Judgment Day: Symbolic AI versus Neuro-Symbolic AI in a Usability Study with UK Junior Doctors

Mercedes Arguello Casteleiro[1,*], Chloe Henson[2], Manoj Kulshrestha[2], Diego Maseda Fernandez[2], Julio Des Diz[3], Carlos Sevillano Torrado[3], Nava Maroto[4], Maria Jesus Fernandez Prieto[5], Tim Furmston[1], Chris Wroe[6], John Keane[1] and Robert Stevens[1]

[1]*Department of Computer Science, School of Engineering, University of Manchester, UK*
[2]*Mid Cheshire Hospital Foundation Trust (MCHFT), NHS England, UK*
[3]*Hospital do Salnés, Villagarcía de Arousa, Spain*
[4]*Depto. Lingüística Aplicada a la Ciencia y a la Tecnología, Universidad Politécnica de Madrid, Spain*
[5]*Salford Languages, University of Salford, UK*
[6]*BMJ, UK*

Abstract

This paper investigates two Artificial Intelligence (AI) approaches for transforming a body of evidence into knowledge graphs (KGs) supporting question answering from junior doctors. The manual symbolic AI approach focuses on COVID-19 and follows the traditional knowledge engineering approach of CommonKADS. The semi-automatic neuro-symbolic AI approach focuses on disease-treatment correlations and exploits prior knowledge and a type of 4-term analogy with embeddings from deep learning. Both AI approaches leveraged on nanopublication and micropublication ontologies (statement-based formalisations) to underpin KGs. The paper reports the results of a usability testing with 13 UK junior doctors.

Keywords

Artificial Intelligence, knowledge graphs, symbolic AI, neuro-symbolic AI, usability study

1. Introduction

How to improve quality of care for patients is a major challenge for healthcare organisations. The gap between "*what is known*" and "*what is done*" is an evidence to practice gap [1]. Evidence is the growing body of scientifically sound research available to clinicians and is not limited to information in PubMed/MEDLINE [2]. There are multiple types of rapidly updating, dynamic information sources. Evidence includes clinical point-of-care summaries, such as the British Medical Journal (BMJ) Best Practice [3] documents. Evidence also includes local/hospital-based and national clinical practice guidelines. The emergence of the Coronavirus Disease 2019 (COVID-19) has highlighted the need for timely support (best practice/evidence) for clinicians as they manage severely ill patients. In this paper, we investigate the real-world problem of transforming a body of evidence into knowledge graphs (KGs) [4] supporting Question Answering (QA) from resident doctors (a.k.a. junior doctors). However, making scientific evidence computer-interpretable brings some challenges:

- Managing changing knowledge and converting unstructured content into KGs [5].
- Lexical gaps, ambiguity and complex questions make difficult QA over KGs [6].

This paper addresses some of the above-mentioned challenges to build two proof-of-concept Web/mobile applications (called here demos) as QA systems powered by evidence-based KGs. The two demos have different Artificial Intelligence (AI) paradigms. The COVID-19 demo was developed following a symbolic AI (knowledge-based) approach, while the SemDeep demo was developed following a neuro-symbolic

SWAT4HCLS 2025: The 16th International Conference on Semantic Web Applications and Tools for Health Care and Life Sciences, February 24–27, 2025, Barcelona, Spain
"m.arguellocasteleiro@gmail.com" (M. A. Casteleiro)
0000-0001-9469-2068 (M. A. Casteleiro)

AI approach [7], combining artificial neural networks (i.e. word2vec neural language models [8]) with symbolic approaches (analogical reasoning with embeddings using prior knowledge). Both AI approaches (symbolic and neuro-symbolic) aim to solve the real-world problem of transforming a body of evidence from human-readable to computer-interpretable (actionable knowledge). This study focused on building QA demos powered by evidence-based KGs and investigating to what extend the QA demos are usable by junior doctors.

In UK, junior doctors are "*qualified doctors in postgraduate training who have not yet reached consultant or general practitioner status*" [9]. They are half of the England National Health Service (NHS) medical workforce, but require supervision, studying often outside of working hours [9].

This work sits within the vision of "*the Web of linked data*" (Semantic Web) [10] and utilised semantic web technologies: Resource Description Framework (RDF) [11] and Web Ontology Language (OWL) [12]. RDF is a graph-based representation of knowledge [11]. An OWL ontology may include descriptions of concepts and structural features (properties and relationships). We formally represent KGs in RDF and use as schema two ontologies in OWL: nanopublication [13] and micropublication [14].

1.1. Related Work

A recent in-depth review of studies implementing neuro-symbolic AI is in [15]. A survey of neuro-symbolic reasoning on KGs [16] recognised as future directions the inclusion of analogical reasoning. The System Usability Scale (SUS) [17] is a standard instrument that can provide an overall score for the usability of interactive systems. There are recent usability studies for KGs applications using modified SUS. We highlight: ALOHA [18] that is a graph-based visualization to browse an integrated dietary supplement knowledge base curated from scientific resources; and GoodTimes [19] that is a neuro-symbolic AI application (an interactive multimodal photo album leveraging on a neural language model and a KG) supporting personalised reminiscence therapy. According to 2024 AI Index Report [20]: "*current AI technology still has significant problems. It cannot reliably deal with facts, perform complex reasoning, or explain its conclusions*". Grounding is defined as "*the process of connecting the model to verifiable sources of information*" [21], aiming to enhance its trustworthiness by anchoring the model responses to information sources.

2. Materials and Methods

In healthcare, change is vital for delivering care based on the best available evidence or best practice [22], and thus, we considered the Lewin's model of change [22, 23]: 1) "*unfreezing*" (examine status quo); 2) "*moving*" (action research), and 3) "*refreezing*" (make changes permanent).

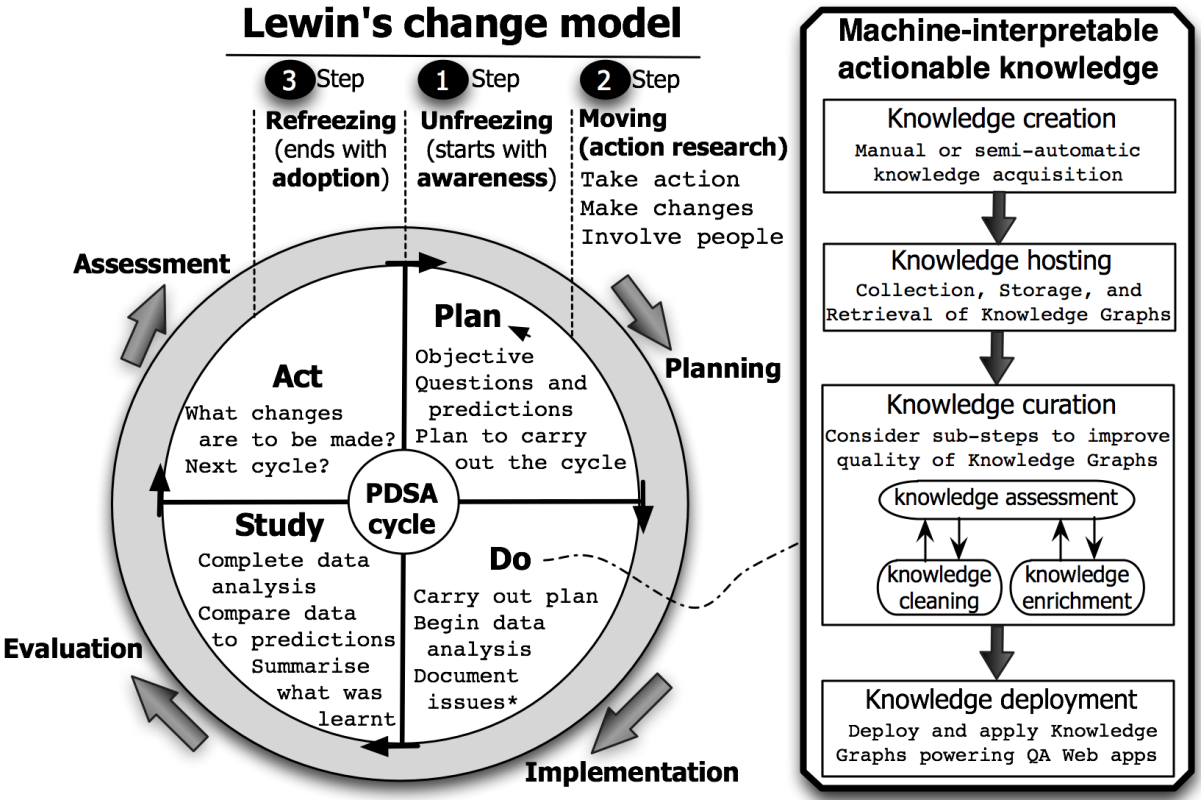
We related the Lewin's model of change to the four stages of the Plan–Do–Study–Act (PDSA) cycle. The PDSA cycle is a more detailed plan of how to generate change in healthcare [24] (left-side Figure 1): assessment, planning, implementation and evaluation. The PDSA cycle is used widely in healthcare improvement [25, 26] and is an agile method [27]. Figure 1 (right-side) outlines the process of building KGs from [6] that we followed. We also adhered to five principles from user-centred agile software development [28] as usability by UK junior doctors of QA demos is paramount.

