

# SUPPLEMENTARY MATERIALS & ANNEXES

---

SUPPLEMENTARY MATERIALS .....	2
CHAPTER II .....	2
<i>Suppl. Table II-1</i> .....	2
CHAPTER III .....	4
<i>Suppl. Note III-1</i> .....	4
<i>Suppl. Note III-2</i> .....	5
<i>Suppl. Note III-3</i> .....	6
<i>Suppl. Note III-4</i> .....	6
<i>Suppl. Table III-1</i> .....	7
<i>Suppl. Table III-2</i> .....	8
<i>Suppl. Table III-3</i> .....	9
<i>Suppl. Table III-4</i> .....	20
CHAPTER IV .....	25
<i>Suppl. Table IV-1</i> .....	25
<i>Suppl. Table IV-2</i> .....	26
<i>Suppl. Figure IV-1</i> .....	28
<i>Suppl. Figure IV-2</i> .....	31
<i>Suppl. Figure IV-3</i> .....	34
<i>Suppl. Figure IV-4</i> .....	37
CHAPTER V .....	38
<i>Suppl. Note V-1</i> .....	38
<i>Suppl. Note V-2</i> .....	38
<i>Suppl. Table V-1</i> .....	42
<i>Suppl. Figure V-1</i> .....	45
<i>Suppl. Figure V-2</i> .....	48
CHAPTER VI .....	51
<i>Suppl. Note VI-1</i> .....	51
<i>Suppl. Note VI-2</i> .....	51
<i>Suppl. Table VI-1</i> .....	52
<i>Suppl. Table VI-2</i> .....	54
<i>Suppl. Table VI-3</i> .....	55
<i>Suppl. Table VI-4</i> .....	56
SUPPLEMENTARY FILES .....	58
LIST OF ANNEXES .....	59

# Supplementary materials

---

## Chapter II

Code	Total	Code	Total
Team as ressource - drawing in expertise	19	Evidence-based ressources	3
Lack of experience	14	Failure to contextualise	3
Assessing severity	13	Feedback	3
Experience	13	Focusing only on the usual	3
Observations	13	Guidelines	3
Anticipation	12	Immediate action	3
History	12	Information or cognitive overload	3
Trajectories	12	Integrate prior	3
Examination	11	Integration of information	3
Expected clinical course	11	Interventions	3
Labs	11	Jumping to conclusion or action	3
Organisational factors	11	Lack of surgical exposure	3
Task fixation - tunnel vision	11	Look outside the box	3
Likelihood - logical thinking	10	Missed detail	3
Background monitoring	9	No consensus between surgeons	3
Give direction - recommend investigations	9	Patient self-reporting	3
Imaging	9	Re-assessment	3
Surgical history	9	Responsibility for patient	3
Defined competency	8	Simulate experience	3
Escalation	8	track progress	3
Flag up changes	8	Adapt to audience	2
Information gathering	8	Awareness of available ressources	2
Provide differential	8	Communication with patient	2
Threshold for action	8	Could add to complexity	2
A to E assessment	7	Data masking reality	2
Differential diagnose	7	Differentiate signal to noise	2
Everyone is a senior	7	Errors needed for learning	2
Fear of the system	7	Fear of missing something	2
Flag up missed items	7	Focusing on one timepoint	2
Knowledge and guidance	7	Identify definitive care	2
Severity assessment - Backing for escalation	7	Inform about norm	2
Intuition and heuristic	6	Insight into the operation	2
Lack of senior support	6	Junior at the frontline	2
Management plan	6	No or ambiguous feedback	2
Safety - Stabilise first	6	Not always consensus between specialties	2
Checklist	5	Not prepared by med school	2
Communication and handover	5	Not that hard	2

## Supplementary materials

Expectation	5	Patient variability	2
Flag singularity	5	Planning support	2
Lack of feedback	5	Prioritisation	2
Lack of staff ressources	5	Prioritisation - triage	2
Parallelisation	5	Remote access to information	2
Patient note	5	Take more time	2
Prioritisation	5	Taking things for granted	2
Rule out approach	5	Technological blocks	2
Silent start of deterioration	5	Time pressure	2
Stress about situation	5	Treatment challenge	2
Anchoring biases	4	Ward specialty and competency	2
Broaden vision	4	Alarm fatigue	1
Correlation investigation-clinic	4	Cannot replace examination	1
Evolution prediction	4	Ceiling of care	1
Finding the overall pattern	4	Clinician pysical factors	1
Focusing on non essential tasks	4	Commitment	1
Further investigations	4	Cumbersome display	1
Handover	4	False positive	1
Material ressources available	4	Fear of consequences	1
Not anticipating	4	Lack of teaching	1
Perseverating in the wrong direction	4	Must be customisable	1
Restarting from beginning	4	Provide feedback	1
Seniority difference	4	Reinforcement bubble	1
Anchoring bias	3	Resistance from senior staff	1
Checklists	3	Responsibility	1
Colloquium	3	slow to adapt to changes	1
Competing duties	3	Supply delivery	1
Contradicting options	3	Testing without understanding	1
Ego	3	Too prescriptive	1
Electronic records	3		

**Suppl. Table II-1: codes description and frequency in the interviews of the surgeons needs study.** One code can be present more than once in an interview.

