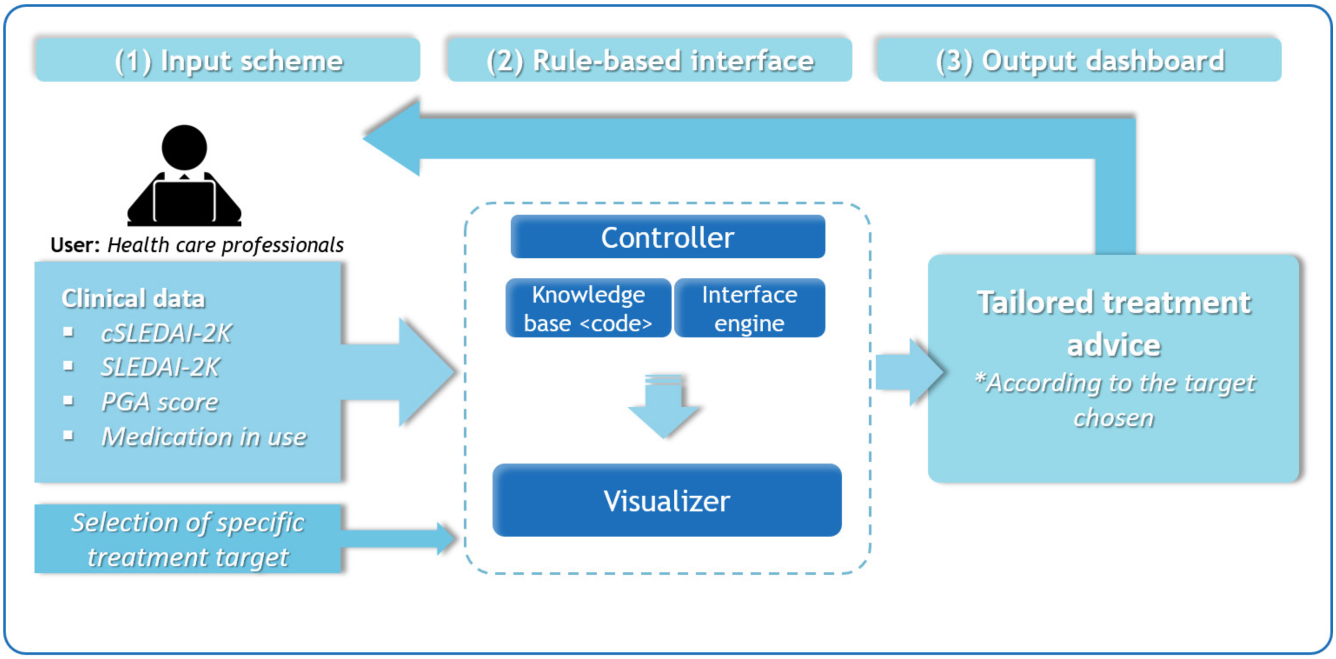

Figure 1 Overview of SLE-T2T CDSS tool architecture. cSLEDAI-2k, Clinical Systemic Lupus Erythematosus Disease Activity Index 2000; PGA: Physician Global Assessment; SLEDAI-2K, SystemicLupus Erythematosus Disease Activity Index 2000.

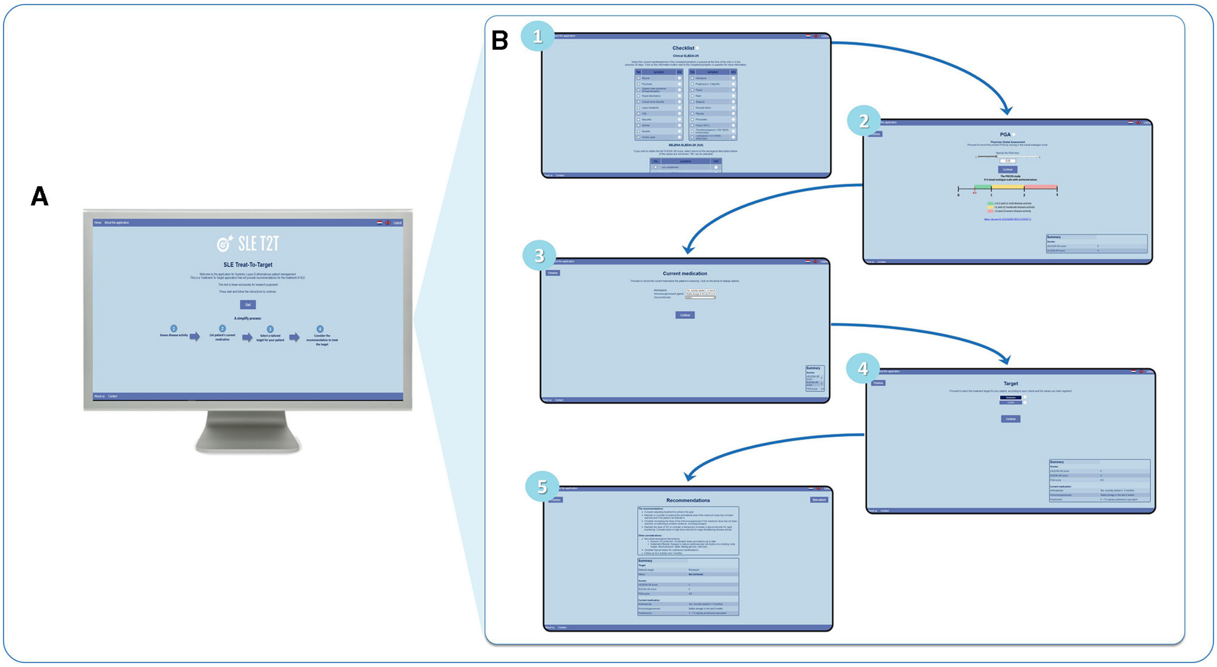

Figure 2 Desktop view screenshots of the SLE-T2T web-based application (Amsterdam UMC, all rights reserved). (A) Home page. (B) Sequence of screenshots following the evaluation process, as follow: 1. SLEDAI-2K checklist; 2. PGA visual scale from 0 to 3; 3. patient's current medication list, divided in antimalarial, immunosuppressive therapy (including biologics) and glucocorticoids (prednisolone dosage); 4. target selection page, among remission and LLDAS; 5. output page, describing the recommendations. LLDAS: Lupus Low Disease Activity State; PGA: Physician Global Assessment; SLEDAI-2K, Systemic Lupus Erythematosus Disease Activity Index 2000.

was 79, on a scale of 0 (worst) to 100 (best), categorising the application as 'good' (in the adjectives and acceptability categories associated with SUS scores), indicating the need for minor improvements to the design. Table 2 depicts the distribution of answers for the SUS rating.

Qualitative analysis

The qualitative data were obtained through unstructured feedback from the participants during the evaluation calls and their comments in the SUS form and classified the eHealth tool as practical and simple to use. In terms of the system strengths, participants perceived the web-based application as an advantage, simple and intelligible as exemplified below:

> I think the website is well-made and provides an easy to use SLEDAI-2K score form… for physicians who do not see patients with SLE that often, an easy to use SLEDAI-2K calculator and general treatment advices might be very useful (Clinician—Researcher in the field of SLE).
>
> I really like that the advice is (a little) personalised (Rheumatologist).
>
> Easy to use. It could save me some time in the daily practice…(Rheumatologist)

Some of the anticipated barriers were related to the migration of the data inputted and the advice generated to the electronic record environment:

> Overall easy to use. How to implement into EPIC? Would be great if we can see changes in scores in a figure in EPIC during follow up (Rheumatologist).

Based on this, a 'summary table' was added and can be seen as the user input data through the whole evaluation process. Once completed, it appears at the output screen, below the recommendations. This summary table can be easily copied into electronic records to keep track of the patient evaluation.

On the other hand, the participants identified the lack of patient opinion as a barrier to know the patient's preference when it comes to the target selection:

> It would be of great value to add PROMS/patient opinion about T2T to this project, as discussed (Rheumatologist)

In spite of this, SLE-T2T is intended for healthcare professionals as users, thus, including the collection of patient-reported outcome measures (PROMs) from the patients, at this stage, was not possible. We have suggested that during the clinical evaluation, the HCPs discuss together with the patient the selection of a treatment target. Based also on this comments, the record of PROMs manually will be included during the subsequent study, to further understand the patients' need in a T2T context.

Finally, in terms of design recommendations, most of the participants agreed that more visual aid will help to sift through the page easily.

Table 2 System usability average scores given by the participants and SUS final score

Number	Item	n = 7 Mean (SD)
1	I think that I would like to use this system frequently.	4 (0.78)
2	I found the system unnecessarily complex.	2 (0.53)
3	I thought the system was easy to use.	4 (0.53)
4	I think that I would need the support of a technical person to be able to use this system.	1 (0.37)
5	I found the various functions in this system were well integrated.	4 (0.48)
6	I thought there was too much inconsistency in this system.	2 (0.89)
7	I would imagine that most people would learn to use this system very quickly.	5 (0.53)
8	I found the system very cumbersome to use.	2 (0.48)
9	I felt very confident using the system.	4 (0.69)
10	I needed to learn a lot of things before I could get going with this system.	2 (0.89)
	SUS Score*	79.28

*The SUS score is computed by summing the score contributions from each item. Each item's score contribution ranges from 0 to 4. For statements Q1, Q3, Q5, Q7 and Q9 (phrased in a positive way), the score contribution is the scale position (from 1 to 5) minus 1. For statements Q2, Q4, Q6, Q8 and Q10 (phrased in a negative way), the contribution is 5 minus the scale position. Then, the sum of the scores is multiplied by 2.5 to obtain an overall system usability score ranging from 0 to 100.
SUS, System Usability Score.

> For Physician global assessment (PGA) scale would be helpful to indicate which side of the scale is good/bad in a more visual way. Make tables for remission and LLDAS goals next to each other so it is easier to compare what the differences are (Clinician—Researcher in the field of SLE).

In this sense, the graphical design of the SLEDAI-2K table and PGA visual scale were modified and made more eye catching, which translated into an easier way to navigate the site and fill in the required data.

The participants also reported some clarifications needed in the prototype web-based application, these in terms of grammatical typos, definition and specification of cut-off levels for some measurements, which were applied to the beta version of the e-health tool.

DISCUSSION

This study evaluated the performance and usability of 'SLE-T2T', a CDSS created to assist clinician in the management of patients with SLE in the context of a T2T strategy. Although well established in the software and development sector, usability testing is less commonplace within the healthcare context. Nonetheless, it has been gradually implemented in various areas where specific CDSS are developed for the improvement of clinical management. Schaaf *et al*[19] have carried out similar assessment process for a CDSS in the field of rare diseases. Using 'think-aloud' protocols in combination with SUS, testing the usability of CDSS, allowed them to reach system improvements in design, user interface and user experience (UX). More recently, in the field of rheumatology, Rheuma Care Manager (RCM)—a CDSS tool to support the management of rheumatoid arthritis applying T2T—was similarly evaluated in terms of accuracy, effectiveness, usability and acceptance.[20] RCM usability (SUS) was rated as good and was well accepted, showing that CDSS usage could support physicians by decreasing assessment deviations and increasing treatment decision confidence.[20]

In the context of SLE, eHealth technologies for the management of SLE are still a relatively new and unexplored topic, with potential for future investigation and development of such tools. Current eHealth tools for SLE are limited to educational tools, patient-reporting system, disease activity calculators and interactive online communities.[21] These have been described as of poor quality and limited functionality, and the literature examining this area is scarce.[21]

Our development and first evaluation process of a CDSS for T2T in SLE involved a small number of users who were used to paper-based indices to measure disease activity state in SLE. Conventions of usability testing support our small sample,[22] and the overall testing process was highly beneficial to the design and development for several reasons: participants had a wide age range and experience in secondary and tertiary care, and since the testing occurred early in development, it allowed us to identify the needed changes in design elements to arrive to a final version of the web-based application. The qualitative 'think aloud' method provided us with specific data and suggestions that we were able to integrate to improve the tool, especially related to UX and technical aspects.