2.1. Building Evidence-Based KGs for QA demos

Figure 1 (right-side) provides an overview of the several steps for building KGs [6]. There are differences and communalities between the symbolic AI and the neuro-symbolic AI approaches. **Knowledge creation**. We took a manual and semi-automatic approach to create models with the underpinning evidence for our KGs:

- **Symbolic AI** (manual approach) – the CommonKADS knowledge models [29] are top-down models built manually. They contain "*domain knowledge*", i.e. biomedical facts and recommendations (best available evidence/practice). We considered for COVID-19: clinical classification, clinical

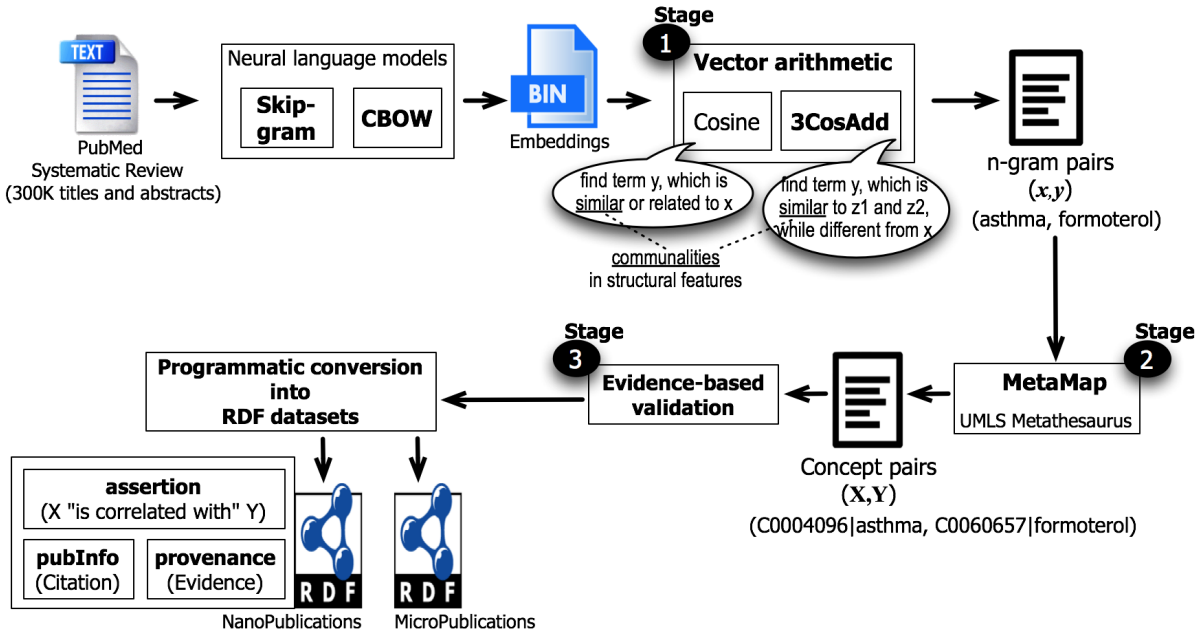
Figure 1: Methodological framework: the PDSA cycle (left) and the process of building KGs (right)



presentation (e.g. symptoms and laboratory abnormalities), diagnostic tests, clinical management including experimental treatments such as drugs/medications, complications of COVID-19 patients, and prognosis (clinical assessment scales). CommonKADS knowledge model construction has 3-stages. *Stage 1 Knowledge Identification:* Determine useful information sources, i.e. four chronological versions of "BMJ Best Practice COVID-19" and "BMJ Best Practice COVID-19 coexisting conditions" (documents released around 10th of May, June, July, and August 2020) [3] along with two NHS England COVID-19 speciality guidelines [30] unchanged from March to September 2020. *Stage 2 Knowledge Specification:* Build a specification of the knowledge model. The emphasis in this stage is on the domain schema, i.e. nanopublication [13] and micropublication [14]. This stage constructs an initial domain conceptualisation and produces a partial set of knowledge instances. *Stage 3 Knowledge Refinement:* Validate the knowledge model (paper simulation) and produce a complete set of knowledge instances.

- **Neuro-symbolic AI** (semi-automatic approach) – word2vec neural language models [8] (a.k.a. static embeddings) are bottom-up models containing representations of the meaning of words learnt from their distributions in texts. We followed the semi-automatic approach of Semantic Deep Learning (SemDeep for short) from [31]. SemDeep leverages on prior knowledge and a type of 4-term analogy (3CosAdd formula [32], to acquire correlations between well-known diseases (Dx) and treatments (Tx) from word2vec embeddings. SemDeep has 3-stages [31] depicted in Figure 2. *Stage 1 Knowledge Acquisition:* Vector arithmetic formulas are applied to n-gram embeddings created with CBOW and Skip-gram using 301,201 PubMed/MEDLINE systematic reviews (titles and available abstracts) [2]. We used prior knowledge and a type of 4-term analogy $y = -x + z1 + z2$ with n-gram embeddings to acquire correlations between diseases x and treatments $z1, z2, y$. For example: $(x, y) = (asthma, formoterol)$ with prior knowledge as evidence that z1 and z2 treat asthma. *Stage 2 Knowledge Organization (explicit conceptualization of the meaning of terms):* using MetaMap [33] to map n-gram pairs (x,y) to concept pairs (X, Y) from

Figure 2: Overview of the 3-stages for SemDeep, i.e. the neuro-symbolic AI approach followed



the UMLS (Unified Medical Language System) Metathesaurus [34]. *Stage 3 Knowledge Validation (an evidence-based evaluation followed by human audit):* Manual literature searches are part of the evidence-based validation for UMLS concept pairs (X, Y) to assign to concept Y some evidence (e.g. quotes from [3]) along with an evidence-based category.