## Chapter III

### Suppl. Note III-1: main literature search strategy

- 1: \*Decision Making, Computer-Assisted/
- 2: exp Diagnosis, Computer-Assisted/
- 3: \*Therapy, Computer-Assisted/
- 4: Drug Therapy, Computer-Assisted/
- 5: exp Decision Support Systems, Clinical/
- 6: \*Algorithms/
- 7: (CDSS\* or CCDSS\* or "decision support" or "decision making" or "diagnos\* support" or "computer aided" or CAD\* or "computer assisted" or "digital assistance" or algorithm\*).ab,kw,ti.
- 8: 1 or 2 or 3 or 4 or 5 or 6 or 7
- 9: exp Artificial Intelligence/
- 10: exp Latent Class Analysis/
- 11: exp Pattern Recognition, Automated/
- 12: ("artificial intelligence" or AI or "machine learning" or "deep learning" or "neural network" or "support vector machine" or "Bayesian network" or "nearest neighbour" or "decision tree" or "random forest" or "patient similarity" or "pattern recognition" or "natural language processing" or (supervised adj2 learning) or (unsupervised adj2 learning) or "reinforcement learning").ab,kw,ti.
- 13: 9 or 10 or 11 or 12
- 14: (doctor\* or residen\* or physician\* or clinician\* or surgeon\* or registrar\* or "house officer\*" or fellow\* or medics or consultant\* or attending or practitioner\* or oncologist\* or pathologist\* or radiologist\* or ophthalmologist\* or neurologist\* or cardiologist\* or urologist\* or gynecologist\* or gastroenterologist\* or pneumologist\* or dermatologist\* or endocrinologist\* or psychiatrist\* or pediatrician\* or internist\* or anesthesiologist\* or orthopedist\*).ab,kw,ti.
- 15: (safety or trust or usability or confidence or reliability or performance or outperform\* or metrics or measure\* or evaluat\* or assess\* or effective\* or precision or recall or accuracy or "patient\* outcome\*" or "clinical outcome\*" or "surgical outcome\*" or "term outcome\*" or mortality or morbidity or complication\*).ab,kw,ti.
- 16: 8 and 13 and 14 and 15
- 17: limit 16 to (editorial or letter or "review" or "systematic review")
- 18: 16 not 17

**Suppl. Note III-2: grey literature search strategy for conference abstracts**

- 1:     \*Decision Making, Computer-Assisted/
- 2:     exp Diagnosis, Computer-Assisted/
- 3:     \*Therapy, Computer-Assisted/
- 4:     Drug Therapy, Computer-Assisted/
- 5:     exp Decision Support Systems, Clinical/
- 6:     \*Algorithms/
- 7:     (CDSS\* or CCDSS\* or "decision support" or "decision making" or "diagnos\* support" or "computer aided" or CAD\* or "computer assisted" or "digital assistance" or algorithm\*).ab,kw,ti.
- 8:     1 or 2 or 3 or 4 or 5 or 6 or 7
- 9:     exp Artificial Intelligence/
- 10:    exp Latent Class Analysis/
- 11:    exp Pattern Recognition, Automated/
- 12:    ("artificial intelligence" or AI or "machine learning" or "deep learning" or "neural network" or "support vector machine" or "Bayesian network" or "nearest neighbour" or "decision tree" or "random forest" or "patient similarity" or "pattern recognition" or "natural language processing" or (supervised adj2 learning) or (unsupervised adj2 learning) or "reinforcement learning").ab,kw,ti.
- 13:    9 or 10 or 11 or 12
- 14:    (doctor\* or residen\* or physician\* or clinician\* or surgeon\* or registrar\* or "house officer\*" or fellow\* or medics or consultant\* or attending or practitioner\* or oncologist\* or pathologist\* or radiologist\* or ophthalmologist\* or neurologist\* or cardiologist\* or urologist\* or gynecologist\* or gastroenterologist\* or pneumologist\* or dermatologist\* or endocrinologist\* or psychiatrist\* or pediatrician\* or internist\* or anesthesiologist\* or orthopedist\*).ab,kw,ti.
- 15:    (safety or trust or usability or confidence or reliability or performance or outperform\* or metrics or measure\* or evaluat\* or assess\* or effective\* or precision or recall or accuracy or "patient\* outcome\*" or "clinical outcome\*" or "surgical outcome\*" or "term outcome\*" or mortality or morbidity or complication\*).ab,kw,ti.
- 16:    8 and 13 and 14 and 15
- 17:    limit 16 to (editorial or letter or "review" or "systematic review")
- 18:    16 not 17
- 19:    limit 18 to (conference abstracts and yr = "2017-2019")