Although there is a growing need and desire for eHealth technologies, the availability of apps designed specifically for SLE and the evidence for their efficacy are still limited. Accelerating the shift from traditional healthcare models to digital solutions remains a challenge faced by patients, their physicians and healthcare systems.[23] SLE-T2T CDSS represents a first step to tackle this unmet need. In the future, comprehensive multidisciplinary partnerships between clinical researchers, patients and app developers are critical to continue shifting digital health.

CONCLUSION

SLE-T2T CDSS is the first eHealth tool to be designed for the management of patients with SLE in a T2T context. The SUS score and unstructured feedback showed high acceptance of this digital instrument, and clinicians strongly supported the implementation of this kind of eHealth tools in the outpatient care setting. A CDSS specifically designed to support the T2T strategy in SLE appears to be both needed and likely to come with significant benefits. The final version reached after the improvements identified through the participants will be used for implementation in a larger T2T pilot study.

Contributors ARPS and RvV conceived the presented idea. ARPS developed the theory and proposal. RvV, AV and MWPT-A-S verified the methods and feasibility, and supervised the findings of this work. MGG, MA, PD, MH, and EJ carried out the development phase and conceived the mhealth tool. HJJ encouraged and supervised MGG, MA, PD, MH and EJ work. ARPS performed the experimental testing and wrote the manuscript with support from MWPT-A-S, RvV, AV and MGG. ARPS is responsible for the overall content as the guarantor.

Funding The work of ARPS is supported by the European Union's Horizon 2020 research and innovation program ('ARCAID'; www.arcaid-h2020.eu; grant agreement number 847551). This project was supported by a research grant from AstraZeneca.

Competing interests RV declares that he has received research support (institutional grants) from BMS, GSK, Lilly and UCB and support for educational programs from Pfizer and Roche. RV declares that he has also received consulting fees from AbbVie, AstraZeneca, Biogen, Biotest, BMS, Galapagos, Gilead, Janssen, Pfizer, Sanofi, Servier, UCB and Vielabio and personal honoraria as a speaker from AbbVie, Galapagos, GSK, Janssen, Pfizer and UCB. AV declares that he has received research support (institutional grants) from GSK and UCB. AV declares that he has also received consulting fees from GSK, AstraZeneca, Roche and personal honoraria as a speaker from GSK.

Patient consent for publication Not applicable.

Ethics approval This study involves human participants and was approved by Medisch Ethische Toetsingscommissie AMCMETC number 2022.0449. Participants gave informed consent to participate in the study before taking part.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data sharing not applicable as no datasets generated and/or analysed for this study. All data relevant to the study are included in the article or uploaded as supplementary information.

ORCID iD

Agner Russo Parra Sanchez http://orcid.org/0000-0002-9553-0447

REFERENCES

1 Dörner T, Furie R. Novel paradigms in systemic lupus erythematosus. *Lancet* 2019;393:2344–58.

2 Narváez J. Systemic lupus erythematosus 2020. *Med Clin (Barc)* 2020;155:494–501.

3 Bruce IN, O'Keeffe AG, Farewell V, *et al*. Factors associated with damage accrual in patients with systemic lupus erythematosus: results from the systemic lupus international collaborating clinics (SLICC) inception cohort. *Ann Rheum Dis* 2015;74:1706–13.

4 Parra Sánchez AR, Voskuyl AE, van Vollenhoven RF. Treat-to-target in systemic lupus erythematosus: advancing towards its implementation. *Nat Rev Rheumatol* 2022;18:146–57.

5 Tunnicliffe D, Singh-Grewal D, Craig J, *et al*. 384 Multi-specialists' perspectives on clinical decision making in systemic lupus erythematosus: an interview study. LUPUS 2017 & ACA 2017, (12th International Congress on SLE &, 7th Asian Congress on Autoimmunity); March 2017:A174–7

6 Ugarte-Gil MF, Burgos PI, Alarcón GS. Treat to target in systemic lupus erythematosus: a commentary. *Clin Rheumatol* 2016;35:1903–7.

7 van Vollenhoven RF, Mosca M, Bertsias G, *et al*. Treat-to-target in systemic lupus erythematosus: recommendations from an international task force. *Ann Rheum Dis* 2014;73:958–67.

8 Muhiyaddin R, Abd-Alrazaq AA, Househ M, *et al*. The impact of clinical decision support systems (CDSS) on physicians: a scoping review. *Stud Health Technol Inform* 2020;272:470–3.

9 Sutton RT, Pincock D, Baumgart DC, *et al*. An overview of clinical decision support systems: benefits, risks, and strategies for success. *NPJ Digit Med* 2020;3:17.

10 Lugtenberg M, Zegers-van Schaick JM, Westert GP, *et al*. Why don't physicians adhere to guideline recommendations in practice? An analysis of barriers among Dutch general practitioners. *Implement Sci* 2009;4:54.

11 Fanouriakis A, Kostopoulou M, Alunno A, *et al*. Update of the EULAR recommendations for the management of systemic lupus erythematosus. *Ann Rheum Dis* 2019;78:736–45.

12 Wasylewicz ATM. Scheepers-Hoeks A: clinical decision support systems. In: Kubben P, Dumontier M, DekkerA, eds. *Fundamentals of clinical data science*. Cham, CH, 2019: 153–69.

13 Sommerville I. *Software engineering*. Pearson, 2016.

14 Svanaes D, Das A, Alsos OA. The contextual nature of usability and its relevance to medical Informatics. *Stud Health Technol Inform* 2008;136:541–6.

15 Brooke J. *SUS: a quick and dirty usability scale*. Redhatch Consulting Ltd, 1995: 189.

16 Lewis J. *Usability testing*. 2006: 1275–316.

17 van Waes L. Thinking aloud as a method for testing the usability of websites: the influence of task variation on the evaluation of hypertext. *IEEE Trans Profess Commun* 2000;43:279–91.

18 Bangor A, Kortum PT, Miller JT. An empirical evaluation of the system usability scale. *International Journal of Human-Computer Interaction* 2008;24:574–94.

19 Schaaf J, Sedlmayr M, Sedlmayr B, *et al*. Evaluation of a clinical decision support system for rare diseases: a qualitative study. *BMC Med Inform Decis Mak* 2021;21:65.

20 Labinsky H, Ukalovic D, Hartmann F, *et al*. An AI-powered clinical decision support system to predict flares in rheumatoid arthritis: a pilot study. *Diagnostics (Basel)* 2023;13:148.

21 Dantas LO, Weber S, Osani MC, *et al*. Mobile health technologies for the management of systemic lupus erythematosus: a systematic review. *Lupus* 2020;29:144–56.

22 Rubin J, Chisnell D, Spool J. *Handbook of usability testing: how to plan, design, and conduct effective test*. Wiley, 2011.

23 Bergier H, Duron L, Sordet C, *et al*. Digital health, big data and smart technologies for the care of patients with systemic autoimmune diseases: where do we stand? *Autoimmun Rev* 2021;20:102864.

24 van Vollenhoven R, Bertsias G, Doria A, *et al*. 2021 DORIS definition of remission in SLE: final recommendations from an international task force. *Ann Rheum Dis* 2021;80:181.

BMJ Health & Care Informatics

Long short-term memory model identifies ARDS and in-hospital mortality in both non-COVID-19 and COVID-19 cohort

Jen-Ting Chen ,[1,2] Rahil Mehrizi,[3] Boudewijn Aasman,[4] Michelle Ng Gong,[2] Parsa Mirhaji[5]

To cite: Chen J-T, Mehrizi R, Aasman B, *et al*. Long short-term memory model identifies ARDS and in-hospital mortality in both non-COVID-19 and COVID-19 cohort. *BMJ Health Care Inform* 2023;**30**:e100782. doi:10.1136/bmjhci-2023-100782

► Additional supplemental material is published online only. To view, please visit the journal online (http://dx.doi.org/10.1136/bmjhci-2023-100782).

Received 19 April 2023
Accepted 21 August 2023

[1]Department of Medicine, Division of Pulmonary and Critical Care Medicine, UCSF, San Francisco, California, USA
[2]Department of Medicine, Division of Critical Care Medicine, Montefiore Medical Center, Bronx, New York, USA
[3]Department of Medicine, Albert Einstein College of Medicine, Bronx, New York, USA
[4]Center for Health Data Innovations, Montefiore Medical Center, Albert Einstein College of Medicine, Bronx, New York, USA
[5]Department of Genetics, Albert Einstein College of Medicine, Bronx, New York, USA

Correspondence to
Dr Jen-Ting Chen;
tina.chen@ucsf.edu

ABSTRACT

Objective To identify the risk of acute respiratory distress syndrome (ARDS) and in-hospital mortality using long short-term memory (LSTM) framework in a mechanically ventilated (MV) non-COVID-19 cohort and a COVID-19 cohort.

Methods We included MV ICU patients between 2017 and 2018 and reviewed patient records for ARDS and death. Using active learning, we enriched this cohort with MV patients from 2016 to 2019 (MV non-COVID-19, n=3905). We collected a second validation cohort of hospitalised patients with COVID-19 in 2020 (COVID+, n=5672). We trained an LSTM model using 132 structured features on the MV non-COVID-19 training cohort and validated on the MV non-COVID-19 validation and COVID-19 cohorts.

Results Applying LSTM (model score 0.9) on the MV non-COVID-19 validation cohort had a sensitivity of 86% and specificity of 57%. The model identified the risk of ARDS 10 hours before ARDS and 9.4 days before death. The sensitivity (70%) and specificity (84%) of the model on the COVID-19 cohort are lower than MV non-COVID-19 cohort. For the COVID-19 + cohort and MV COVID-19 + patients, the model identified the risk of in-hospital mortality 2.4 days and 1.54 days before death, respectively.

Discussion Our LSTM algorithm accurately and timely identified the risk of ARDS or death in MV non-COVID-19 and COVID+ patients. By alerting the risk of ARDS or death, we can improve the implementation of evidence-based ARDS management and facilitate goals-of-care discussions in high-risk patients.

Conclusion Using the LSTM algorithm in hospitalised patients identifies the risk of ARDS or death.

WHAT IS ALREADY KNOWN ON THIS TOPIC

⇒ Acute respiratory distress syndrome (ARDS) is commonly under-recognised in clinical settings, which can lead to delays in evidence-based management.

WHAT THIS STUDY ADDS

⇒ A long short-term memory algorithm trained on mechanically ventilated patients can identify the risk of ARDS development or in-hospital mortality using structured electronic health record data without the use of chest X-ray analysis. SARS-CoV-2 infection has a noted high incidence of ARDS. The model, trained on mechanically ventilated non-COVID-19 patients, performed well on COVID-19 patients, with an evaluation of 1.82 patients needed to identify 1 patient at risk of ARDS or death in the hospital.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

⇒ Being able to identify the risk of ARDS, regardless of COVID-19 status, early can improve compliance with evidence-based management and allow prognostication.