Knowledge curation. We merged knowledge creation with curation before hosting. Both manual and semi-automatic approaches performed knowledge cleaning and enrichment, sub-steps for knowledge assessment (see Figure 1). **Knowledge hosting**. We combined the nanopublication [13] and micropublication [14] ontologies into a single OWL ontology. We also re-used axioms from different ontologies (see Figure 3). The curated knowledge is stored as N-Quads [35], allowing SPARQL queries [36].

Knowledge deployment. N-Quads RDF datasets are converted into JSON-LD (JSON for Linking Data) [37] with Apache Jena scripts [38]. RDF datasets follow the FAIR principles [39]: findable, accessible, interoperable, and reusable.

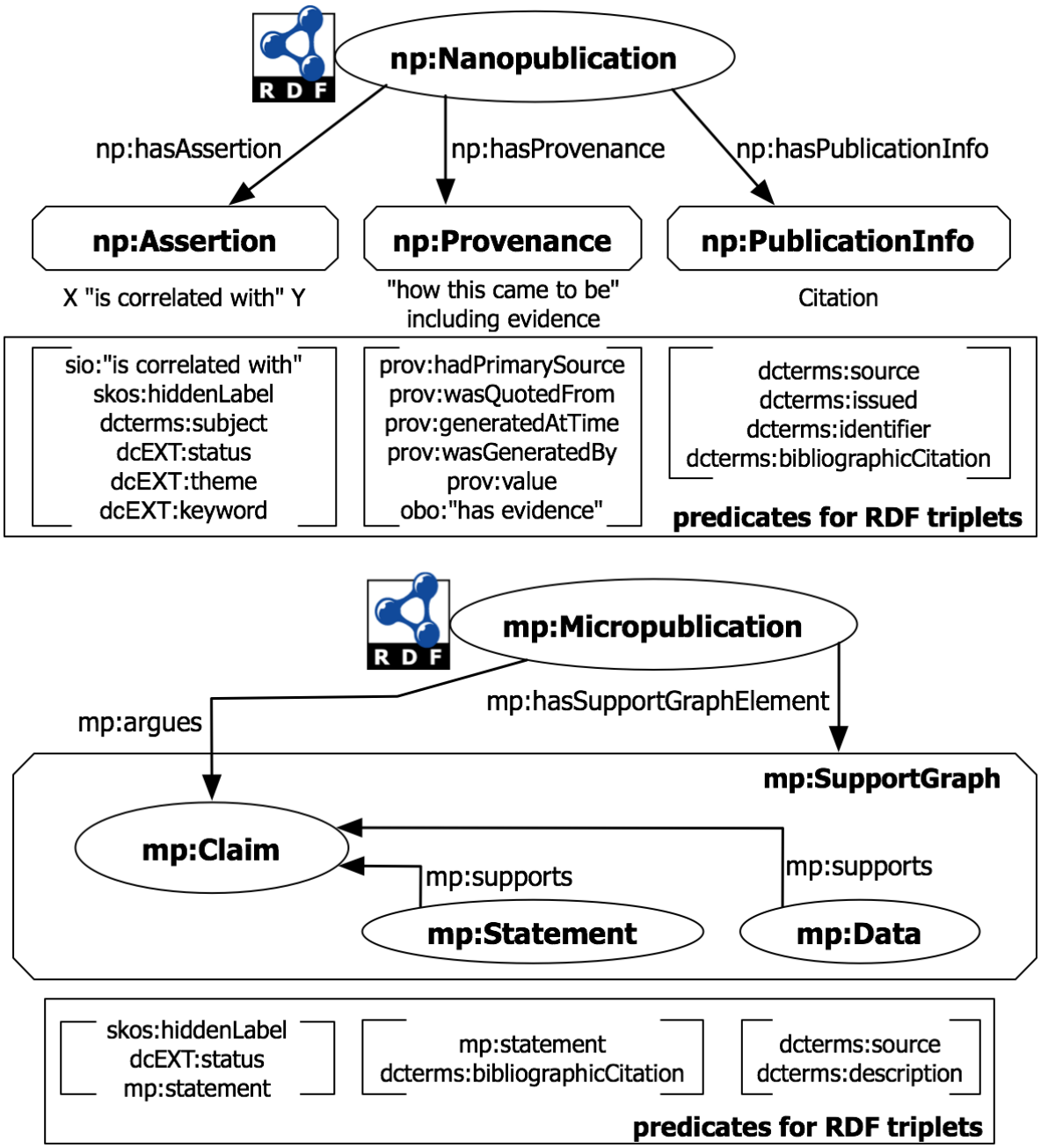
Knowledge deployment: QA demos. The COVID-19 demo has KGs created from the symbolic AI approach, i.e. CommonKADS knowledge models. The SemDeep demo has KGs created from neuro-symbolic AI approach, i.e. correlations between well-known diseases and treatments acquired from analogical reasoning with neural embeddings using prior knowledge. QA demos were built following five user-centred agile software development principles from [40].

2.2. Usability Evaluation

Medical consultants formulated 14 clinical questions (i.e. 7 questions per QA demo) as appropriate for junior doctors. The clinical questions for the COVID-19 demo (symbolic AI approach):

- How often is conjunctivitis reported in the clinical presentation for COVID-19?
- For coagulation screen, what are the most common abnormalities for COVID-19?
- What corticosteroids are recommended for the treatment of COVID-19?
- What are the considerations/recommendations for COVID patients with rheumatoid arthritis?
- Is National Early Warning Score 2 (NEWS2) recommended for use in COVID-19 patients?
- What clinical information can be an indicator of poor prognosis for COVID-19?

Figure 3: The nanopublication and micropublication ontologies with the modifications introduced



- What are the considerations/recommendations for COVID patients with comorbidities such as diabetes mellitus (any type), asthma, COPD, hypertension?

The clinical questions for the SemDeep demo (neuro-symbolic AI approach):

- Is anticholinergic treatment recommended in people with asthma?
- Why nebivolol is recommended to treat high blood pressure?
- Are dihydropyridine agents recommended in people with heart failure?
- Are tricyclic antidepressants recommended in patients with epilepsy?
- Is tramadol recommended in patients with osteoarthritis?
- Why recombinant erythropoietin is recommended in patients with CKD?
- Why iStent is recommended for patients with open-angle glaucoma?

The SUS is a 10-item questionnaire:

1. I think that I would like to use this system frequently.
2. I found the system unnecessarily complex.
3. I thought the system was easy to use.
4. I think that I would need the support of a technical person to be able to use this system.
5. I found the various functions in this system were well integrated.
6. I thought there was too much inconsistency in this system.
7. I would imagine that most people would learn to use this system very quickly.
8. I found the system very cumbersome ("awkward") to use.
9. I felt very confident using the system.
10. I needed to learn a lot of things before I could get going with this system.

The SUS can provide an overall score [17] for each QA demo. A participant assigns to each questionnaire item a value within a numbered 5-point scale [17] (a.k.a. raw scores), which can be plotted in a radar chart resembling a "*five-pointed star*" [41]. The appropriate number of participants in usability testing is controversial, although there is compelling evidence of the reliability of the 10-item SUS with sample sizes of at least 12 participants [42].

We also asked the junior doctors 5 additional questions Q1 to Q5 [43-46] requesting some feedback (see Appendix): Q1 and Q2 gather information about the device used; Q3 is about task completion [43]; Q4 relates to QA performance [44]; and Q5 assesses subjective satisfaction [43].

3. Results

This section starts with the outcome of the steps followed from [6] for building KGs. Next, we processed the 13 responses from UK junior doctors to the usability questionnaire.