**Suppl. Note III-3: grey literature search strategy for the Cochrane Central Register of Controlled Trials (CENTRAL)**

- 1: MeSH descriptor: [Decision Making, Computer-Assisted] this term only
- 2: MeSH descriptor: [Diagnosis, Computer-Assisted] explode all trees
- 3: MeSH descriptor: [Therapy, Computer-Assisted] this term only
- 4: MeSH descriptor: [Drug Therapy, Computer-Assisted] explode all trees
- 5: MeSH descriptor: [Decision Support Systems, Clinical] explode all trees
- 6: MeSH descriptor: [Algorithms] this term only
- 7: CDSS\* or CCDSS\* or "decision support" or "decision making" or "diagnos\* support" or "computer aided" or CAD\* or "computer assisted" or "digital assistance" or algorithm\*
- 8: #1 or #2 or #3 or #4 or #5 or #6 or #7
- 9: MeSH descriptor: [Artificial Intelligence] explode all trees
- 10: MeSH descriptor: [Latent Class Analysis] explode all trees
- 11: MeSH descriptor: [Pattern Recognition, Automated] explode all trees
- 12: "artificial intelligence" or AI or "machine learning" or "deep learning" or "neural network" or "support vector machine" or "Bayesian network" or "nearest neighbour" or "decision tree" or "random forest" or "patient similarity" or "pattern recognition" or "natural language processing" or (supervised adj2 learning) or (unsupervised adj2 learning) or "reinforcement learning"
- 13: #9 or #10 or #11 or #12
- 14: doctor\* or residen\* or physician\* or clinician\* or surgeon\* or registrar\* or "house officer\*" or fellow\* or medics or consultant\* or attending or practitioner\* or oncologist\* or pathologist\* or radiologist\* or ophthalmologist\* or neurologist\* or cardiologist\* or urologist\* or gynecologist\* or gastroenterologist\* or pneumologist\* or dermatologist\* or endocrinologist\* or psychiatrist\* or pediatrician\* or internist\* or anesthesiologist\* or orthopedist\*
- 15: safety or trust or usability or confidence or reliability or performance or outperform\* or metrics or measure\* or evaluat\* or assess\* or effective\* or precision or recall or accuracy or "patient\* outcome\*" or "clinical outcome\*" or "surgical outcome\*" or "term outcome\*" or mortality or morbidity or complication\*
- 16: #8 and #13 and #14 and #15
- 17: limit #16 to date from Jan 2010 to May 2019

**Suppl. Note III-4: grey literature search strategy for the World Health Organization (WHO) International Clinical Trials Registry Platform (ICTRP)**

artificial intelligence and CDSS or artificial intelligence and decision support or artificial intelligence and CAD or machine learning and CDSS or machine learning and decision support or machine learning and CAD or deep learning and CDSS or deep learning and decision support or deep learning and CAD or algorithm\* and CDSS or algorithm\* and decision support or algorithm\* and CAD.