INTRODUCTION

Acute respiratory distress syndrome (ARDS) affects nearly a quarter of all acute respiratory failure patients requiring mechanical ventilation. It contributes to high morbidity and mortality of critically ill patients.[1] ARDS is consistently under-recognised, leading to delays in implementing evidence-based best practices, such as the use of lung-protective ventilation strategies.[2 3] The onset of the COVID-19 pandemic overwhelmed the healthcare system in the USA, and patients with severe to critical SARS-CoV-2 infections had a high incidence of ARDS and high mortality. This was especially true early in the pandemic, before the discovery of using early steroids and other immunosuppressants for treatment.[4 5] An electronic health record (EHR)-based decision support system that accurately identifies patients with ARDS can improve the management and escalation of these critically ill patients.[6] Different machine learning techniques, such as L2-logistic regression, artificial neural networks and XGBoost gradient boosted tree models, have leveraged EHR to identify or predict ARDS, yielding robust statistical discrimination as reported in studies.[7–9] In a distinct study, Zeiberg *et al* applied L2-regularised logistic regression to structured EHR data sourced from a single-centre population within the initial 7 days of hospitalisation. A meticulous two-physician chart review established the gold standard diagnosis of ARDS. Despite the rarity of ARDS occurrences (2.5%) within the testing cohort of this investigation, the area under the receiver operating curve (AUROC) attained an impressive value of 0.81.[7] Other investigations centred on using the Medical Information Mart for the ICU databases.[10 11] These endeavours relied on diverse data sources such as free-text entries, diagnostic codes and radiographic reports for both the diagnosis and prediction of ARDS.[10 11]

We aimed to train a deep learning model using long short-term memory (LSTM) framework and active learning method using a historic dataset from a mechanically ventilated (MV) non-COVID-19 cohort to identify patients with risk of ARDS or in-hospital mortality. We validated the model on an MV non-COVID-19 cohort, a COVID+ cohort and a subgroup of MV COVID+ cohort.

MATERIALS AND METHODS

The study was conducted at Montefiore Medical Center, encompassing three hospital sites.

COHORT ASSEMBLY

MV non-COVID-19 cohorts

Non-COVID-19 cohort 1 was constructed between 1 January 2017 and 31 August 2018 (figure 1). We included MV adults in the ICU with ages greater than 18. Each patient's chart was reviewed for ARDS.

Ground truth labelling: ARDS gold-standard identification

We defined ARDS using the Berlin criteria: hypoxaemia (arterial oxygen tension (PaO2) to fractional inspired oxygen (FiO2) ratio (PFR)≤300 with positive pressure ventilation ≥5cm$H_2$0), bilateral infiltrates on chest radiographs by independent review and a presence of ARDS risk factors (sepsis, shock, pancreatitis, aspiration, pneumonia, drug overdose and trauma/burn) not solely due to heart failure.[12] We used the first date and time of PFR≤300 with confirmed bilateral infiltrates within 24 hours as the time of ARDS presentation (ToP of ARDS).

Active learning

We used the 'active learning' technique to provide additional adult MV patients from July 2016 to December 2016 and September 2018 to December 2019 (AL-cohort).[13] A preliminary recurrent neural network was developed using the LSTM model and trained with the original non-COVID-19 cohort 1. Next, we applied the preliminary model to the AL-cohort. We used pool-based sampling and uncertainty techniques to identify records from AL-cohort to be reviewed and labelled by clinicians.[13] The uncertainty technique includes patients whose scores are very close to the cut-off, which means the model is least confident about them. We chose a cut-off of 0.80 and selected all records with a score between 0.75 and 0.85. We created the MV non-COVID-19 cohort 2 using the top 1% of the highest, lowest 1% and medium scores of the AL cohort. This allowed us to enrich MV non-COVID-19 cohort 2 with patients with ARDS or those who died in the hospital.

COVID-19 validation cohort

We included all hospitalised adult patients with and without mechanical ventilation with a positive SARS-Cov-2 transcription-mediated amplification assay from 1 March 2020 to 17 April 2020 in the COVID-19 cohort.

Training and validation cohort splitting

MV non-COVID-19 cohorts 1 and 2 were combined as the MV non-COVID-19 cohort. We randomly selected 80% of patients for training (MV non-COVID training cohort) and validation to learn model parameters and find optimal hyperparameters. The trained model was validated on the remaining 20% of the non-COVID-19 cohort (MV non-COVID-19 validation cohort), the COVID-19 cohort and the MV COVID-19 cohort separately (figure 1).

EHR DATA COLLECTION AND PROCESSING

Clinical data were collected through automated abstraction of EHR data. Raw EHR data for each admission were abstracted, sampled and validated (online supplemental table 2).

Sampling

Raw longitudinal EHR data were sampled every hour. Sampling was necessary since the different variables were recorded at different timestamps with different frequencies to aggregate the longitudinal data into hourly snapshots. If the data were recorded multiple times within 1 hour, we computed the minimum and maximum based on all recorded measurements. If it was not recorded at all within the 1-hour time frame, we considered it as 'missing'. For data that were recorded exactly once during an hour, the minimum and maximum would be the same.

Data validation

Data validation was performed by range checking (online supplemental table 2). If the recorded measure was outside the valid range, we discarded it and treated it as a missing value.

Missing data

The missing data were handled by 'forward imputing', where the most recent value fills the missing value. If there were no data available for imputation, we used normal values. We used the lower bound of the normal range as the minimum and the upper bound as the maximum value for those timestamps. A feature vector of size 132 represents each timestamp.

MODEL TRAINING

LSTM network is a paradigm of recurrent neural networks that can capture the temporal information

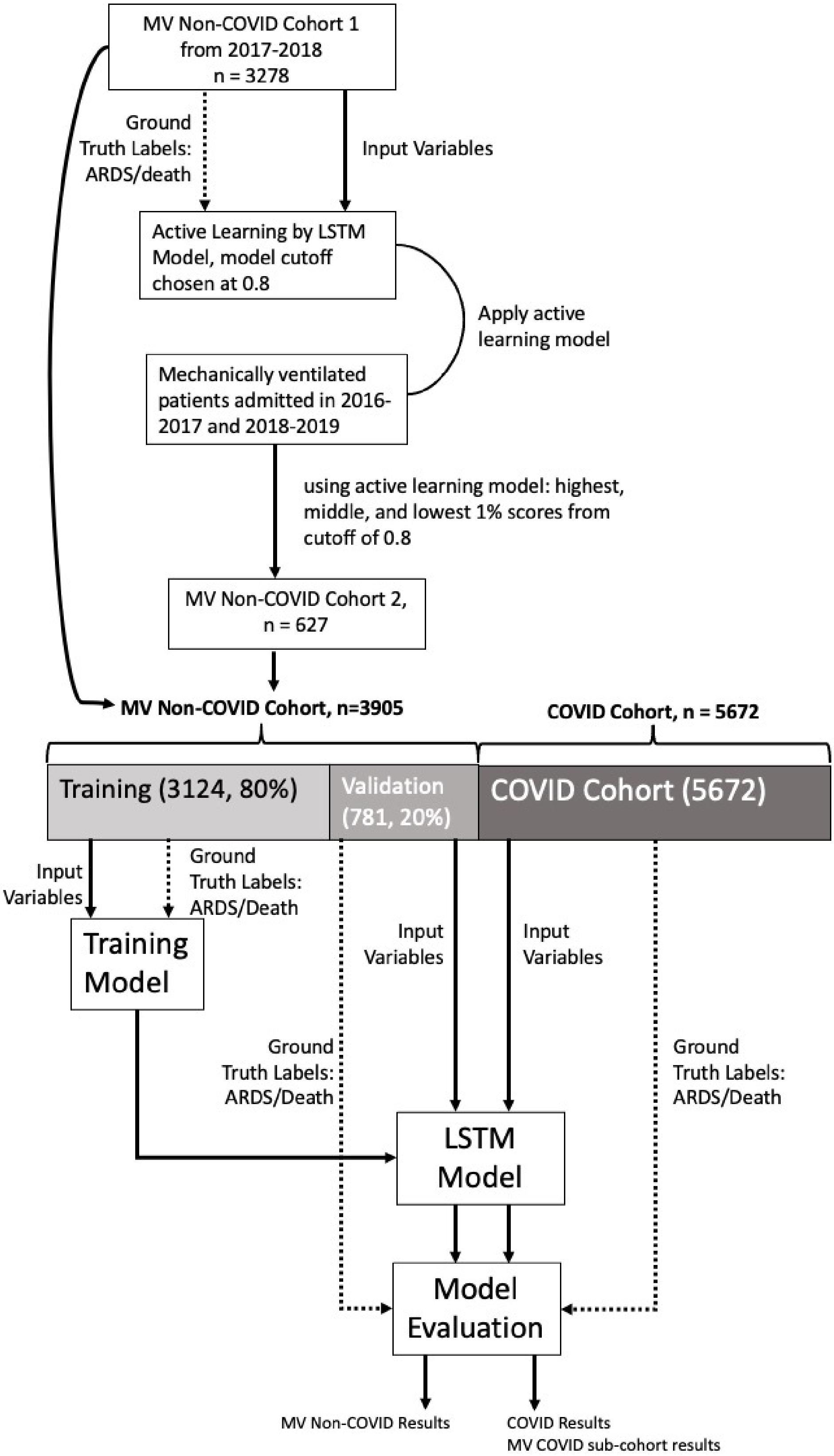
MV Non-COVID Cohort 1 from 2017-2018 n = 3278
Ground Truth Labels: ARDS/death
Input Variables
Active Learning by LSTM Model, model cutoff chosen at 0.8
Apply active learning model
Mechanically ventilated patients admitted in 2016-2017 and 2018-2019
using active learning model: highest, middle, and lowest 1% scores from cutoff of 0.8
MV Non-COVID Cohort 2, n = 627
MV Non-COVID Cohort, n=3905
COVID Cohort, n = 5672
Training (3124, 80%)
Validation (781, 20%)
COVID Cohort (5672)
Input Variables
Ground Truth Labels: ARDS/Death
Training Model
Input Variables
Input Variables
Ground Truth Labels: ARDS/Death
Ground Truth Labels: ARDS/Death
LSTM Model
Model Evaluation
MV Non-COVID Results
COVID Results
MV COVID sub-cohort results

Figure 1 Cohort assembly and model training. ARDS, acute respiratory distress syndrome; LSTM, long short-term memory; MV, mechanically ventilated.

of sequential data.[14] We used the EHR data, including the previous 12 hours, as the network inputs to train a model that can generate a predictive score for every patient at every hour. The network consisted of an LSTM unit with 10 filters, followed by a drop-out layer with 50% probability of keeping.[15] The network ended

Table 1 Cohorts characteristics

		Training	Validation		
Variables	MV non-COVID-19 cohort	MV non-COVID-19 (training) cohort	Non-COVID-19 (validation) cohort	COVID-19 cohort	MV COVID-19 subcohort
n	3905	3124	781	5672	803
Age, year, mean±SD	65.0±14.7	65.0±14.8	65.3±14.4	60.80±17.2	62.1±13.9
Gender					
Male, n (%)	1741 (44.6)	1437 (46)	328 (42)	2665 (47)	319 (40)
Female, n (%)	2164 (55.4)	1686 (54)	452 (58)	3006 (53)	484 (60)
Race or ethnicity					
White, n (%)	1015 (26)	749 (24)	249 (32)	623 (11)	177 (22)
Black, n (%)	1718 (44)	1405 (45)	320 (41)	2495 (44)	345 (43)
Other, n (%)	1171 (30)	968 (31)	210 (27)	2552 (45)	281 (35)
ARDS determination					
PaO2/FiO2 ratio ≤300, n (%)	3211 (82.2)	2579 (82.6)	632 (80.9)	617 (10.9)	617 (77)
CXR interpretation					
Yes (consistent with ARDS), n (%)	1333 (34.1)	35.4 (35.4)	260 (33.3)	565 (10)	565 (82)
Indeterminant, n (%)	313 (8.0)	7.1 (7.1)	60 (7.7)	18 (.3)	18 (2.2)
No (not consistent with ARDS), n (%)	2259 (57.8)	57.6 (57.6)	461 (59)	34 (.6)	34 (4.2)
Risk factors					
Aspiration, n (%)	407 (10.4)	10.3 (10.3)	86 (11)		
Shock, n (%)	1520 (38.9)	39.2 (39.2)	299 (38.3)		
Pneumonia, n (%)	1530 (39.2)	39.8 (39.8)	288 (36.9)	5672 (100)	803 (100)
Sepsis, n (%)	1885 (48.3)	48.8 (48.8)	362 (46.4)		
Pancreatitis, n (%)	42 (1.1)	1.1 (1.1)	9 (1.2)		
Burn, n (%)	3 (0.1)	3 (0.1)	0 (0)		
Overdose, n (%)	98 (2.5)	2.5 (2.5)	21 (2.7)		
Transfusion, n (%)	1191 (30.5)	30.7 (30.7)	232 (29.7)		
Congestive heart failure, n (%)	914 (23.4)	23.6 (23.6)	178 (22.8)		
Presence of at least one risk factor, n (%)	2739 (70.1)	70.6 (70.6)	362 (46.4)	5672 (100)	803 (100)
Clinical outcomes					
Mechanically ventilated, n (%)	3905	3124	781	803 (14.2)	803
ARDS, n (%)	1646 (42.2)	1326 (42.4)	320 (41.0)	583 (10.3)	583 (72.6)
In-hospital mortality, n (%)	1033 (26.5)	848 (27.1)	185 (23.7)	907 (16.0)	418 (52.1)
ARDS or in-hospital mortality, n (%)	2044 (52.3)	1655 (53.0)	389 (49.8)	1235 (21.9)	746 (92.9)

ARDS, acute respiratory distress syndrome; CXR, chest X-ray; FiO2, fractional inspired oxygen; MV, mechanically ventilated; PaO2, arterial oxygen tension.

with a linear layer and a Sigmoid activation function to output a score from 0 to 1, which is interpreted as the probability of developing ARDS or in-hospital mortality.