3.1. Building Evidence-Based KGs for QA demos

Knowledge creation (including knowledge curation). Both manual and semi-automatic approaches obtained evidence (quoted text) for UMLS Metathesaurus concept pairs (X,Y). Every UMLS Metathesaurus concept has a concept unique identifier (CUI) [34]. Hence, we mapped clinical ideas to CUIs (symbolic AI) and we also mapped n-grams with vector representations of real numbers (neural embeddings) to CUIs (neuro-symbolic AI). For the manual CommonKADS knowledge models the X always refers to COVID-19 disease, whilst the semi-automatic SemDeep considers ten diseases with the n-grams: heart_failure; glaucoma; CKD (i.e. acronym for *Chronic Kidney Disease*); diabetes; asthma; epilepsy; arthritis; osteoarthritis; anaemia; and hypertension.

Knowledge hosting. Table 1 has the number of nanopublications and micropublications for COVID-19 per month (May to August 2020). For example, for the UMLS concept pair (X,Y)=(C5203670|COVID-19, C0015967|Fever) there is evidence from BMJ Best Practice of fever as a symptom commonly reported for COVID-19. Table 1 has two extra columns to differentiate between the status "*new*" (evidence added or changed since the previous version) and status "active" (evidence unchanged). For the semi-automatic SemDeep, Table **??** has the number of nanopublications and micropublications for the ten diseases, where the statement-based formalisations model evidence in support of disease-treatment correlations (i.e. Dx-Tx correlations between disease and treatment). For example, for the concept pair (X,Y)=(C0004096|asthma,C0060657|formoterol) there is evidence from BMJ Best Practice of formoterol as a "*Tx with therapeutic effect*" for asthma.

Knowledge deployment. RDF datasets programmatically created follow the FAIR principles [39]: findability (data is described with ontologies/rich metadata and identified uniquely), accessibility (data is accessible with HTTPS protocol and metadata is accessible without authentication/authorisation), interoperability (both RDF data and OWL metadata can be queried with SPARQL), and reusability (data has detailed provenance and metadata is available openly). UMLS CUIs are stored in the RDF

Table 1
Number of nanopublications and micropublications for COVID-19 per month, starting on 10th May

2020 monthly version	Statements with status = active	Statements with status = new	Number of nanopublications	Number of micropublications
August	691	94	785	267
July	545	169	714	251
June	260	335	595	226
May	406	0	406	181

Table 2
Number of nanopublications and micropublications per disease n-gram x

Disease n-gram x	Number of nanopublications	Number of micropublications
heart failure	32	7
glaucoma	12	5
CKD	9	1
diabetes	14	0
asthma	41	8
epilepsy	54	18
arthritis	24	8
osteoarthritis	56	16
anaemia	23	16
hypertension	38	15

plain literals of property *skos:hiddenLabel* [45] in the assertion of the nanopublication (see Figure 2). The assertion graph of the nanopublication contains the main claim of the nanopublications [13]. The "new" evidence has "new" as the RDF plain literals of property *dcEXT:status* in the assertion of the nanopublication (see Figure 2).

Knowledge deployment: QA demos – Figures 4 and 5 display screenshots of the COVID-19 and the SemDeep demo when they were accessed either from computers/laptops (Figure 4) or from mobile devices/phones (Figure 5). Both demos exploit the RDF datasets converted into JSON-LD with Apache Jena scripts [38]. The ubiquitous use of mobile phones by clinicians was included as a requirement whilst adhering to 5 user-centred agile software development principles from [40].

We developed reduced-compacted QA for mobile devices/phones (Figure 5) as well as enlarged-sophisticated QA for computers/laptops (Figure 4). When an end-user accesses the QA demos, the device is detected, and therefore, the end-user is automatically redirected to the Web-based QA demo for mobile devices/phones or for computers/laptops.

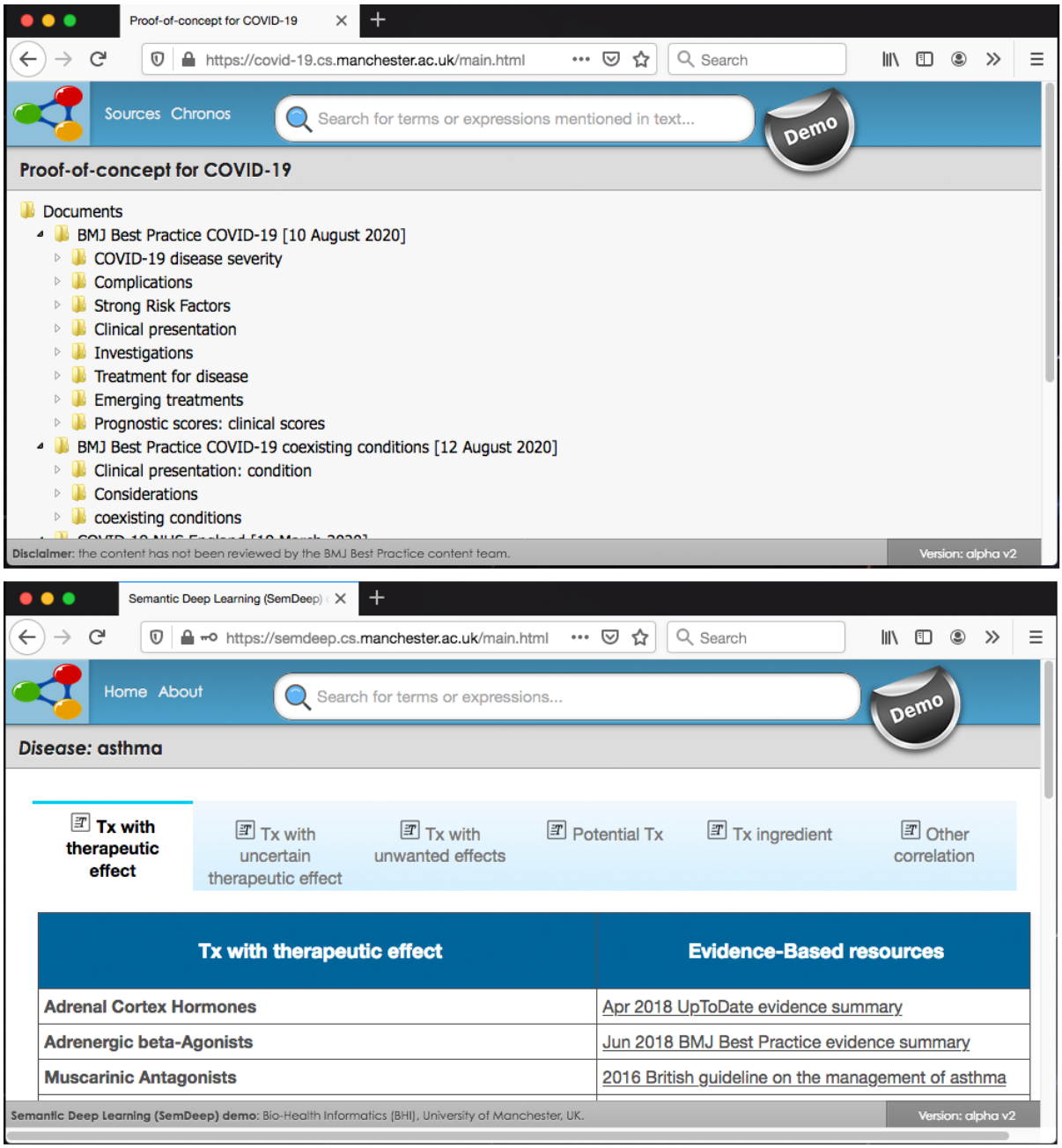
3.2. Usability Evaluation

Figure 6 plots the raw scores (i.e. original scores) from the 10-item SUS in a radar chart for each QA demo. The best system imaginable will have a perfect 5-pointed star. In Figure 6, the 5-pointed star for the COVID-19 demo is closer to perfection than the 5-pointed star for the SemDeep demo.