# Supplementary materials

	Patient selection			Index test			Reference			Flow and timing											
	Selection criteria clearly described?	Consecutive or random sample enrolled?	Avoid inappropriate exclusions?	RoB subscore	Index test blindly interpreted?	Operators with appropriate training?	Technology of the index test unchanged?	Same clinical data?	RoB subscore	Same reference standard?	Described in sufficient detail?	Likely to correctly classify?		Reference blindly interpreted?	RoB subscore	All received a reference standard?	All receive the same reference standard?	Withdrawals from the study explained?	Study free of commercial funding?	All patients included in the analysis?	RoB subscore
	Alissa	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
	Aslantas	yes	unclear	yes	yes	unclear	unclear	unclear	unclear	yes	yes	unclear	yes	yes	yes	yes	yes	N/A	yes	yes	yes
	Bargallo	yes	yes	yes	yes	yes	unclear	yes	yes	N/A	unclear	yes	yes	unclear	yes	no	yes	unclear	unclear	unclear	0
	Barinov	no	no	unclear	yes	yes	unclear	no	yes	N/A	unclear	yes	yes	unclear	yes	yes	yes	N/A	unclear	unclear	1
	Bartolotta	yes	yes	yes	yes	yes	unclear	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	N/A	unclear	yes	2
	Bien	no	no	unclear	yes	unclear	unclear	unclear	yes	yes	yes	yes	yes	yes	yes	yes	yes	N/A	unclear	yes	2
	Biggelaar	yes	yes	yes	yes	yes	unclear	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	N/A	yes	yes	2
	Blackmon	yes	no	yes	yes	unclear	unclear	unclear	yes	yes	yes	yes	unclear	yes	yes	yes	yes	N/A	unclear	yes	2
	Cha	unclear	unclear	yes	yes	unclear	unclear	unclear	yes	yes	yes	yes	yes	yes	yes	yes	yes	N/A	unclear	yes	2
	Chabi	yes	unclear	yes	yes	unclear	unclear	no	0	yes	yes	yes	yes	yes	yes	yes	yes	N/A	unclear	yes	0
	Cho	yes	yes	yes	yes	unclear	unclear	no	0	no	yes	yes	yes	yes	yes	yes	yes	N/A	no	yes	1
	Choi (2018)	yes	yes	yes	yes	yes	unclear	yes	1	no	yes	yes	yes	yes	yes	yes	yes	N/A	yes	yes	2
	Choi (2019)	yes	unclear	yes	yes	unclear	unclear	no	0	no	no	yes	yes	yes	yes	yes	yes	N/A	yes	yes	0
	Cole	yes	no	yes	yes	yes	yes	no	0	yes	yes	yes	yes	yes	yes	yes	yes	unclear	yes	yes	1
	Endo	no	unclear	yes	yes	unclear	unclear	unclear	1	yes	yes	yes	yes	yes	yes	yes	yes	N/A	yes	yes	2
	Engelke	yes	yes	yes	yes	unclear	unclear	unclear	1	yes	no	yes	yes	yes	yes	yes	yes	N/A	unclear	yes	2
	Giannini	yes	no	yes	yes	unclear	unclear	no	0	no	yes	yes	yes	yes	yes	yes	yes	N/A	yes	yes	2
	Gomez	yes	yes	yes	yes	unclear	yes	yes	2	yes	yes	yes	yes	yes	yes	yes	yes	N/A	yes	unclear	1
	Hwang	yes	unclear	yes	yes	unclear	unclear	unclear	1	no	yes	yes	yes	yes	yes	yes	yes	N/A	yes	yes	2
	Lindsey	yes	yes	no	yes	unclear	unclear	unclear	1	yes	no	yes	yes	yes	yes	yes	yes	yes	no	yes	0
	Park	yes	unclear	yes	yes	unclear	unclear	no	0	no	yes	yes	yes	yes	yes	yes	yes	N/A	yes	yes	2
	Rodriguez-Ruiz	yes	no	yes	yes	yes	yes	no	0	no	no	yes	yes	yes	yes	yes	yes	N/A	unclear	yes	1
	Romero	unclear	yes	yes	yes	yes	yes	yes	2	no	unclear	unclear	yes	yes	yes	yes	yes	N/A	yes	yes	2
	Samulski	yes	no	yes	yes	yes	yes	yes	2	yes	yes	unclear	yes	yes	yes	yes	yes	N/A	yes	yes	1
	Sayres	yes	yes	yes	yes	yes	unclear	unclear	1	yes	yes	yes	yes	yes	yes	yes	yes	N/A	yes	yes	2
	Shimauchi	yes	no	yes	yes	yes	unclear	no	0	yes	yes	yes	yes	yes	yes	yes	yes	N/A	yes	yes	1
	Sohns	N/A	N/A	N/A	4	N/A	N/A	N/A	4	N/A	N/A	N/A	N/A	N/A	4	N/A	N/A	N/A	unclear	N/A	4
	Steiner	yes	no	yes	yes	yes	unclear	unclear	1	yes	yes	yes	yes	yes	yes	yes	yes	N/A	yes	yes	1
	Stoffel	yes	no	yes	yes	unclear	yes	no	0	yes	yes	yes	yes	yes	yes	yes	yes	N/A	yes	yes	2
	Sun	yes	yes	yes	yes	unclear	unclear	no	0	unclear	yes	yes	yes	yes	yes	yes	yes	N/A	unclear	yes	2
	Sunwoo	yes	no	yes	yes	unclear	unclear	no	0	yes	yes	yes	yes	yes	yes	yes	yes	N/A	yes	yes	2
	Tang	no	no	unclear	yes	unclear	unclear	unclear	1	yes	yes	yes	yes	yes	yes	yes	yes	N/A	unclear	yes	2
	Taylor	yes	no	unclear	yes	unclear	unclear	unclear	1	yes	yes	yes	yes	yes	yes	yes	yes	N/A	yes	yes	2
	Vassallo	yes	yes	yes	yes	unclear	unclear	unclear	1	yes	yes	yes	unclear	no	0	yes	yes	N/A	yes	yes	2
	Wanatabe	yes	yes	yes	yes	unclear	unclear	unclear	1	yes	yes	yes	yes	yes	yes	yes	yes	N/A	yes	yes	1
	Way	yes	no	yes	yes	unclear	unclear	unclear	0	no	yes	yes	yes	yes	yes	yes	yes	N/A	yes	yes	2
	Zhang	yes	no	yes	yes	unclear	unclear	unclear	1	no	yes	yes	yes	yes	yes	yes	yes	N/A	yes	yes	2

Suppl. Table III-1: QUADAS detailed scores after conflict resolution. RoB = risk of bias; 0 = high, 1 = unclear, 2 = low; N/A = not applicable

# Supplementary materials

		Aissa Aslantas Bargallo Barinov Bartolotta Bien Biggelaar Blackmon Cha Chabi Cho Cho (2018) Cho (2019) Cole Endo Engelke Giannini Gomez Hwang Lindsey Park Rodriguez-Ruiz Romero Samulski Sayres Shimauchi Sohns Steiner Stoffel Sun Sunwoo Tang Taylor Vassallo Wanatabe Way Zhang																											
Confounders and co-interventions	General confounding domains	none																											
	General co-interventions	none none likely <sup>a</sup> none none none none none none none none none none none none none none none unlikely none																											