MODEL EVALUATION

We applied the model on the MV non-COVID-19 validation cohort and COVID-19 cohort hourly to produce the score for that timestamp which is an indication of the probability of ARDS development or death. For each cohort, we calculated the AUROC. We also calculated the sensitivity, specificity, positive predictive value (PPV), negative predictive value, and F1 score at different risk thresholds (cutoffs). We use the highest F1 score to generate a confusion matrix for selecting a score cut-off.

Table 2 Model diagnostics

TREAT-ECARDS model diagnostics	MV non-COVID-19 cohort	COVID-19 cohort	MV COVID-19 subcohort
Sensitivity	0.86	0.7	0.92
Specificity	0.57	0.84	0.23
Positive predictive value	0.66	0.55	0.94
Negative predictive value	0.8	0.91	0.17
Receiver operating curve	0.78	0.83	0.7
F1 score	0.75	0.61	0.93
No needed to evaluate	1.52	1.82	1.06

MV, mechanically ventilated.

The warning time is the first time the score exceeds the predefined cut-off. We continued running the test until the score exceeded the cut-off or discharge time. We evaluated model timeliness based on ARDS and death, ARDS and not death, no ARDS and death, no ARDS and not death and compared the actual ToP ARDS time/death time with the warning time.

FEATURE IMPORTANCE

Feature importance identifies a subset of features that are the most relevant for the accuracy of the model. We used local interpretable model-agnostic explanations (LIME),[16] to determine the importance of each variable to the accuracy of the model. The feature importance value was determined for 200 randomly sampled patients in each cohort using LIME, then calculated the average across all samples.

RESULTS

Cohort description

MV non-COVID-19 cohort 1 included 3278 patients (online supplemental table 1 and figure 1). MV Non-COVID-19 cohort 2 was derived from the active learning, consisting of 627 patients (online supplemental table 1). We combined MV Non-COVID-19 cohorts 1 and 2 to create the MV non-COVID-19 Cohort (n=3905, table 1). COVID-19 cohort included 5672 patients (table 1). Online supplemental table 3 shows the descriptive statistics of all variable fields in the MV non-COVID and COVID-19 cohorts.

MODEL DIAGNOSTICS

MV non-COVID-19 validation cohort

Based on the highest F1 score, we chose a model score cut-off at 0.90. The model diagnostics are presented in table 2, figure 2. The model warned of patient risk at a median of 10 hours (IQR −75 to 4) before ARDS and −225 hours or 9 days (IQR −461 to 101 hours) before death in the hospital (table 3). In ARDS survivors, the majority of the patients had ARDS risk identified before intubation and before ARDS diagnosis (table 3). For ARDS non-survivors, the model warned at 1 hour (IQR −38 to 9) before intubation, −20 hours (IQR −115 to 0.3) before ARDS and at −314 hours (IQR −589 to −128 hours) before death (table 3).

COVID-19 cohort and MV COVID-19 subcohort

Using the same cut-off of 0.9, we applied the model to COVID-19 and MV COVID-19 subcohorts. The model diagnostics are presented in table 2 and figure 2. When the model was applied to the COVID-19 cohort, the PPV was lower and more patients needed to be screened compared with the MV non-COVID-19 validation cohort. Whereas in the MV COVID -19 subcohort patients had a high prevalence of ARDS and in-hospital mortality, the PPV and number needed to evaluate were much lower than in the MV non-COVID-19 Validation Cohort.

In the COVID-19 cohort, the model warned the patient was likely to have ARDS or in-hospital mortality 3 hours after intubation and at ToP ARDS (table 3). Among the non-survivors, the model warned 2.4 days before in-hospital mortality (IQR 4.7–0.83) in COVID-19 patients, and 1.54 days before in-hospital mortality (IQR 3.6–0.46) in MV COVID-19 patients (table 3).

FEATURE IMPORTANCE

For both the MV non-COVID-19 and COVID-19 cohorts, we randomly selected 200 encounters from each cohort and performed LIME (online supplemental figure 1). The top contributors are similar in the MV non-COVID-19 and COVID-19 cohorts. The most important variable to the model was lactate level in discriminating the clinical outcome. The model consistently used lactate, age, cryoprecipitate transfusion, dopamine, bicarbonate level and epinephrine as important input variables (online supplemental figure 1).

DISCUSSION

From a cohort of pre-COVID-19 pandemic patients on mechanical ventilation, we developed and validated an LSTM model to identify patients at risk for ARDS or in-hospital mortality. This model was successfully integrated into EHR and identified patients at risk for ARDS

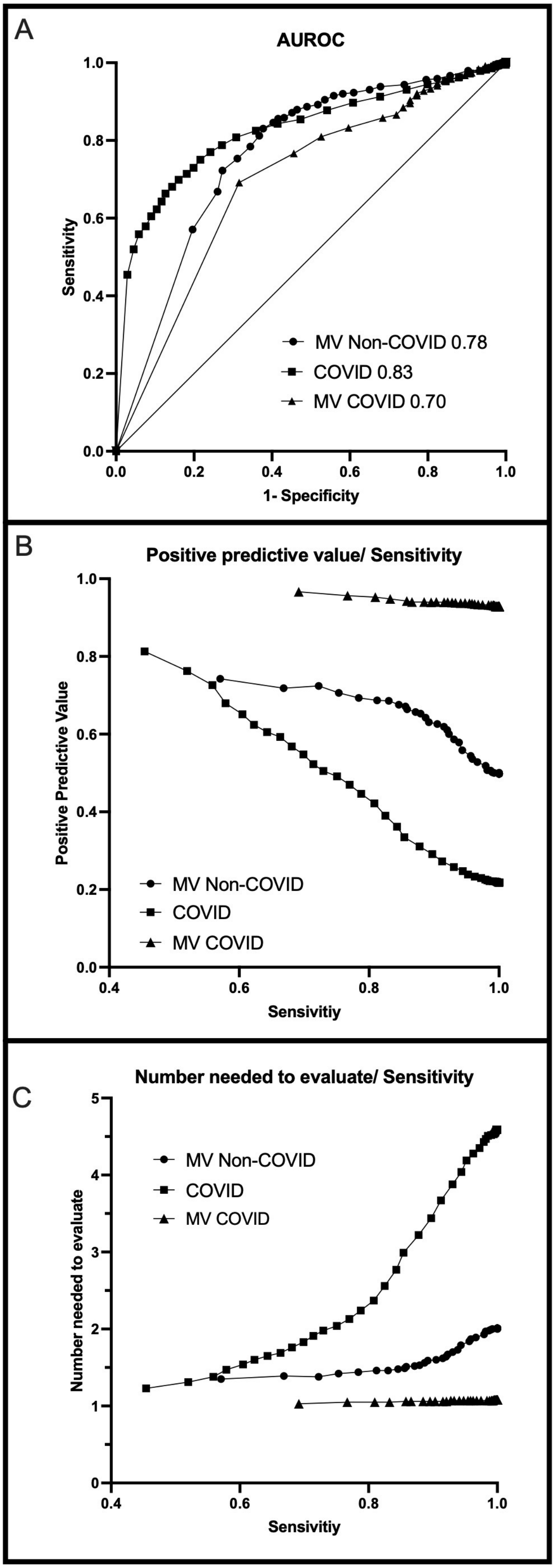

Figure 2 Model diagnostics, AUROC, PPV with sensitivity and NNE with sensitivity. AUROC, area under the receiver operating curve; MV, mechanically ventilated; NNE, number needed to evaluate; PPV, positive predictive value.

or in-hospital mortality in all adults hospitalised with and without COVID-19 infection, regardless of mechanical ventilation status. The model was also able to warn well before the events of ARDS or death in both the MV non-COVID-19 and COVID-19 cohorts. The timeliness of the model allows clinicians to modify management and implement evidence-based practices promptly.

This is the first utilisation of an LSTM network for identifying the risk of ARDS and in-hospital mortality. The LSTM is a recurrent neural network that uses feedback layers to capture temporal aspects such as sequences and trends. This approach is well suited for this study because past events and the progression of patient status are often valuable to determine the probability of ARDS or death. As in the reality of managing critically ill patients, physiological observations at each time point are taken into account. Their change and progression or regression inform the decisions at the subsequent processing of this information. This is well suited for dynamically changing situations to monitor and identify patients progressing to ARDS or in-hospital mortality. LSTM models have been used to predict heart failure, transfusion needs in the ICU, and mortality in the neonatal ICU, all with better predictive utility than traditional logistic regression models.[17–19] We chose to include ARDS diagnosis and in-hospital mortality as our patient-centred outcomes of interest instead of ARDS or in-hospital mortality alone, as in previous ARDS prediction studies.[6 7 20] Identifying the risk of ARDS or in-hospital mortality has shown real clinical implications when managing patients, mitigating the ambiguity that sometimes can exist in ARDS clinical diagnosis based on shifting diagnostic criteria.[7 8 20–22]

This cohort is one of the largest validated ARDS gold standards developed by manual chart review and active learning from a single centre. We did not rely on ICD-10 diagnosis codes or radiology reports to identify ARDS. Instead, we followed the Berlin criteria using PFR, independent review of chest X-ray for the presence of bilateral infiltrates and risk factors of ARDS in the patients' chart. Our model performed similarly to previously reported models using other machine learning methods, ranging from 0.71 to 0.90.[7 9–11 21] We forgo chest X-ray interpretation as input variables, as in Zeiberg *et al*.[7] Other large-scale ARDS identification studies which used natural language processing of radiology reports and diagnostic codes in clinical settings would delay ARDS recognition and rely heavily on clinician decisions.[9 11] Using chest radiographs for the diagnosis of ARDS has its limitations, as studies show high interobserver variabilities despite training.[12 23] In addition, radiology report turn-around times can range from 15 min to 26 hours, depending on the study location, availability of staff and hospital resources.[24 25] This reliance on chest radiograph interpretations may delay ARDS diagnosis.