A lower variability of the participants' replies for the COVID-19 demo is observed from Figure 7 that plots the different values of participants' SUS score (grey lines) against SUS score benchmarks [42]: grading scale; adjective rating scale; and acceptability ranges. The arithmetic mean of the individual SUS scores for the 13 junior doctors is 80.38 for the COVID-19 demo and 73.46 for the SemDeep demo (see Appendix for the calculation details).

According to the grading scale (a.k.a. school grading scale) from [42] (Figure 7), the mean SUS scores from 13 junior doctors are: grade A- (SUS score 80.38) for the COVID-19 demo; and grade B- (SUS score 73.46) for the SemDeep demo. Both QA demos have SUS scores above 70, i.e. they both have "*acceptable*" SUS score in the acceptability ranges scale [42].

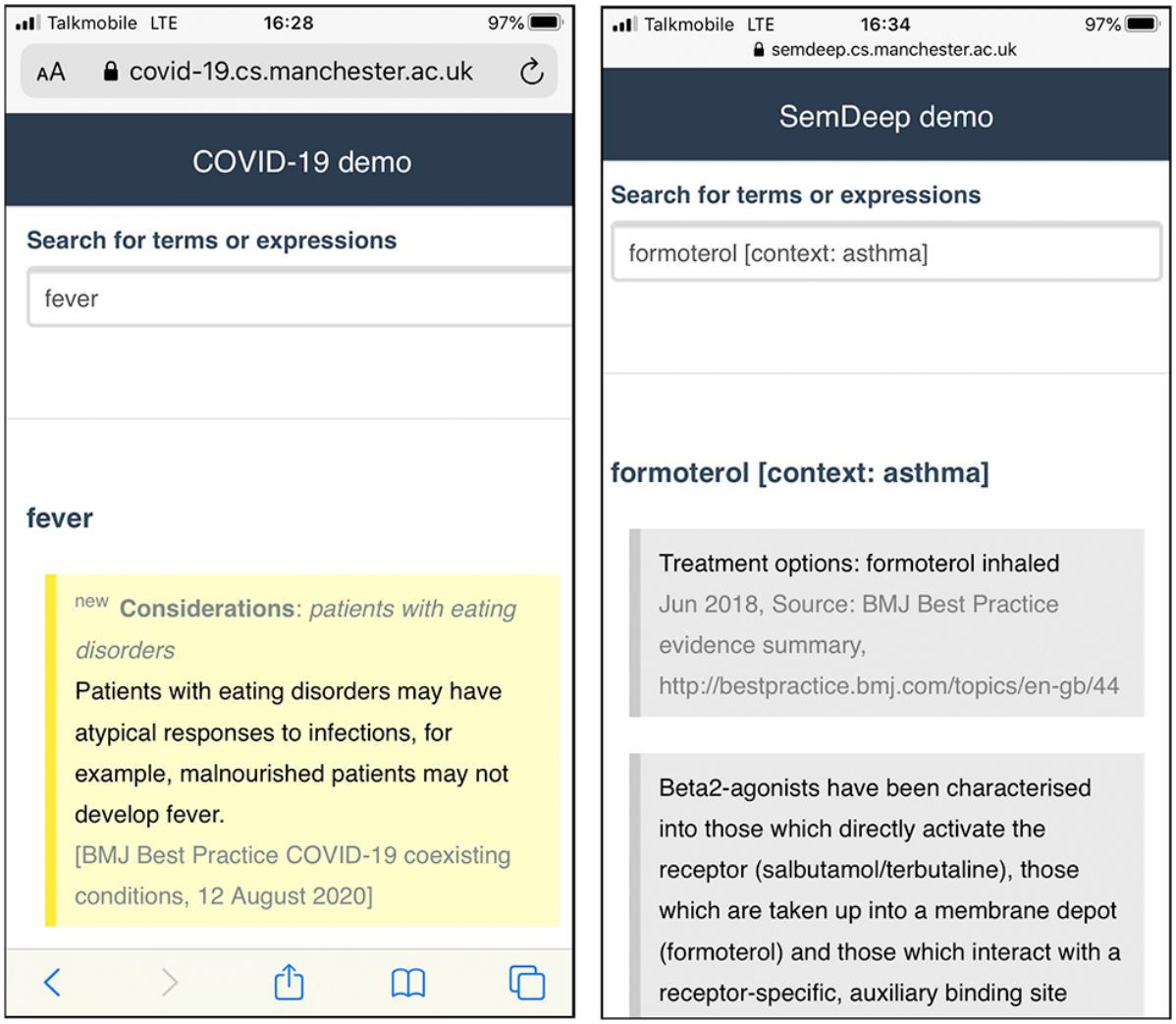
Figure 4: Accessing COVID-19 demo (top) and SemDeep demo (bottom) from computers/laptops



According to the adjective rating scale from [42] (Figure 7), for the best system imaginable the SUS score is 100 (the maximum SUS score), and for the worst system imaginable the SUS score is 25. For both the COVID-19 demo and the SemDeep demo, the mean SUS scores are greater than 72.75 "*Good SUS score*" and less than 85.58 "*Excellent SUS score*".

Answers to Q1 to Q5 questions requesting some feedback. 3 of the 13 junior doctors used mobile phones (answer Q1 and Q2). According to the replies to Q3: 9 of 13 junior doctors found answers for all clinical questions using the COVID-19 demo, and 10 of 13 junior doctors found answers for all clinical questions using the SemDeep demo. The task completion time (answer Q4) is consistent with other studies [46, 47]. 9 or 10 of the 13 junior doctors (answer Q5) would like other documents (e.g. BMJ Best Practice documents) accessible in the same way.

Figure 5: Accessing COVID-19 demo (left) and SemDeep demo (right) from mobile phone



4. Conclusions

Junior doctors already act on knowledge gaps they have perceived, and seek biomedical facts and recommendations from best evidence/practice. This study looked into AI approaches for transferring best evidence/practice from generators/providers (e.g. BMJ) into healthcare services (e.g. NHS Trust). We investigated a symbolic AI approach (based on CommonKADS) and a neuro-symbolic AI approach (based on SemDeep) to transform a body of evidence into evidence-based KGs underpinned by nanopublication and micropublication ontologies. The paper has shown the viability of deploying QA demos powered by evidence-based KGs. We developed reduced-compacted QA demos for mobile devices/phones that co-exist with enlarged-sophisticated QA demos for computers/laptops.

We investigated the usability of two QA demos with 13 UK junior doctors. 9 or 10 of the 13 junior doctors answered all the clinical questions with the aid of the QA demos. The COVID-19 demo (symbolic AI) achieved a mean SUS score of 80.38 (grade A-), which is higher than the mean SUS score of 73.46 (grade B-) obtained for the SemDeep demo (neuro-symbolic AI). Both QA demos have “acceptable” SUS score (above 70) that are also greater than 72.75 “Good SUS score”.