## Supplementary materials

First author	Year	Gold standard comparison	Subgroup*	Outcome	Human alone performance	Assisted human performance	Statistical significance
Aissa	2018	1 radiologist (detection) 2 radiologists (follow up) all also study subjects	all participants together (3)	number of solid nodules detected	326	458	yes
				number of true positive nodules detected	326	418	yes
				number of ground glass opacities detected	25	8	yes
Aslantas	2016	one experienced physician	all participants together (1)	accuracy in %	95.38	96.9	NA
				sensitivity in %	97.95	98	NA
				specificity in %	87.5	90.6	NA
Bargallo	2014	positive biopsy results for the positive cases, no information for the negative cases	all participants together (9 without CDSS, 4 with CDSS)	recall rate in %	3.94	7.02	NA
				biopsy rate in %	0.9	1.02	NA
				cancer detection rate in %	5.25	6.1	NA
				PPV of recall in %	13.32	8.69	NA
				breast cancer stage at diagnosis	0: 25.3%, I: 52.6%, II: 15.8%, III: 3.2%, IV: 1.6%, NA: 0.8%	0: 21.5%, I: 55.4%, II: 17.7%, III: 3.1%, IV: 0%, NA: 2.3%	NA
Barinov	2019	pathology results after biopsy or 1 year follow-up	1 radiologist with 20+ years experience	AUROC (second reader)	0.76	0.79	no
				AUROC (first reader)	0.76	0.82	yes
				sensitivity in % (first reader, OPS)	97.5	98.2	NA
				specificity in % (first reader, OPS)	62	55	NA
			1 radiologist with 10+ years experience	AUROC (second reader)	0.75	0.77	no
				AUROC (first reader)	0.75	0.83	yes
				sensitivity in % (first reader, OPS)	95.9	98.2	NA
				specificity in % (first reader, OPS)	59	47.5	NA
			1 radiologist with 5+ years experience	AUROC (second reader)	0.73	0.79	no
				AUROC (first reader)	0.73	0.8	yes
				sensitivity in % (first reader, OPS)	92.4	97	NA
				specificity in % (first reader, OPS)	54.5	53.5	NA
all participants together (3)	inter-reader variability, Kendall's tau b (second reader)	0.42-0.55	0.56-0.66	yes			
	inter-reader variability, Kendall's tau b (first reader)	0.42-0.55	0.62-0.75	yes			
Bartolotta	2018	core-biopsy or 24 months follow-up	2 radiologists with 20+ years experience	cases correctly classified	257	273	no
				sensitivity in %	91.8	97.5	NA
				specificity in %	81.5	86.5	NA
				PPV in %	77.2	83.2	NA
				NPV in %	93.6	98.1	NA
				number of lesions in each BI-RADS class	NA	NA	no
				AUROC	0.93	0.95	no

## Supplementary materials

			2 radiology residents	res 1 - AUROC	0.85	0.88	yes
				res 2 - AUROC	0.83	0.87	yes
				intra-observer agreement - res 1, kappa	0.69	0.78	NA
				intra-observer agreement - res 2, kappa	0.69	0.81	NA
				inter-observer agreement - baseline, kappa	0.67	0.7	NA
				inter-observer agreement - 3 months, kappa	0.63	0.77	NA
Bien	2018	3 board certified MSK radiologists (consensus)	7 radiologists and 2 orthopedists	sensitivity (abnormality)	0.896	0.916	no
				sensitivity (ACL)	0.914	0.91	no
				sensitivity (meniscus)	0.776	0.831	no
				specificity (abnormality)	0.825	0.851	no
				specificity (ACL)	0.917	0.996	yes
				specificity (meniscus)	0.856	0.849	no
				accuracy (abnormality)	0.883	0.905	no
				accuracy (ACL)	0.916	0.939	no
				accuracy (meniscus)	0.815	0.836	no
				inter-rate reliability (abnormality), kappa	0.571	0.64	NA
				inter-rate reliability (ACL), kappa	0.754	0.84	NA
				inter-rate reliability (meniscus), kappa	0.526	0.621	NA
			the same 7 radiologists	sensitivity (abnormality)	0.905	0.926	no
				sensitivity (ACL)	0.906	0.902	no
				sensitivity (meniscus)	0.82	0.829	no
				specificity (abnormality)	0.844	0.864	no
				specificity (ACL)	0.933	0.977	yes
				specificity (meniscus)	0.882	0.88	no
van den Biggelaar	2010	histopathology after surgery or 1 year follow up	all participants together (2)	accuracy (abnormality)	0.894	0.916	no
				accuracy (ACL)	0.92	0.94	no
				accuracy (meniscus)	0.849	0.846	no
				sensitivity in %	84	84	no
				specificity in %	95	95	no
				PPV in %	45	44	no
Blackmon	2011	3 experienced radiologists using the CAD output (consensus)	all participants together (2)	NPV in %	99	99	no
				diagnostic odd ratio	96	90	no
				number of positive cases	94	96	NA
				sensitivity in % (patient)	84.4	92.2	no
				specificity in % (patient)	92.6	88.3	NA
				sensitivity in % (PEs total)	50	70.6	yes
				PPV in % (PEs total)	80.4	80.8	NA
				PPV in % (patient)	88.6	84.3	NA
				NPV in % (patient)	89.7	94.3	NA
				false positive PEs (per patient)	0.18	0.25	no
				accuracy in % (double detection, patient)	39.2	48.1	yes
				sensitivity in % (double detection, PEs total)	32.8	61.3	yes
				sensitivity in % (double detection, central)	84.6	92.3	no
				sensitivity in % (double detection, lobar)	81.8	90.9	no
				sensitivity in % (double detection, segmental)	28.6	58.9	yes
				sensitivity in % (double detection, subsegmental)	26.3	57.9	yes