Despite the different clinical characteristics of the study cohorts, being MV patients non-COVID-19 versus non-MV COVID-19 patients, important features in risk identification were broadly consistent between the cohorts using

Table 3 Timeliness of model

Cohort, (n)	Correctly identifies, n (%)	Time from intubation, median (IQR), hours	Before intubation, n (%)	After intubation, n (%)	Time from ARDS label, median (IQR), hours	Before ARDS, n (%)	After ARDS, n (%)	Time from death median (IQR), hours
MV non-COVID-19 cohort								
ARDS**,** (204)	166 (81.4)	0 (–12.8 to 26.0)	87 (52.4)	79 (47.6)	0 (–43.8 to 12.0)	115 (69.3)	51 (30.7)	
Death, (69)	60 (87)							225.5 (–461.3 to –101.3)
ARDS and death, (116)	108 (93.1)	-1 (–38.8 to 9.3)	68 (63.0)	40 (37.0)	20 (–115.5 to 0.3)	81 (75.0)	27 (25.0)	314 (–588.5 to –127.8)
ARDS or death, (389)	274 (70.4)	1 (–17.8 to 15.0)	155 (56.6)	119 (43.4)	10.0 (–75.5 to 4.0)	196 (71.5)	78 (28.5)	225.5 (–461.3 to –101.3)
No ARDS or death, (392)	223 (56.9)							
COVID-19 cohort								
ARDS, (328)	318 (97)	3 (–8.8 to 11.0)	136 (42.8)	182 (57.2)	0 (–16 to 9.0)	141 (44.3)	128 (40.3)	
Death, (652)	308 (47.2)							58 (–112 to –20)
ARDS and death, (255)	237 (92.9)	4 (–1 to 18)	86 (36.3)	156 (65.8)	0 (–12 to 10)	125 (52.7)	112 (47.3)	112 (–211.3 to –52)
ARDS or death, (1235)	555 (44.9)	3 (–3.5 to 13.0)	222 (40.0)	333 (60.0)	0 (–14.0, 10.0)	266 (47.9)	240 (43.2)	58 (–112 to –20)
No ARDS or death, (4437)	3724 (83.9)							
MV COVID-19 subcohort								
ARDS, (328)	318 (97)	3 (–8.8 to 11.0)	136 (42.8)	182 (57.2)	0 (–16.0 to 9.0)	141 (44.3)	128 (40.3)	
Death, (163)	128 (78.5)							37 (–87 to –11)
ARDS and death, (255)	237 (92.9)	4 (–1, 18)	86 (36.3)	156 (65.8)	0 (–12 to 10)	125 (52.7)	112 (47.3)	112 (–211. 3 to –52)
ARDS or death, (746)	555 (74.4)	3 (–3.5 to 13.0)	222 (40.0)	333 (60.0)	0 (–14.0 to 10.0)	266 (47.9)	240 (43.2)	37 (–87 to –11)
No ARDS or death, (57)	13 (22.8)							

ARDS, acute respiratory distress syndrome; MV, mechanically ventilated.

lactate, age, cryoprecipitate transfusion, dopamine, bicarbonate level and epinephrine as important input variables. LIME can directly associate model features to increased or decreased risk of ARDS or death in an individual, on a patient-by-patient-level.[26 27] We randomly sampled 200 patients in each cohort and obtained an average of the absolute LIME values to understand what features were generally used. This does not provide a clinical explanation and rationale for why features may relate to higher or lower scores. Instead, it sheds light on important features that the model needs as its input data to predict a score accurately, whether additive or subtractive, to the risk. Norepinephrine was the most commonly used vasopressor for both cohorts; intriguingly, it did not contribute to the model consideration. The model rarely used vasopressors such as dopamine and epinephrine to discriminate the outcome of ARDS and/or in-hospital mortality. Oxygen support devices were also not deemed important on average; we postulate that our gold standard labelling required mechanical ventilation for ARDS identification, making oxygen support devices less important in the discrimination.

In clinical practice, ARDS is underdiagnosed, which leads to increased exposures in management that are detrimental to patients, such as high tidal volume ventilation and delayed implementation of evidence-based practices that are helpful.[2 3 28–31] We used continuous data at 1-hour intervals starting at hospital admission to identify the early risk of an adverse outcome. Indeed, in the non-COVID-19 cohort, we identified ARDS hours before intubation and at the time of ToP ARDS. The majority of patients (56.5%) had been identified before ARDS diagnosis in the MV non-COVID-19 cohort, and this remained the case in the COVID+ cohort (43%). Implemented and delivered as a clinical decision support system, the early recognition would allow clinicians to initiate treatment such as LTVV as early as possible, when it may more positively impact outcomes.[3]

Furthermore, the model identified the risk of in-hospital mortality 9 days in advance in the non-COVID-19 cohort and 2 days in advance in the COVID-19 cohort. This has significant implications for triaging patients during surge capacity. In the MV non-COVID-19 cohort, there was no concern for ventilator or ICU resource allocation. Early identification of risk for death would alert the clinician to implement aggressive management and allow the treating physician to consider early palliation intervention/conversation. In the setting of a high volume surge of respiratory illness, such as the onset of the COVID-19 pandemic, where the incidences of ARDS and death are high, identifying adverse outcomes days in advance could help the clinician in making necessary triage decisions for resource allocation.[32–34]

Our study has some limitations. First, our cohorts were constructed from a single centre in the Bronx, and the patients' characteristics may not be generalisable to other centres and populations. However, our medical centre consists of three hospitals ranging from community and academic to tertiary transplant centres, thus spanning a wide spectrum of disease severity. In addition, we validated the algorithm in the COVID-19 cohort regardless of the respiratory support type, demonstrating consistent model performance across different cohorts. Second, although we were able to determine feature importance using LIME on 200 samples from each cohort, we were unable to discern the actual direction of association with the risk of ARDS or death. We cannot discern if the individual variables increase or decrease the risk of ARDS or death, despite their importance to the overall model. However, the consistency in features used to determine risk between the validation cohorts is reassuring. Ultimately, the variables that we included in models are variables known to be clinically associated with ARDS or death; therefore, the direction of influence on risk assessment is less germane. The strength of our study lies in the predictive nature of this algorithm and the timeliness of its predictions. Using longitudinal data from admission allowed the LSTM model to learn from the progression of the patient's clinical status over time. This model also was flexible to have similar diagnostic performance in patients with different clinical characteristics.

In conclusion, our LSTM model identified risk for ARDS and in-hospital mortality on patients with or without COVID-19 regardless of mechanical ventilator support. The model identified patients early, which implies management changes can be implemented early.

Acknowledgements The authors thank Mr. Daniel Ceusters for his administrative assistance for this project.

Contributors J-TC: data collection and monitoring, data analysis and manuscript preparation. RM: data analysis and manuscript preparation. BBA: data collection and data analysis. MNG: idea and project generation, data analysis, manuscript review. PM: idea and project generation, data analysis, manuscript review. PM is the guarantor of the overall content.

Funding This study was supported by Agency for Healthcare Research and Quality (AHRQ), grant number R18HS026188 and National Insitute of Health/ National Center for Advancing Translational Science (NCATS) Einstein-Montefiore CTSA Grant Number 1UM1TR004400.

Competing interests None declared.

Patient consent for publication Not applicable.

Ethics approval This study was approved by Albert Einstein College of Medicine Internal Review Board.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data sharing not applicable as no datasets generated and/or analysed for this study. All data relevant to the study are included in the article or uploaded as supplementary information.

ORCID iD

Jen-Ting Chen http://orcid.org/0000-0002-8530-2520

REFERENCES

1 Cartin-Ceba R, Kojicic M, Li G, *et al*. Epidemiology of critical care syndromes, organ failures, and life-support interventions in a suburban US community. *Chest* 2011;140:1447–55.

2 Bellani G, Laffey JG, Pham T, *et al*. Epidemiology, patterns of care, and mortality for patients with acute respiratory distress syndrome in intensive care units in 50 countries. *JAMA* 2016;315:788–800.

3 Needham DM, Yang T, Dinglas VD, *et al*. Timing of low tidal volume ventilation and intensive care unit mortality in acute respiratory distress syndrome. A prospective cohort study. *Am J Respir Crit Care Med* 2015;191:177–85.

4 Berlin DA, Gulick RM, Martinez FJ. Severe COVID-19. *N Engl J Med* 2020;383:2451–60.

5 Richardson S, Hirsch JS, Narasimhan M, *et al*. Presenting characteristics, Comorbidities, and outcomes among 5700 patients hospitalized with COVID-19 in the New York City area. *JAMA* 2020;323:2052–9.

6 Wayne MT, Valley TS, Cooke CR, *et al*. “Electronic “Sniffer” systems to identify the acute respiratory distress syndrome”. *Ann Am Thorac Soc* 2019;16:488–95.

7 Zeiberg D, Prahlad T, Nallamothu BK, *et al*. Machine learning for patient risk stratification for acute respiratory distress syndrome. *PLoS One* 2019;14:e0214465.

8 Wong A-KI, Cheung PC, Kamaleswaran R, *et al*. Machine learning methods to predict acute respiratory failure and acute respiratory distress syndrome. *Front Big Data* 2020;3:579774.

9 Le S, Pellegrini E, Green-Saxena A, *et al*. Supervised machine learning for the early prediction of acute respiratory distress syndrome (ARDS). *J Crit Care* 2020;60:96–102.

10 Taoum A, Mourad-Chehade F, Amoud H. Early-warning of ARDS using novelty detection and data fusion. *Comput Biol Med* 2018;102:191–9.

11 Apostolova E, Uppal A, Galarraga JE, *et al*. Towards reliable ARDS clinical decision support: ARDS patient analytics with free-text and structured EMR data. *AMIA Annu Symp Proc* 2019;2019:228–37.

12 Sjoding MW, Hofer TP, Co I, *et al*. Interobserver reliability of the Berlin ARDS definition and strategies to improve the reliability of ARDS diagnosis. *Chest* 2018;153:361–7.

13 Settles B. Active learning literature survey. Computer Sciences Technical Report 1648. University of Wisconsin–Madison; 2009.

14 Hochreiter S, Schmidhuber J. Long short-term memory. *Neural Comput* 1997;9:1735–80.

15 Srivastava N, Hinton G, Krizhevsky A, *et al*. Dropout: A simple way to prevent neural networks from Overfitting. *J Mach Learn Res* 2014;15:1929–58.

16 Ribeiro MT, Singh S, Guestrin C. Why should I trust you?": explaining the predictions of any Classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; San Francisco, California, USA: Association for Computing Machinery, 2016:1135–44

17 Shung D, Huang J, Castro E, *et al*. Neural network predicts need for red blood cell transfusion for patients with acute gastrointestinal bleeding admitted to the intensive care unit. *Sci Rep* 2021;11:8827.

18 Hagan R, Gillan CJ, Spence I, *et al*. Comparing regression and neural network techniques for personalized predictive Analytics to promote lung protective ventilation in intensive care units. *Comput Biol Med* 2020;126:104030.

19 Maheshwari S, Agarwal A, Shukla A, *et al*. A comprehensive evaluation for the prediction of mortality in intensive care units with LSTM networks: patients with cardiovascular disease. *Biomed Tech (Berl)* 2020;65:435–46.

20 Ding X-F, Li J-B, Liang H-Y, *et al*. Predictive model for acute respiratory distress syndrome events in ICU patients in China using machine learning Algorithms: a secondary analysis of a cohort study. *J Transl Med* 2019;17:326.

21 Fei Y, Gao K, Li WQ, *et al*. Prediction and evaluation of the severity of acute respiratory distress syndrome following severe acute Pancreatitis using an artificial neural network algorithm model. *HPB (Oxford)* 2019;21:891–7.

22 Sinha P, Delucchi KL, McAuley DF, *et al*. Development and validation of parsimonious Algorithms to classify acute respiratory distress syndrome phenotypes: a secondary analysis of randomised controlled trials. *Lancet Respir Med* 2020;8:247–57.

23 Goddard SL, Rubenfeld GD, Manoharan V, *et al*. The randomized educational acute respiratory distress syndrome diagnosis study: A trial to improve the radiographic diagnosis of acute respiratory distress syndrome. *Crit Care Med* 2018;46:743–8.