Figure 6: Plotting the raw scores in a radar chart for each QA demo

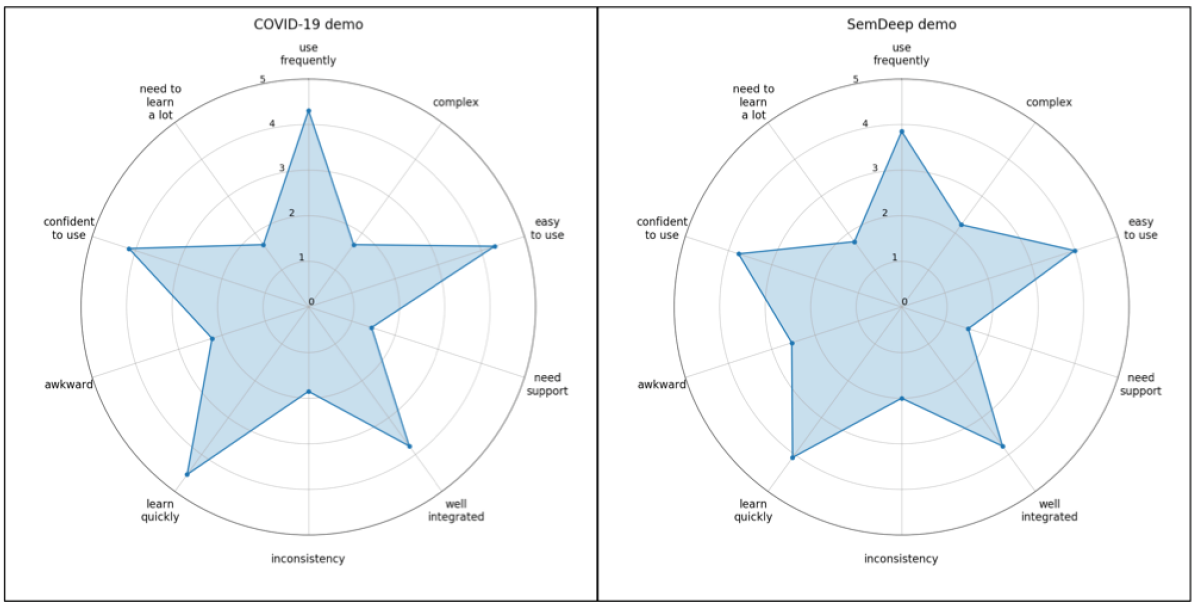
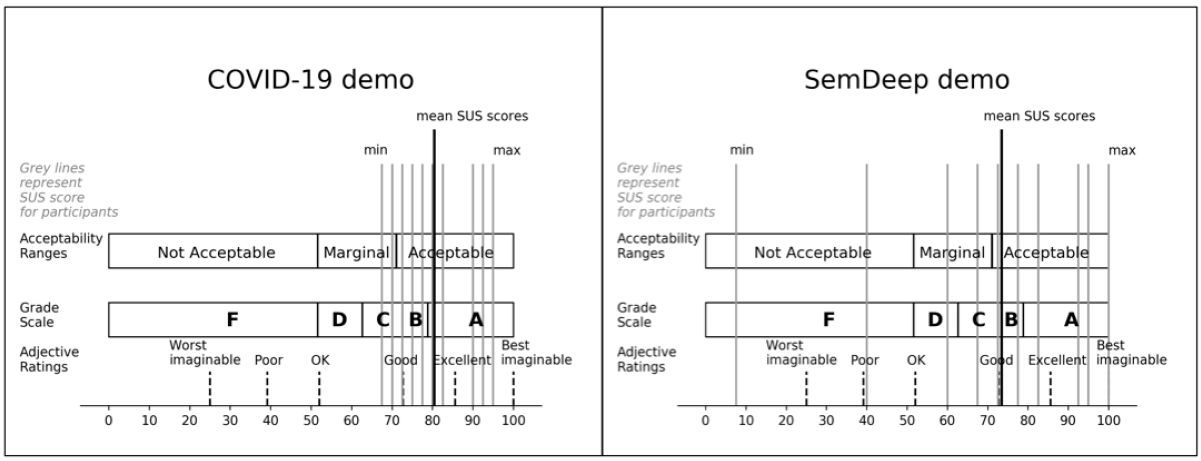


Figure 7: Comparison of SUS scores for each QA demo with SUS benchmarks



5. Acknowledgements

This study is an outcome of the EPSRC IAA Proof of Concept Scheme "Semantic deep learning and Knowledge Graphs to support clinicians: a user evaluation study" with BMJ and MCHFT as partners. The study obtained the approval of the University of Manchester Research Ethics Committee (ref.2018-3442-5107 and ref.2020-3442-15691). CW works for BMJ that produces the clinical decision support tool BMJ Best Practice. All other authors have no conflict of interest to declare.

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] J. H. Elliott, T. Turner, O. Clavisi, J. Thomas, J. P. T. Higgins, C. Mavergames, R. L. Gruen, Living systematic reviews: An emerging opportunity to narrow the evidence-practice gap, PLOS Medicine 11 (2014) 1–6. doi:10.1371/journal.pmed.1001603.

[2] Pubmed/medline, 2024. URL: https://pubmed.ncbi.nlm.nih.gov/.

[3] Bmj best practice, 2024, 2024. URL: https://bestpractice.bmj.com/info/.

[4] Making science computable, 2024. URL: https://confluence.hl7.org/display/CDS/HEvKA-Health+Evidence+Knowledge+Accelerator.

[5] N. Noy, Y. Gao, A. Jain, A. Narayanan, A. Patterson, J. Taylor, Industry-scale knowledge graphs: lessons and challenges, Commun. ACM 62 (2019) 36–43. doi:10.1145/3331166.

[6] D. Fensel, U. Simsek, K. Angele, E. Huaman, E. Karle, O. Panasiuk, I. Toma, J. Umbrich, A. Wahler, Knowledge Graphs : Methodology, Tools and Selected Use Cases, 1st ed. 2020. ed., Springer International Publishing, Cham, 2020.

[7] P. Hitzler, A. Eberhart, M. Ebrahimi, M. K. Sarker, L. Zhou, Neuro-symbolic approaches in artificial intelligence, National Science Review 9 (2022) nwac035. URL: https://doi.org/10.1093/nsr/nwac035. doi:10.1093/nsr/nwac035.

[8] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, International Conference on Learning Representations 2013, 2013. URL: https://arxiv.org/abs/1301.3781. arXiv:1301.3781.

[9] F. K. Lock, D. Carrieri, Factors affecting the uk junior doctor workforce retention crisis: an integrative review, BMJ Open 12 (2022). doi:10.1136/bmjopen-2021-059397.

[10] W3c semantic web, 2006. URL: https://www.w3.org/DesignIssues/LinkedData.

[11] Resource description framework (rdf), 2014. URL: https://www.w3.org/TR/rdf11-concepts/.

[12] Web ontology language (owl), 2012. URL: https://www.w3.org/TR/owl2-overview/.

[13] Nanopublication guidelines, 2024. URL: https://nanopub.net/guidelines/working_draft/.

[14] T. Clark, P. N. Ciccarese, C. A. Goble, Micropublications: a semantic model for claims, evidence, arguments and annotations in biomedical communications, Journal of Biomedical Semantics 5 (2014) 28. doi:10.1186/2041-1480-5-28.

[15] K. Hamilton, A. Nayak, B. Božić, L. Longo, Is neuro-symbolic ai meeting its promises in natural language processing? a structured review, Semantic Web 15 (2024) 1265–1306. doi:10.3233/SW-223228.

[16] J. Zhang, B. Chen, L. Zhang, X. Ke, H. Ding, Neural, symbolic and neural-symbolic reasoning on knowledge graphs, AI Open 2 (2021) 14–35. doi:https://doi.org/10.1016/j.aiopen.2021.03.001.

[17] J. Brooke, Sus-a quick and dirty usability scale. usability evaluation in industry., 1996. URL: http://hell.meiert.org/core/pdf/sus.pdf.

[18] X. He, R. Zhang, R. Rizvi, J. Vasilakes, X. Yang, Y. Guo, Z. He, M. Prosperi, J. Huo, J. Alpert, J. Bian, Aloha: developing an interactive graph-based visualization for dietary supplement knowledge graph through user-centered design, BMC Medical Informatics and Decision Making 19 (2019) 150. doi:10.1186/s12911-019-0857-1.