## Supplementary materials

Cha	2018	1 radiologist with 32 years experience with access to histopathology of resected bladder	all participants together (12)	AUROC (all)	0.74	0.77	yes
				standard deviation of estimates on the % scale (all)	20.4	17,9	yes
				AUROC (easy cases)	0.81	0.84	NA
				standard deviation of estimates on the % scale (easy cases)	14.7	13.4	yes
				AUROC (difficult cases)	0.59	0.62	NA
				standard deviation of estimates on the % scale (difficult cases)	29.1	24.7	yes
Chabi	2012	cytology and/or pathology for all lesions BI-RADS >2	1 radiologist with 20 years experience	sensitivity in % (benign/malignant)	99	99	no
				specificity in % (benign/malignant)	70	46	yes
				sensitivity in % (BI-RADS >=4)	100	100	no
				specificity in % (BI-RADS >=4)	48	31	yes
			1 radiologist with 5 years experience	sensitivity in % (benign/malignant)	87	96	yes
				specificity in % (benign/malignant)	80	58	yes
				sensitivity in % (BI-RADS >=4)	87	96	yes
				specificity in % (BI-RADS >=4)	80	58	yes
			1 radiologist with 1 year experience	sensitivity in % (benign/malignant)	88	95	no
				specificity in % (benign/malignant)	69	57	no
				sensitivity in % (BI-RADS >=4)	97	95	no
				specificity in % (BI-RADS >=4)	34	54	yes
Cho	2017	histopathology after needle biopsy, operative resection or 2 years follow up	1 radiologist with 7 year experience	sensitivity in %	94.4	87	no
				specificity in %	49.2	86.2	yes
				PPV in %	60.7	83.9	yes
				NPV in %	91.4	88.9	no
				accuracy in %	69.8	86.6	yes
				AUROC	0.887	0.895	no
			1 radiologist with 1 year experience	sensitivity in %	94.4	83.3	yes
				specificity in %	55.4	87.7	yes
				PPV in %	63.8	84.9	yes
				NPV in %	92.3	86.4	no
				accuracy in %	73.1	85.7	yes
				AUROC	0.901	0.901	no
Choi J.-H.	2018	histopathology or 2 years follow up	2 radiologists with 5 years experience	sensitivity in %	91.7	91.7	no
				specificity in %	76.6	80.3	no
				PPV in %	20	22	yes
				NPV in %	99.3	99.3	no
				accuracy in %	77	81	no
				AUROC	0.84	0.86	no
			2 radiologists with 1 week of training	sensitivity in %	75	83.3	no
				specificity in %	71.8	77.1	no
				PPV in %	14.5	18.9	yes
				NPV in %	97.8	98.6	no
				accuracy in %	72	77.5	no
				AUROC	0.73	0.8	no

## Supplementary materials

			all participants together (4)	inter-observer agreement, kappa	0.337	0.457	yes
Choi J. S.	2019	histopathology for all lesions BI-RADS >3, or 3 with palpable mass, or growing, or on patient request, stable imaging in follow up for the rest	2 radiologists with 11 and 3 years experience	Rad 1 - sensitivity in %	88.8	86.3	no
				Rad 1 - specificity in %	72.8	93.1	yes
				Rad 1 - accuracy in %	77.9	90.9	yes
				Rad 1 - PPV in %	60.2	85.2	yes
				Rad 1 - NPV in %	93.3	93.6	no
				Rad 1 - AUROC	0.884	0.919	yes
				Rad 2 - sensitivity in %	86.3	90	no
				Rad 2 - specificity in %	83.2	90.2	yes
				Rad 2 - accuracy in %	84.2	90.1	yes
				Rad 2 - PPV in %	70.4	80.1	yes
				Rad 2 - NPV in %	92.9	95.1	no
				Rad 2 - AUROC	0.919	0.942	yes
			2 radiologists with <1 year of training	Rad 3 - sensitivity in %	88.8	95	no
				Rad 3 - specificity in %	75.1	82.1	yes
				Rad 3 - accuracy in %	79.4	86.2	yes
				Rad 3 - PPV in %	62.3	71	yes
				Rad 3 - NPV in %	93.5	97.3	no
				Rad 3 - AUROC	0.906	0.951	yes
				Rad 4 - sensitivity in %	81.3	86.3	no
				Rad 4 - specificity in %	92.5	89	yes
				Rad 4 - accuracy in %	88.9	88.1	yes
				Rad 4 - PPV in %	83.3	78.4	yes
				Rad 4 - NPV in %	91.4	93.3	no
				Rad 4 AUROC	0.895	0.914	yes
			all participants together (4)	sensitivity in %	81.3-88.8	86.3-95.0	no
				specificity in %	72.8-92.5	82.1-93.1	yes
				accuracy in %	77.9-88.9	86.2-90.9	yes
				PPV in %	60.2-83.3	71.0-85.2	yes
				NPV in %	91.4-93.5	93.3-97.3	no
				AUROC	0.884-0.919	0.914-0.951	yes
				interobserver agreement, kappa	0.538-0.706	0.632-0.788	NA
Cole	2014	histopathology or 1 years follow up	all participants together (Image Checker <sup>†</sup> , 15)	AUROC	0.71	0.72	no
				sensitivity in %	51	53	no
				specificity in %	87	86	no
			all participants together (SecondLook <sup>†</sup> , 14)	AUROC	0.71	0.72	no
				specificity in %	89	87	no
Endo	2012	histopathology after surgery or 2 years follow up for benign lesions	2 lung specialists	accuracy in % (average)	76.7	85	NA
			1 radiologist	accuracy in %	80	93.3	NA
			all participants together (3)	accuracy in % (average)	74.4	76.7	NA
Engelke	2010	2 experienced radiologists (consensus)	1 "experienced" radiologist	percentual pulmonary embolism severity index	26.75	27.14	yes
				percentual scoring errors	4.9	3.2	yes
				correct stratifications	55	56	NA
				overestimates	0	0	NA
				underestimates	3	2	NA