24 Towbin AJ, Iyer SB, Brown J, *et al*. Practice policy and quality initiatives: decreasing variability in turnaround time for radiographic studies from the emergency Department. *Radiographics* 2013;33:361–71.

25 Chan KT, Carroll T, Linnau KF, *et al*. Expectations among academic Clinicians of inpatient imaging turnaround time: does it correlate with satisfaction *Acad Radiol* 2015;22:1449–56.

26 Ribeiro MT, Singh S, Guestrin C. Why should I trust you?": explaining the predictions of any Classifier. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD ’16); New York, NY, USA: Association for Computing Machinery, 2016

27 Elshawi R, Al-Mallah MH, Sakr S. On the Interpretability of machine learning-based model for predicting hypertension. *BMC Med Inform Decis Mak* 2019;19:146.

28 Weiss CH, Baker DW, Weiner S, *et al*. Low tidal volume ventilation use in acute respiratory distress syndrome. *Crit Care Med* 2016;44:1515–22.

29 Qadir N, Bartz RR, Cooter ML, *et al*. Variation in early management practices in moderate-to-severe ARDS in the United States: the severe ARDS: generating evidence study. *Chest* 2021;160:1304–15.

30 Duggal A, Rezoagli E, Pham T, *et al*. Patterns of use of Adjunctive therapies in patients with early moderate to severe ARDS: insights from the LUNG SAFE study. *Chest* 2020;157:1497–505.

31 Brower RG, Matthay MA, Morris A, *et al*. Ventilation with lower tidal volumes as compared with traditional tidal volumes for acute lung injury and the acute respiratory distress syndrome. *N Engl J Med* 2000;342:1301–8.

32 Laventhal N, Basak R, Dell ML, *et al*. The ethics of creating a resource allocation strategy during the COVID-19 pandemic. *Pediatrics* 2020;146:e20201243.

33 Emanuel EJ, Persad G, Upshur R, *et al*. Fair allocation of scarce medical resources in the time of COVID-19. *N Engl J Med* 2020;382:2049–55.

34 Aliberti MJR, Szlejf C, Avelino-Silva VI, *et al*. COVID-19 is not over and age is not enough: using frailty for prognostication in hospitalized patients. *J Am Geriatr Soc* 2021;69:1116–27. 10.1111/jgs.17146 Available: https://onlinelibrary.wiley.com/toc/15325415/69/5

BMJ Health & Care Informatics

Implementer report: ICD-10 code F44.5 review for functional seizure disorder

Sana F Ali ,[1,2,3] Yarden Bornovski,[4] Margaret Gopaul ,[1] Daniela Galluzzo,[4] Joseph Goulet,[2,3] Stephanie Argraves,[2,3] Ebony Jackson-Shaheed,[5] Kei-Hoi Cheung,[2,3] Cynthia A. Brandt,[2,3] Hamada Hamid Altalib[1,2,3]

To cite: Ali SF, Bornovski Y, Gopaul M, *et al.* Implementer report: ICD-10 code F44.5 review for functional seizure disorder. *BMJ Health Care Inform* 2023;**30**:e100746. doi:10.1136/bmjhci-2023-100746

► Additional supplemental material is published online only. To view, please visit the journal online (http://dx.doi.org/10.1136/bmjhci-2023-100746).

Received 03 February 2023
Accepted 21 August 2023

Check for updates

[1]Neurology, Yale School of Medicine, New Haven, Connecticut, USA
[2]Neurology, VA Connecticut Healthcare System West Haven VA Medical Center, West Haven, Connecticut, USA
[3]Biomedical Informatics and Data Science, Yale School of Medicine, New Haven, Connecticut, USA
[4]Neurology, Westchester Medical Center Health Network, Valhalla, New York, USA
[5]Department of Health and Human Services, Connecticut Department of Public Health, Bridgeport, Connecticut, USA

Correspondence to
Dr Hamada Hamid Altalib;
hamada.hamid@yale.edu

ABSTRACT

Objective The study aimed to measure the validity of International Classification of Diseases, 10th Edition (ICD-10) code F44.5 for functional seizure disorder (FSD) in the Veterans Affairs Connecticut Healthcare System electronic health record (VA EHR).

Methods The study used an informatics search tool, a natural language processing algorithm and a chart review to validate FSD coding.

Results The positive predictive value (PPV) for code F44.5 was calculated to be 44%.

Discussion ICD-10 introduced a specific code for FSD to improve coding validity. However, results revealed a meager (44%) PPV for code F44.5. Evaluation of the low diagnostic precision of FSD identified inconsistencies in the ICD-10 and VA EHR systems.

Conclusion Information system improvements may increase the precision of diagnostic coding by clinicians. Specifically, the EHR problem list should include commonly used diagnostic codes and an appropriately curated ICD-10 term list for 'seizure disorder,' and a single ICD code for FSD should be classified under neurology and psychiatry.

INTRODUCTION

Epilepsy is the fourth most common neurological disorder after Alzheimer disease, migraine and stroke.[1] Overall, 20%–30% of people seen at epilepsy centers for drug-resistant seizures are diagnosed with functional seizure disorder (FSD).[2] FSD is often misdiagnosed as epilepsy with several years of delay before a correct diagnosis.[3] Subsequently, FSD is incorrectly documented and miscoded as epilepsy in the electronic health record (EHR). The International Classification of Diseases, 10th Edition (ICD-10) introduced specific codes for the diagnosis of FSD and epileptic seizures, respectively, ICD-10 code F44.5—FSD, conversion disorder with seizures (code F44.5) and ICD-10 code G40.9—epilepsy, unspecified.[4] The differentiation of a code for FSD in the ICD-10 was intended to improve the validity of FSD diagnostic coding in the EHR.

A problem list is a compilation of diagnoses selected by clinicians during patient encounters and updated when a diagnosis changes.[5] Outpatient records rely on clinician-inputted problem lists in the EHR to identify and document medical conditions.[5] A single diagnosis may be represented by multiple, ICD-coded diagnostic terms. Correct diagnostic coding requires active maintenance of EHR problem lists and clinician judgement.[6] An assessment of the quality of diagnostic coding supports better patient care and improved outcomes. The study aimed to measure the precision of code F44.5 in the VA Healthcare System (VA) EHR.

METHODS

Setting

An informatics search tool and a natural language processing (NLP) algorithm identified potential cases of FSD through data extraction of VA inpatient, outpatient and pharmacy EHR charts across 170 VA medical centers in fiscal years 2002–2018.[3 7] The development and validation of the NLP tool is described elsewhere.[3] Briefly, the NLP classifier was validated using 2200 notes of veterans evaluated for seizure disorders. Reviewers used Yale cTakes Extension to annotate syntactic constructs, named entities and their negation context in the EHR. These annotations are passed to a classifier to detect NES patients. The achieved a positive predictive value (PPV) of 93%, a sensitivity of 99% and a F-score of 96%.

Sample

Of the 12 000 veterans diagnosed with FSD or epilepsy, a sample of 876 veterans coded with F44.5 were manually reviewed.[5] FSD classification was based on the International League Against Epilepsy (ILAE) Nonepileptic Seizures Task Force levels: definite (clinically established diagnosis of FSD with video electroencephalogram (vEEG)), probable (seizure witnessed by a neurologist), possible (some mention of FSD in the chart),

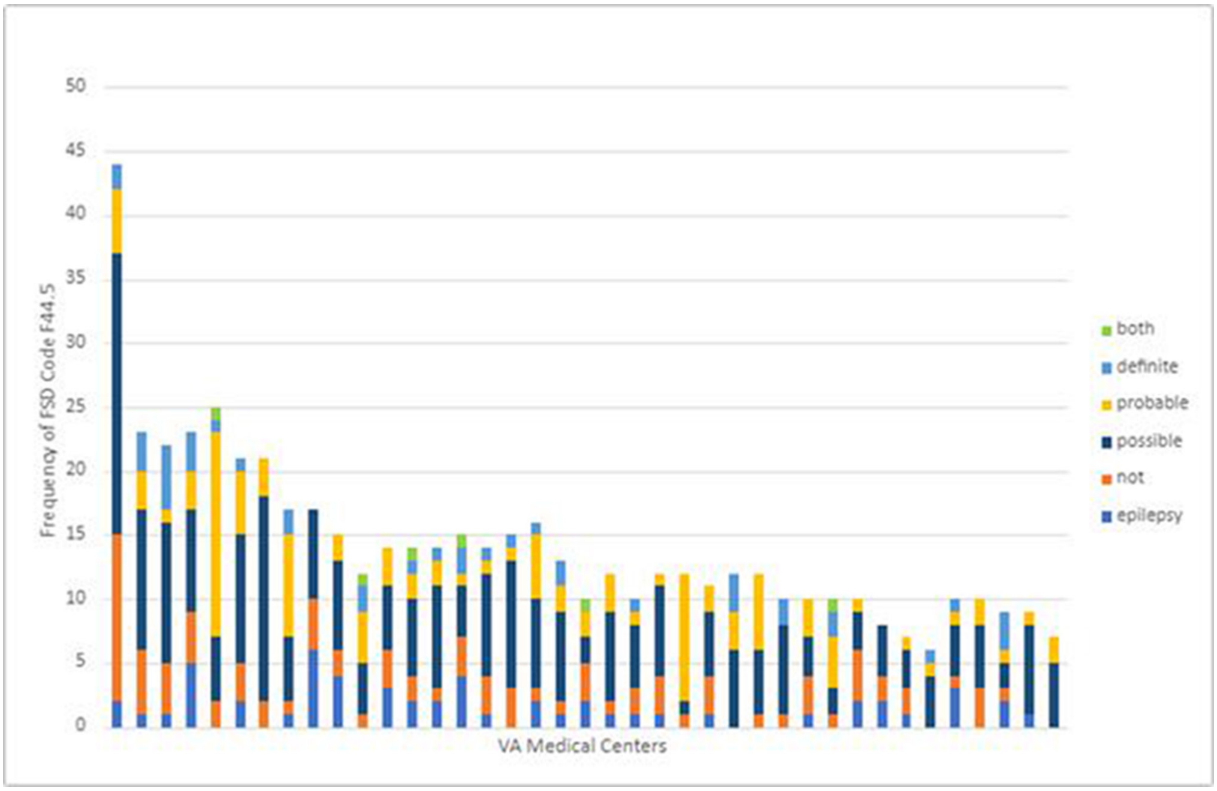

Figure 1 Frequency of FSD code F44.5 by Veterans Affairs Connecticut Healthcare System West Haven Medical Center. FSD, functional seizure disorder.

not (not FSD), epilepsy and both (mention of epilepsy and FSD in the chart).[8]

Statistical analysis

The PPV of code F44.5 was calculated with the true positive value to include definite (vEEG) and probable (seizure witnessed by a neurologist) groups, while the false positive (FP) value included not (not FSD) and epilepsy groups. Although the classification groups both and possible capture some cases of F44.5, the groups were excluded from the overall definition of F44.5 due to the possibility of captured FPs. Code F44.5 was used by 39 medical centres. Those medical centers were deidentified and stratified according to the frequency of code F44.5 usage (figure 1). Patient charts with missing data (n=3) for FSD classification and code F44.5 were removed from the analysis.

RESULTS

Results indicated a PPV of ~0.439 with a 95% CI of (0.391 to 0.487). This PPV demonstrated a low precision rate for code F44.5 in the VA EHR. The sample of patients (N=876) included: definite n=99 (11%), probable n=128 (15%), possible n=347 (40%), not n=206 (24%), epilepsy n=83 (9%) and both n=10 (1%) (online supplemental figure 2). Among the medical centers, the highest accuracy was 65% (17/26) (figure 1). Conversely, the medical center with the most FSD diagnoses had a poor accuracy of 14% (7/48) (figure 1).