[19] X. Wang, J. Li, T. Liang, W. U. Hasan, K. T. Zaman, Y. Du, B. Xie, C. Tao, Promoting personalized reminiscence among cognitively intact older adults through an ai-driven interactive multimodal photo album: Development and usability study, JMIR Aging 7 (2024) e49415. doi:10.2196/49415.

[20] Ai index report, 2024. URL: https://aiindex.stanford.edu/report/.

[21] Grounding, 2024. URL: https://ai.google.dev/gemini-api/docs/grounding.

[22] R. Moen, C. Norman, Circling back, Quality Progress 43 (2010) 22–28.

[23] G. Mitchell, Selecting the best theory to implement planned change, Nurs. Manag. (Harrow) 20 (2013) 32–37.

[24] M. J. Taylor, C. McNicholas, C. Nicolay, A. Darzi, D. Bell, J. E. Reed, Systematic review of the application of the plan–do–study–act method to improve quality in healthcare, BMJ Quality & Safety 23 (2014) 290–298. doi:10.1136/bmjqs-2013-001862.

[25] C. Larman, V. Basili, Iterative and incremental developments. a brief history, Computer 36 (2003) 47–56. doi:10.1109/MC.2003.1204375.
[26] A. Pearson, B. Vaughan, M. Fitzgerald, Nursing Models for Practice, Butterworth-Heinemann Medical, 1996.
[27] N. Ivankova, N. Wingo, Applying mixed methods in action research: Methodological potentials and advantages, American Behavioral Scientist 62 (2018) 978–997. doi:10.1177/0002764218772673.
[28] V. L. Plano Clark, N. V. Ivankova, Mixed methods research: A guide to the field, 2016. doi:10.4135/9781483398341.
[29] Commonkads, 2024. URL: https://commonkads.org/.
[30] Nhs england specialty guides, 2024. URL: https://www.england.nhs.uk/coronavirus/.
[31] M. Arguello Casteleiro, J. Des Diz, N. Maroto, M. J. Fernandez Prieto, S. Peters, C. Wroe, C. Sevillano Torrado, D. Maseda Fernandez, R. Stevens, Semantic deep learning: Prior knowledge and a type of four-term embedding analogy to acquire treatments for well-known diseases, JMIR Med. Inform. 8 (2020) e16948.
[32] O. Levy, Y. Goldberg, Linguistic regularities in sparse and explicit word representations, in: R. Morante, S. W.-t. Yih (Eds.), Proceedings of the Eighteenth Conference on Computational Natural Language Learning, Association for Computational Linguistics, Ann Arbor, Michigan, 2014, pp. 171–180. URL: https://aclanthology.org/W14-1618/. doi:10.3115/v1/W14-1618.
[33] Metamap, 2024. URL: https://lhncbc.nlm.nih.gov/ii/tools/MetaMap.html.
[34] Umls metathesaurus, 2024. URL: https://www.nlm.nih.gov/research/umls.
[35] Rdf 1.2 n-quads, 2024. URL: https://www.w3.org/TR/rdf12-n-quads/.
[36] Sparql 1.2, 2024. URL: https://www.w3.org/TR/sparql12-update/.
[37] Json-ld, 2024. URL: https://json-ld.org/.
[38] Rdf in apache jena, 2024. URL: https://jena.apache.org/documentation/io/.
[39] Fair principles, 2024. URL: https://www.go-fair.org/fair-principles/.
[40] M. Brhel, H. Meth, A. Maedche, K. Werder, Exploring principles of user-centered agile software development: A literature review, Information and Software Technology 61 (2015) 163–181. doi:https://doi.org/10.1016/j.infsof.2015.01.004.
[41] D. Hariyanto, A. C. Nugraha, A. Asmara, H. Liu, An asynchronous serial communication learning media: Usability evaluation, Journal of Physics: Conference Series 1413 (2019) 012018. URL: https://doi.org/10.1088/1742-6596/1413/1/012018. doi:10.1088/1742-6596/1413/1/012018.
[42] J. R. Lewis, The system usability scale: Past, present, and future, International Journal of Human–Computer Interaction 34 (2018) 577–590. doi:10.1080/10447318.2018.1455307.
[43] K. Hornbæk, Current practice in measuring usability: Challenges to usability studies and research, International Journal of Human-Computer Studies 64 (2006) 79–102. doi:https://doi.org/10.1016/j.ijhcs.2005.06.002.
[44] J. Sauro, J. R. Lewis, Introduction and how to use this book, in: Quantifying the User Experience, Elsevier, 2012, pp. 1–8. doi:10.1016/C2010-0-65192-3.
[45] Skos, 2009. URL: https://www.w3.org/TR/skos-reference/.
[46] G. Del Fiol, T. E. Workman, P. N. Gorman, Clinical questions raised by clinicians at the point of care: a systematic review, JAMA Intern. Med. 174 (2014) 710–718.
[47] A. van der Vegt, G. Zuccon, B. Koopman, A. Deacon, How searching under time pressure impacts clinical decision making, J. Med. Libr. Assoc. 108 (2020) 564–573.

A. Calculating the SUS score

A participant assigns to each SUS questionnaire item a value within a numbered 5-point scale, where 1 means "*Strongly disagree*" and 5 means "*Strongly agree*" [17]. A SUS questionnaire item with value 3 means "neither agree nor disagree" (i.e. "*neutral*", the centre of the rating scale). The original scores for the 10-item SUS questionnaire [17] are called the raw scores, which can be visualised with a radar

chart. For the best system imaginable (i.e. a system with the maximum SUS score of 100), the radar chart will show a perfect *5-pointed star* [41]

The score contributions within the range 0 to 4 are calculated from the original scores (raw scores) within the range 1 to 5. To calculate the score contributions [17]:

- For questionnaire item 1, 3, 5, 7 and 9 the score contribution is the raw score minus 1.
- For questionnaire item 2, 4, 6, 8 and 10 the score contribution is 5 minus the raw score.

COVID-19 demo: *calculating the SUS score for the symbolic AI approach* – Table 3 shows the original scores per participant (the so-called raw scores) after the evaluation sessions with junior doctors. In Table 3, cells with red background color have an unexpected value.

- Odd-numbered SUS items 1, 3, 5, 7, and 9 are expected to have the highest values 4 or 5
- Even-numbered SUS items 2, 4, 6, 8, 10 are expected to have the lowest values 1 or 2

Table 3
The raw scores (original scores assigned) of 13 UK junior doctors for the COVID-19 demo

Identifier	Participant													Arithmetic mean	Standard deviation
	1	2	3	4	5	6	7	8	9	10	11	12	13		
SUS **Q1**	4	4	4	4	5	5	4	5	4	4	4	5	4	4.31	0.48
SUS **Q2**	2	2	2	1	1	2	1	1	2	**3**	1	2	2	1.69	0.63
SUS **Q3**	4	5	**3**	5	4	4	5	5	4	4	5	4	4	4.31	0.63
SUS **Q4**	1	1	1	2	2	1	1	1	1	2	1	1	4	1.46	0.88
SUS **Q5**	**3**	4	4	4	**3**	4	**3**	**3**	4	4	5	4	4	3.77	0.60
SUS **Q6**	4	2	2	2	2	1	2	2	1	2	1	1	2	1.85	0.80
SUS **Q7**	5	4	4	4	4	5	5	5	4	5	5	5	4	4.54	0.52
SUS **Q8**	**3**	4	4	2	2	1	2	1	**3**	**3**	1	1	2	2.23	1.09
SUS **Q9**	**3**	4	**3**	5	4	5	4	5	4	4	4	5	4	4.15	0.69
SUS **Q10**	1	2	2	2	1	2	2	1	2	2	1	2	2	1.69	0.48

The last 2 columns in Table 4 show the arithmetic mean and standard deviation per SUS item. The last 2 rows in Table 4 have the total score and the SUS score calculated multiplying the total score by 2.5. The SUS score for the COVID-19 demo is 80.38 (the arithmetic mean of the individual SUS scores for the 13 participants) with standard deviation of 9.29.