## Supplementary materials

			1 "experienced" radiologist	percentual pulmonary embolism severity index	25.85	27.04	yes
				percentual scoring errors	6	4	yes
				correct stratifications	55	56	NA
				overestimates	0	0	NA
				underestimates	3	2	NA
			1 "inexperienced" radiologist	percentual pulmonary embolism severity index	20.63	23.33	yes
				percentual scoring errors	37.9	27.2	yes
				correct stratifications	42	51	NA
				overestimates	0	0	NA
				underestimates	16	7	NA
			1 "inexperienced" radiologist	percentual pulmonary embolism severity index	21.2	23.24	yes
				percentual scoring errors	31.9	28.1	yes
				correct stratifications	44	49	NA
				overestimates	0	0	NA
				underestimates	14	9	NA
			2 "experienced" radiologists	blant & Altman interobserver limits of agreement	-5.45 to 3.03	-3.67 to 2.03	NA
			2 "inexperienced" radiologists	blant & Altman interobserver limits of agreement	-19.71 to 7.47	-9.49 to 5.35	NA
Giannini	2017	biopsy or PSA follow up	all participants together (4)	sensitivity in %	77-93	84-95	yes (individually)
				interobserver agreement, kappa	0.74	0.83	yes
			all participants together (3)	sensitivity in % (patient)	80.9	87.6	no
				sensitivity in % - GS = 6 (patient)	80.6	80.6	no
				sensitivity in % - GS > 6 (patient)	81.2	91.3	yes
				sensitivity in % - diameter 4-9 mm (patient)	77.8	82.2	no
				sensitivity in % - diameter >10 mm (patient)	80	95	yes
				specificity in % (patient)	75.3	78.4	no
				PPV in % (patient)	68	72.4	no
				NPV in % (patient)	85.9	90.7	no
				sensitivity in % (lesion)	70.9	74.4	no
				sensitivity in % - GS = 6 (lesion)	69.2	71.8	no
				sensitivity in % - GS > 6 (lesion)	71.8	75.6	no
				sensitivity in % - diameter 4-9 mm (lesion)	64.9	57.9	no
				sensitivity in % - diameter >10 mm (lesion)	76.7	90	yes
				reading time in second	220	60	yes
				inter-observer agreement, kappa (patient)	0.55	0.63	no
				inter-observer agreement, kappa (lesion)	0.46	0.57	no
				reader 1 - AUROC	0.84	0.85	no
				reader 2 - AUROC	0.82	0.91	yes
				reader 3 - AUROC	0.84	0.88	no
Hwang	2019	5 board-certified radiologists in each institution with 7-14 years experience and access to CT examinations	5 thoracic radiologists	AUROC	0.932	0.958	yes
				area under the JAFROC	0.907	0.938	yes
				sensitivity (average)	0.876	0.924	yes
				specificity (average)	0.946	0.948	no

## Supplementary materials

			5 board-certified radiologists	AUROC	0.896	0.939	yes
				area under the JAFROC	0.87	0.919	yes
				sensitivity (average)	0.812	0.893	yes
				specificity (average)	0.948	0.948	no
			5 non-radiologists physicians	AUROC	0.814	0.904	yes
				area under the JAFROC	0.781	0.873	yes
				sensitivity (average)	0.699	0.835	yes
				specificity (average)	0.901	0.924	yes
			all participants together (24)	sensitivity in %	82.7	92.5	yes
				specificity in %	87.4	94.1	yes
				Relative reduction in misinterpretation	NA	-47%	NA
Lindsey	2018	label by subspecialized orthopedic surgeons (alone or consensus)					
Park	2019	histopathology after needle biopsy or follow up	1 radiologist with 8-10 years experience	sensitivity in %	85.4	90.2	no
				specificity in %	52.5	66.1	yes
				PPV in %	55.6	64.9	yes
				NPV in %	83.8	90.7	no
				accuracy in %	66	74	yes
				AUROC (based on malignancy score)	0.856	0.907	yes
			1 radiologist with 8-10 years experience	sensitivity in %	92.7	90.2	no
				specificity in %	54.2	66.1	yes
				PPV in %	58.5	64.9	yes
				NPV in %	91.4	90.7	no
				accuracy in %	70	76	no
				AUROC (based on malignancy score)	0.889	0.904	no
			1 first year fellowship trainee	sensitivity in %	65.9	97.6	yes
				specificity in %	27.1	23.7	no
				PPV in %	38.6	47.1	yes
				NPV in %	53.3	93.3	yes
				accuracy in %	43	54	yes
				AUROC (based on malignancy score)	0.623	0.828	yes
			1 first year fellowship trainee	sensitivity in %	75.6	85.4	no
				specificity in %	50.8	66.1	yes
				PPV in %	51.7	63.6	yes
				NPV in %	75	86.7	yes
				accuracy in %	61	74	yes
				AUROC (based on malignancy score)	0.702	0.823	yes
			1 first year fellowship trainee	sensitivity in %	87.8	97.6	no
				specificity in %	27.1	30.5	no
				PPV in %	45.6	49.4	no
				NPV in %	76.2	94.7	yes
				accuracy in %	51	58	no
				AUROC (based on malignancy score)	0.759	0.839	yes
			2 radiologists with 8-10 years experience	interobserver variability, kappa (BI-RADS category)	0.26	0.51	yes
			3 first year fellowship trainees	interobserver variability, kappa (BI-RADS category)	0.186	0.412	yes
			all participants together (5)	interobserver variability, kappa (BI-RADS category)	0.221	0.32	yes