DISCUSSION

FSD is poorly documented in the VA EHR, as evidenced by the 44% precision rate for code F44.5. Many people with FSD who are misdiagnosed with epilepsy are prescribed unnecessary medications that are harmful and costly to the patient. Correct diagnostic coding of FSD leads to appropriate, timely treatment, as well as the appropriate allocation of healthcare resources. After auditing the documentation workflow, we speculated that the low precision rate for code F44.5 is in part due to coding errors in the lookup diagnosis and problem list functions.

Most EHR systems provide a lookup diagnosis function. This function allows clinicians to search for a keyword which yields a problem list of diagnostic terms to select from. When a clinician uses the lookup diagnosis function for a keyword search, some problem lists yield a lengthy list of diagnoses. A problem list with too many diagnoses to scroll through may overwhelm the user.[5] For instance, a lookup diagnosis for the keyword epilepsy yielded a lengthy problem list with diagnoses ordered alphabetically. Conversion disorder with seizures or convulsions (a diagnostic term for FSD) was listed first (online supplemental figure 3). Clinicians may have inadvertently coded some epilepsy patients with an FSD diagnosis due to its convenient placement on top of the problem list.

In contrast to lengthy problem lists, some problem lists exclude relevant diagnoses. When a lookup diagnosis yields a problem list with a single diagnosis, that diagnosis may be selected by default. For example, a lookup diagnosis for seizure disorder resulted in a problem list populated with only one term—conversion disorder with seizures or convulsions (FSD)—which was unequivocally wrong in many cases (online supplemental figure 4). The selection of an incorrect diagnosis by default suggests that some patients with seizure disorder or epilepsy were miscoded for FSD. This default selection is due to the exclusion of relevant diagnoses in a problem list. The optimisation of the lookup diagnosis and problem list functions may improve clinician coding.

ILAE, ICD and Diagnostic and Statistical Manual of Mental Disorders DSM) have different diagnostic criteria for FSD.[4] This lack of consistency may lead to diagnostic ambiguity and coding error. For example, the ILAE diagnostic classification system has distinct codes for FSD, epilepsy and seizure disorder. However, ICD-9 did not have a code for FSD and ICD-10 classified FSD under psychiatry instead of neurology.[2 4] The shift in ICD-10 classification of FSD from a purely psychiatric disorder to a functional neurological disorder in the ICD-11 has not yet aligned with the DSM-5 classification of FSD as a mental health condition.[9] Consequently, the variability among classification systems and in coding practices of clinicians have likely undermined the validity of EHR-coded data.[2]

The lessons learned from this implementer report demonstrate the necessity of routine audits on ICD coding for real-world healthcare system applications. In fact, due to the proven inaccuracy of FSD coding, the VA did not include FSD within their internal VHA Support Service Center Neurology Cube, a web-based capital project application and tracking database.[10] Additionally, organisations should incentivise and support clinicians to maintain problem lists. Problem lists help facilitate patient care among clinicians and organisations. The standardisation of EHR problem lists and clinician coding practices can improve the quality of EHR-coded data and clinical processes.[10] Finally, the low precision rate of code F44.5 suggests that the EHR-coded data for the differential diagnoses of seizures (ie, epilepsy, focal seizures, generalised seizures) may be inaccurate (online supplemental figures 3,4).

There are some limitations to the study and to this assessment. First, the errors in the lookup diagnosis function were not tracked by individual medical centres. Thus, which medical centers were impacted by which errors are unknown. Second, the problem list errors identified in this report were of one medical center's VA EHR, and problem lists vary across medical centres. Finally, the unavailability of data on false negative diagnoses of FSD made it impossible to calculate the accuracy of code F44.5.

CONCLUSION

The low precision rate of FSD code F44.5 was affected by errors in the VA EHR's lookup diagnosis and problem list functions, and by variations in FSD criteria across diagnostic classification systems. This implementor report demonstrated a health informatics approach to troubleshooting data validity. In brief, three key recommendations to promote FSD code validity emerged from the analysis: the problem lists should be composed of the most common and most inclusive diagnostic codes; the problem list results of the lookup diagnosis function for seizure disorder must yield all relevant ICD-10 terms; and a single ICD code for FSD should be classified under neurology and psychiatry. Overall, implementing information system improvements will increase the validity of diagnostic coding by clinicians and of EHR-coded data.

Contributors HHA, CB and YB conceived of the presented idea. HHA, YB and SFA developed the theory and performed the computations. YB, DG and EJ-S conducted medical chart review. SA facilitated data mining and management. YB, EJ-S, SA, JG, MG and SFA verified the analytical methods. CB encouraged the investigation of using the NLP method and supervised the findings of this work. SFA took the lead in writing the manuscript with support from HHA. All authors provided critical feedback and helped shape the research, analysis and 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 applicable.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available on reasonable request.

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.

ORCID iDs
Sana F Ali http://orcid.org/0000-0003-0555-4208
Margaret Gopaul http://orcid.org/0000-0002-1329-3258

REFERENCES

1 Hamid H, Fodeh SJ, Lizama AG, *et al*. Validating a natural language processing tool to exclude psychogenic Nonepileptic seizures in electronic medical record-based epilepsy research. *Epilepsy Behav* 2013;29:578–80.
2 Jette N, Beghi E, Hesdorffer D, *et al*. ICD coding for epilepsy: past, present, and future--a report by the international league against epilepsy task force on ICD codes in epilepsy. *Epilepsia* 2015;56:348–55.
3 Daskivich TJ, Abedi G, Kaplan SH, *et al*. Electronic health record problem lists: accurate enough for risk adjustment *Am J Manag Care* 2018;24:e24–9.
4 Martin PM, Sbaffi L. Electronic health record and problem lists in Leeds, United kingdom: variability of general practitioners' views. *Health Informatics J* 2020;26:1898–911.
5 Altalib HH, Galluzzo D, Argraves S, *et al*. Managing functional neurological disorders: protocol of a cohort study on psychogenic non-epileptic seizures study. *Neuropsychiatr Dis Treat* 2019;15:3557–68.

6 LaFrance WC, Baker GA, Duncan R, *et al*. Minimum requirements for the diagnosis of psychogenic nonepileptic seizures: a staged approach. *Epilepsia* 2013;54:2005–18.
7 Aybek S, Perez DL. Diagnosis and management of functional neurological disorder. *BMJ* 2022;376:64.
8 Ertan D, Aybek S, LaFrance WC, *et al*. Functional (psychogenic non-epileptic/dissociative) seizures: why and how?. *J Neurol Neurosurg Psychiatry* 2022;93:144–57.
9 VA National Neurology Cube. VA Connecticut health system. 2022. Available: https://vssc.med.va.gov/VSSCMainApp/
10 Williams R, Kontopantelis E, Buchan I, *et al*. Clinical code set engineering for reusing EHR data for research: a review. *J Biomed Inform* 2017;70:1–13.

BMJ Health & Care Informatics

How to organise a datathon for bridging between data science and healthcare? Insights from the Technion-Rambam machine learning in healthcare datathon event

Jonathan Sobel [1], Ronit Almog[2], Leo Celi[3], Michal Yablowitz[4], Danny Eytan[2], Joachim Behar[1]

To cite: Sobel J, Almog R, Celi L, *et al*. How to organise a datathon for bridging between data science and healthcare? Insights from the Technion-Rambam machine learning in healthcare datathon event. *BMJ Health Care Inform* 2023;**30**:e100736. doi:10.1136/bmjhci-2023-100736

Received 31 January 2023
Accepted 25 July 2023

[1]Biomedical Engineering, Technion Israel Institute of Technology, Haifa, Israel
[2]Epidemiology and Pediatric Critical Care, Rambam Health Care Campus, Haifa, Israel
[3]Institute for Medical Engineering and Science, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA
[4]TIMNA- Israel's Ministry of Health Big Data Platform, State of Israel Ministry of Health, Jerusalem, Israel

Correspondence to
Dr Jonathan Sobel;
jsobel@campus.technion.ac.il

INTRODUCTION

A datathon is a time-constrained information-based competition involving data science applied to one or more challenges.[1–7] Datathons and hackathons differ in their focus, with datathons prioritising data analysis and modelling, while hackathons concentrate on building prototypes. Furthermore, hackathons can encompass a broad range of topics, spanning from software development to hardware design, whereas datathons are more narrowly focused on data analysis. In-person datathons offer the unique opportunity to learn alongside a community of fellow students and researchers, as well as to directly interact with clinicians and medical professionals. This is in contrast to Kaggle like competitions, which are often self-learning experiences.

Context of the event

A joint event organised by the Technion, Rambam Healthcare Campus and the MIT Critical Data group in March 2022 provided a unique opportunity to understand the challenges faced by leading researchers and clinicians working in the field of medical data science. The Technion is a leading science and technology research institutes and Rambam is the largest hospital in the north of Israel. It was organised as the inaugural event of a new joint Technion-Rambam initiative in medical AI (TERA), which aims to serve as an academic centre for medical AI committed to advanced medical and clinical research, with significant and actionable benefit to patient care.[3] The initiative opening event entitled 'Technion-Rambam Hack: Machine Learning in Healthcare,' was attended by about 250 people. The first two days consisted of a collaborative information-based competition that focused on solving real-world clinical problems through interdisciplinary teams and access to real data.[1–7] The datathon was followed by a one day conference with lectures delivered by researchers from the Technion, Rambam, MIT, the Israeli Ministry of Health (MOH), Clalit Health Services, GE Healthcare, and Roche.

The datathon days

The planning of the datathon and the conference began approximately six months before the event. After an initial brainstorming between the scientific committee, which included Technion principal scientists, Rambam clinicians and MIT scientists, a fundraising campaign was launched as list of potential speakers for the conference day was drawn up and invitations were extended. Communication around the event was initiated in November 2021 via social media platforms (Twitter, LinkedIn and Facebook). Students interested in the datathon were asked to apply to the event and were asked to complete a survey about their skills, their interests and their level of education (Bsc, Msc, Ph.D, alumni) and specialty (engineering or bio/med). We accepted approximately 70% of the applicants and the participation rate exceeded 95%. To ensure commitment from registrants to participate in the datathon, we required a registration fee of $25. In parallel, we contacted clinicians from Rambam and asked them to propose projects consisting of a medical question and to provide a relevant dataset to research the question. Four challenges proposed by clinicians who had collected large datasets in recent years and who presented challenging scientific questions which could be tackled by ML were selected. The projects were (1) Prediction of newborn birth weight by maternal parameters and previous newborn siblings birthweights,[8] (2) ML-based predictive model for bloodstream infections during hematopoietic stem cell transplantation,[9] (3) Prediction of recurrent hospitalisation in heart failure patients[10] and (4) Risk factor and severity prediction in hospitalised COVID-19 patients.[11 12] Project leaders were required to provide an agreement for their dataset, following the standard Hospital Institutional Review Board (IRB) process.

Two competing teams composed of 5–7 participants were assigned to each project. This approach was adopted for two reasons: first, to increase the likelihood of obtaining interesting results from at least one of the teams, and second, due to the resource-intensive nature of dataset creation, which involves extraction, curation, and anonymization processes. The projects were designed to have comparable difficulty in terms of the structured (tabular) medical data provided, and we intentionally limited the number of variables to prevent overwhelming teams with an excessive amount of data. We had participants from diverse fields, comprising 1/3 biologists/medical professionals and 2/3 engineers, computer scientists, statisticians, or mathematicians. Ethical agreement was requested from all participants during the subscription process. Each participant signed a consent and a non-disclosure agreement. Each team was assigned a clinical mentor from the Rambam and a data science mentor from either the Technion or the industry. Participants were selected based on their interests and competency (studies and skills). Our goal was to have mixed teams in terms of data analysis capacity and field knowledge to work on each challenge. Each team had a separate virtual machine with personal, secured access for each team member. During the 2 days of the event, the teams were split in several rooms at the Technion Faculty of Biomedical Engineering. Each team was asked to present its work at the end of the second day. Thereafter, using an external jury comprised of a principal investigator from the Technion, clinicians, Rambam epidemiology and IT department, and industrial partners, the three best teams were selected for the competition final, which took place on the conference day.