SemDeep demo: *calculating the SUS score for the neuro-symbolic AI approach* – We repeated the calculations (Table 5 and 6). The SUS score for the SemDeep demo is 73.46 (the arithmetic mean of the individual SUS scores for the 13 participants) with standard deviation of 25.73.

B. Requesting some feedback

Figure 8 has 5 closed-ended questions convertible into numbers (i.e. quantitative data):

- Q1 and Q2 gather information about the device used to access the QA demos
- Q3 relates to task completion, which measures effectiveness (objective usability) [43]
- Q4 gathers the time on task, which is a good way to assess QA performance [44]. This question for task completion time measures efficiency (objective usability) [43].
- Q5 assesses subjective satisfaction [43]

Table 7 reports the percent agreement on positive ratings (yes answer) for questions Q1 to Q5. For example, for question Q3: 9 of 13 junior doctors found answers for all clinical questions using the COVID-19 demo, and 10 of 13 junior doctors found answers for all clinical questions using the SemDeep demo. An answer of 15 minutes or less (about 2 minutes for 7 clinical questions) to Q4 is re-interpreted as the "yes" answer for Q4. Previous studies from the literature [46, 47] reported less than 2 to 3 minutes as the time that clinicians spent seeking an answer to a specific question.

Table 4
The score contributions to calculate the SUS score for the COVID-19 demo

	Participant													Arithmetic mean	Standard deviation
	1	2	3	4	5	6	7	8	9	10	11	12	13		
SUS **Q1**	3	3	3	3	4	4	3	4	3	3	3	4	3	3.31	0.48
SUS **Q2**	3	3	3	4	4	3	4	4	3	2	4	3	3	3.31	0.63
SUS **Q3**	3	4	2	4	3	3	4	4	3	3	4	3	3	3.31	0.63
SUS **Q4**	4	4	4	3	3	4	4	4	4	3	4	4	1	3.54	0.88
SUS **Q5**	2	3	3	3	2	3	2	2	3	3	4	3	3	2.77	0.60
SUS **Q6**	1	3	3	3	3	4	3	3	4	3	4	4	3	3.15	0.80
SUS **Q7**	4	3	3	3	3	4	4	4	3	4	4	4	3	3.54	0.52
SUS **Q8**	2	1	1	3	3	4	3	4	2	2	4	4	3	2.77	1.09
SUS **Q9**	2	3	2	4	3	4	3	4	3	3	3	4	3	3.15	0.69
SUS **Q10**	4	3	3	3	4	3	3	4	3	3	4	3	3	3.31	0.48
Total score	28	30	27	33	32	36	33	37	31	29	38	36	28	32.15	3.72
SUS score	70.0	75.0	67.5	82.5	80.0	90.0	82.5	92.5	77.5	72.5	95.0	90.0	70.0	80.38	9.29

Table 5
The raw scores (original scores assigned) of 13 UK junior doctors for the SemDeep demo

Identifier	Participant													Arithmetic mean	Standard deviation
	1	2	3	4	5	6	7	8	9	10	11	12	13		
SUS **Q1**	4	4	4	4	1	5	4	5	5	2	5	2	5	3.85	1.34
SUS **Q2**	2	**3**	2	2	5	1	1	1	1	4	1	4	2	2.23	1.36
SUS **Q3**	4	5	**3**	5	1	4	4	5	5	2	5	4	5	4.00	1.29
SUS **Q4**	1	1	1	2	4	1	1	1	1	**3**	1	1	2	1.54	0.97
SUS **Q5**	**3**	**3**	4	5	1	5	**3**	5	4	**3**	5	**3**	5	3.77	1.24
SUS **Q6**	2	2	2	2	4	1	2	1	1	**3**	1	**3**	2	2.00	0.91
SUS **Q7**	5	4	4	5	1	5	5	5	4	**3**	4	4	4	4.08	1.12
SUS **Q8**	2	4	4	2	5	2	1	1	2	4	1	**3**	2	2.54	1.33
SUS **Q9**	**3**	4	**3**	4	1	5	4	5	5	**3**	4	4	4	3.77	1.09
SUS **Q10**	1	1	2	2	4	1	2	1	1	**3**	1	2	2	1.77	0.93

Table 6
The score contributions to calculate the SUS score for the SemDeep demo

	Participant													Arithmetic mean	Standard deviation
	1	2	3	4	5	6	7	8	9	10	11	12	13		
SUS **Q1**	3	3	3	3	0	4	3	4	4	1	4	1	4	2.85	1.34
SUS **Q2**	3	2	3	3	0	4	4	4	4	1	4	1	3	2.77	1.36
SUS **Q3**	3	4	2	4	0	3	3	4	4	1	4	3	4	3.00	1.29
SUS **Q4**	4	4	4	3	1	4	4	4	4	2	4	4	3	3.46	0.97
SUS **Q5**	2	2	3	4	0	4	2	4	3	2	4	2	4	2.77	1.24
SUS **Q6**	3	3	3	3	1	4	3	4	4	2	4	2	3	3.00	0.91
SUS **Q7**	4	3	3	4	0	4	4	4	3	2	3	3	3	3.08	1.12
SUS **Q8**	3	1	1	3	0	3	4	4	3	1	4	2	3	2.46	1.33
SUS **Q9**	2	3	2	3	0	4	3	4	4	2	3	3	3	2.77	1.09
SUS **Q10**	4	4	3	3	1	4	3	4	4	2	4	3	3	3.23	0.93
Total score	31	29	27	33	3	38	33	40	37	16	38	24	33	29.38	10.29
SUS score	77.5	72.5	67.5	82.5	7.5	95.0	82.5	100.0	92.5	40.0	95.0	60.0	82.5	73.46	25.73

Figure 8: The questions included in the usability questionnaire for requesting feedback

Q1: Did you use a mobile or a computer/laptop to access the demo? [mobile: yes / no]
Q2: If you use a computer/laptop, did you explore/try anything other than the term search? [yes / no]
Q3: Did you find answers for all the clinical questions? [yes / no]
Q4: How much time in total did you spend with the demo trying to answer all the clinical questions?
[number minutes aprox.:]
Q5: Would you like other documents (e.g. Trust guidelines or BMJ Best Practice documents) with the information accessible in the same way? [yes / no]

C. Exemplifying how a UK junior doctor may answer a clinical question

Let's illustrate (Figure 9) how a UK junior doctor may use the SemDeep demo to answer the clinical question: "Why iStent is recommended for patients with open-angle glaucoma?". The UK junior may look for "iStent" considering the disease "open-angle glaucoma" as context.

Table 7
The responses from 13 junior doctors to question Q1 to Q5 in Textbox 8

Question	COVID-19 demo: yes	COVID-19 demo: Percent	SemDeep demo: yes	SemDeep demo: Percent
Q1	3	23.1	3	23.1
Q2	7	53.8	5	38.5
Q3	9	69.2	10	76.9
Q4*	12	92.3	13	100.0
Q5	10	76.9	9	69.2

*yes = 15 minutes or less

Figure 9: Answering "Why iStent is recommended for patients with open-angle glaucoma?"