The conference day

The guest talks at the conference aimed to introduce clinical data science to a wide audience and provide a perspective on its future impact on medicine. There was a total of 12 lectures delivered. The lectures were divided into three thematic sessions which are: (1) current trends in machine learning in healthcare, (2) data stakeholders, (3) deployment of machine learning in medical practice. The full list of lectures and speakers is available on the event website for reference (https://technion-hack.github.io/).

HOW TO ORGANISE A DATATHON?

To organise a successful event, several important points should be well thought through before the event (figure 1). The checklist provided below should help any organiser in this process. We further elaborate on some of these key points reflecting on our more mature experience.

Datathon check list

- Venue: Physical/virtual/hybrid, dates, location.
- Logistics: catering, strong WIFI, rooms and amphitheatre.
- Partners: Industrial, NGO, clinicians and academic stakeholders, who may fund some part of the event (awards/venue/infrastructure) as well as deliver relevant talks during the event.
- Projects: Research project call for datasets with ethical consents (IRB) and specified aims/questions.
- IT support and secured computational infrastructure (for sensitive clinical data to be shared with participants).
- Mentors: Clinical and data science. Try to select senior mentors.
- Participants: Who is your targeted audience? (students/medical professionals/data scientists).
- Communication/PR: Information to participants and advertisement of the event (flyers/website/social media).
- Awards: Money for the winning teams or other gifts, and support for the continuation of the project (scientific publication and start-up/spin-off).

One of the first decisions is related to the place and dates of the future event, should it be virtual, in person or hybrid. On-site events offer the advantage of providing a face-to-face experience, facilitating networking opportunities, allowing for more immersive experiences, and creating a sense of community among attendees. Online events offer several advantages, including increased accessibility and convenience, the ability to reach a wider audience regardless of geographic location, reduced costs for both organisers and attendees, and the ability to easily collect and analyse data on attendee engagement and behaviour. Hybrid events offer the advantage of combining the best of both virtual and in-person events, allowing for a wider reach and increased engagement while still maintaining a personal touch. However, hybrid events tend to reduce in-person attendance because of the alternative online option, which may be more convenient. We preferred an in-person event since the main objective was to enable human interaction and initiate a professional local community interested in ML in medicine. Without a doubt, this was the right choice, and a virtual meeting would have had very limited impact. Additionally, two of the talks were delivered as recorded videos. It was noticeable that while these were projected on a large screen with high-quality resolution and sound, the audience did not focus on these presentations at all. Instead, they started consulting their emails or working

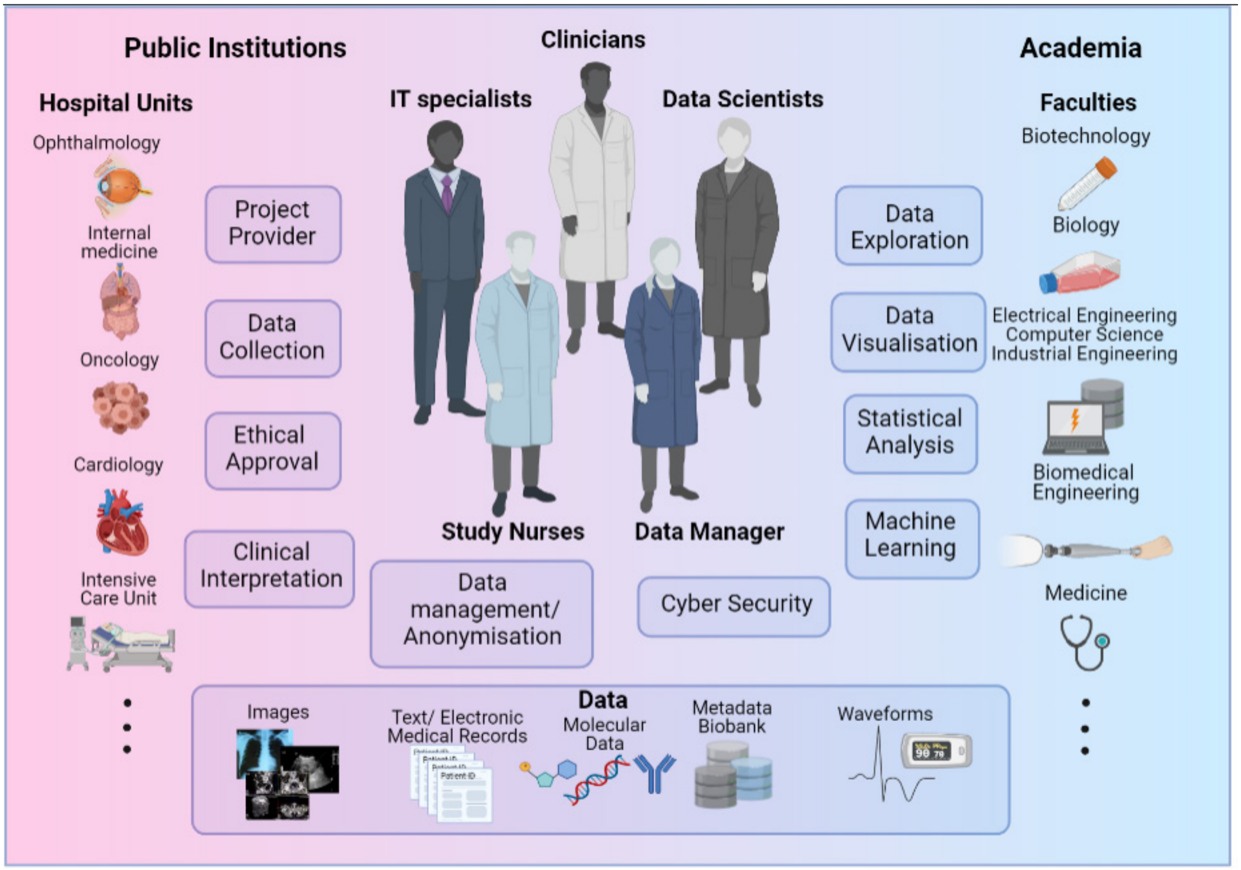

Figure 1 Datathon partners and essential components for sucessful projects. This figure was made with *BioRender*.

on their laptops. Thus, based on our experience, we strongly recommend an in-person meeting for such an event. Finally, we suggest that the lectures can be recorded and later released on YouTube if the organisers choose to widely distribute them – something we did – in order to maximise the impact of the event.

Recruiting sponsors was not the easiest task. Primarily, this was due to reaching out to local industries that had connections with some of the conference organisers. We presented the sponsors with various options. The generous partners had the opportunity to deliver a lecture and thereby promote their own research activities in the field of medical AI. This aligns well with the scientific programme and also helped in financing a portion of the event's cost. The lower-tier sponsorship option involved featuring their logo on our communications such as the website and flyers. However, this option did not attract any sponsors.

Another critical point for a datathon involves finding questions of clinical importance that can be addressed using previously collected data such as in.[13] There are several options here as some events may be more flexible in the sense that participants/projects leader can come with their data and questions or a more directed approach with defined datasets and questions. We chose projects from various medical fields rather than focusing on a specific problem area. This decision was made to foster the development of a professional community in the field of medical AI, promoting diversity in terms of the represented clinical specialties. Additionally, due to the time-constrained nature of the datathon, we aimed for students to work with real hospital data. Therefore, we sought datasets that had been developed by hospital researchers specifically for research purposes. In all cases, the data should be properly anonymized and the ethical statements from the IRB should be provided before the event. Mentors, with a clinical or a data science expertise, who can follow each team during the whole event are necessary to ensure the success of each project. The goal of a datathon is to demonstrate the effectiveness of a multidisciplinary approach where each team member can provide different expertise (medical, technical, social, legal and business). For that reason, we decided, as organisers, to create teams from the pool of participants prior to the datathon event.

We would like to add some additional recommendations based on things we would have done differently in retrospect. These include avoiding recorded lectures entirely. Additionally, it is important to ensure that speakers adhere to their allocated presentation time and to seek permission in advance for the use of pictures and videos from the event for marketing purposes. Finally, after the event, it would be beneficial to request feedback from participants through an online form in order to assess the impact of the event.

Twitter Jonathan Sobel @jonathansobel1

Acknowledgements We acknowledge the financial support of the Technion-Rambam Initiative in Artificial Intelligence in Medicine (TERA), the Technion Machine

Learning and Intelligent Systems (MLIS) Center, the Technion Human Health Initiative (THHI) and the MISTI MIT - Israel Zuckerman STEM Fund. This research was partially supported by The Milner Foundation, founded by Yuri Milner and his wife Julia. We are grateful to the Placide Nicod Foundation for their financial support (J.S.). We are grateful to the medical and technical staff from the IT and epidemiology departments at Rambam HCC. We are thankful to the Technion administrative staff for supporting the organization and communication of the event.

Contributors JS, JB, RA, LAC, and DE were involved in the conception and design. JS: was the coordinator. JS, JB, LAC, MY and DE wrote the first draft. All authors critically reviewed the first draft, and approved the final version.

Competing interests None declared.

Patient consent for publication Not applicable.

Ethics approval Not applicable.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement No data are available.

ORCID iD

Jonathan Sobel http://orcid.org/0000-0002-5111-4070

REFERENCES

1 Aboab J, Celi LA, Charlton P, *et al*. “A "Datathon" model to support cross-disciplinary collaboration”. *Sci Transl Med* 2016;8:333ps8.

2 Anslow C, Brosz J, Maurer F, *et al*. Datathons: an experience report of data Hackathons for data science education. Proceedings of the 47th ACM Technical Symposium on Computing Science Education (SIGCSE ’16); New York, NY, USA: Association for Computing Machinery, 2016:615–20

3 Sapci AH, Sapci HA. Artificial intelligence education and tools for medical and health Informatics students: systematic review. *JMIR Med Educ* 2020;6:e19285.

4 Lyndon MP, Cassidy MP, Celi LA, *et al*. Hacking Hackathons: preparing the next generation for the Multidisciplinary world of Healthcare technology. *Int J Med Inform* 2018;112:1–5.

5 Serpa Neto A, Kugener G, Bulgarelli L, *et al*. First Brazilian Datathon in critical care. *Rev Bras Ter Intensiva* 2018;30:6–8.

6 Pathanasethpong A, Soomlek C, Morley K, *et al*. Tackling regional public health issues using mobile health technology: event report of an mHealth Hackathon in Thailand. *JMIR Mhealth Uhealth* 2017;5:e155.

7 Li P, Xie C, Pollard T, *et al*. Promoting secondary analysis of electronic medical records in China: summary of the PLAGH-MIT critical data conference and health Datathon. *JMIR Med Inform* 2017;5:e43.

8 McCowan LME, Harding JE, Stewart AW. Customised birthweight centiles predict SGA pregnancies with perinatal morbidity. *BJOG* 2005;112:1026–33.

9 Gupta V, Braun TM, Chowdhury M, *et al*. A systematic review of machine learning techniques in hematopoietic stem cell transplantation (HSCT). *Sensors (Basel)* 2020;20:6100.

10 Ouwerkerk W, Voors AA, Zwinderman AH. Factors influencing the predictive power of models for predicting mortality and/or heart failure hospitalization in patients with heart failure. *JACC Heart Fail* 2014;2:429–36.

11 Reiner Benaim A, Sobel JA, Almog R, *et al*. Comparing COVID-19 and influenza presentation and trajectory. *Front Med* 2021;8.

12 Sobel JA, Levy J, Almog R, *et al*. Descriptive characteristics of continuous Oximetry measurement in moderate to severe COVID-19 patients. *Sci Rep* 2023;13:442.

13 Johnson AEW, Pollard TJ, Shen L, *et al*. MIMIC-III, a freely accessible critical care database. *Sci Data* 2016;3:160035.