## Supplementary materials

Rodríguez-Ruiz	2019	1 experienced radiologist with access to histopathology or 1 year follow up	all participants together (14)	AUROC	0.87	0.89	yes
				sensitivity in %	83	86	yes
				specificity in %	77	79	no
				reading time in second	146	149	no
			50% most experienced	AUROC	0.87	0.88	no
Romero	2011	not clearly stated, probably biopsy and follow up	all participants together (2)	50% least experienced	AUROC	0.87	yes
				carcinoma detection rate in ‰ (global)	11.9	14.3	no
				carcinoma detection rate in ‰ (screening)	6.1	5.6	no
				carcinoma detection rate in ‰ (diagnostic)	28.8	31.1	no
				% of DCIS in detected cancer (screening)	21.1	36.8	no
				% of DCIS in detected cancer (diagnostic)	16.1	20	no
				detection rate of microcalcification in % (global)	26.3	68.4	yes
				% of T1 tumor (global)	88	79.8	no
				% of T1 tumor (screening)	94.7	84.2	no
				% of T1 tumor (diagnostic)	83.9	78.2	no
				biopsy rate in ‰ (global)	14.7	17.9	no
				biopsy rate in ‰ (screening)	8.3	7.6	no
				biopsy rate in ‰ (diagnostic)	33.4	37.8	no
				biopsy PPV in % (screening)	73.1	69.2	no
Samulski	2010	biopsy for malignant lesions, no information for benign lesions	4 radiologists certified in mammography	biopsy PPV in % (diagnostic)	86.1	82.1	no
				mean correct localization fraction in the false-positive fraction interval ranging from 0 to 0.1	24.9	29.3	NA
			5 non-radiologists physicians experience in mammography	average reading time per case in seconds	70	72.8	yes
				mean correct localization fraction in the false-positive fraction interval ranging from 0 to 0.1	25.2	39.2	NA
			all participants together (9)	average reading time per case in seconds	97	96.7	no
				mean correct localization fraction in the false-positive fraction interval ranging from 0 to 0.1	25.1	34.8	yes
Sanchez Gomez	2011	not clearly stated, probably biopsy and follow up	all participants together (6)	average reading time per case in seconds	84.7	85.9	no
				recall rate in %	7.2	7.6	no
				biopsy rate in %	NA	NA	no
				PPV of biopsy in %	20.23	20.23	no
				sensitivity in %	96.1	97.1	NA
				specificity in %	93.2	92.8	NA
				PPV in %	6.4	6.1	NA
				NPV in %	99.5	99.5	NA
				cancer detection rate in ‰	4.2	4.3	yes
Sayres	2019	3 experienced ophthalmologists (consensus)	5 retina specialists	number of cases detected	93	94	yes
				accuracy in % (grades)	62.3	appr. 70 <sup>†</sup>	NA
			5 general ophthalmologists	accuracy in % (grades + heatmap)	62.3	appr. 66 <sup>†</sup>	NA
				accuracy in % (grades)	46.3	appr. 58 <sup>†</sup>	NA
				accuracy in % (grades + heatmap)	46.3	appr. 56 <sup>†</sup>	NA

## Supplementary materials

			all participants together (10)	sensitivity in % (grades, average)	79.4	87.5	NA
				specificity in % (grades, average)	96.6	96.1	NA
				accuracy in % (grades)	NA	NA	yes
				sensitivity in % (grades + heatmap, average)	79.4	88.7	NA
				specificity in % (grades + heatmap, average)	96.6	95.5	NA
				accuracy in % (grades + heatmap)	NA	NA	no
				accuracy in % - algorithm correct (grades)	91.1	94.4	yes
				accuracy in % - algorithm correct (grades + heatmaps)	91.1	92.6	yes
				accuracy in % - algorithm incorrect (grades)	37	32.4	no
				accuracy in % - algorithm incorrect (grades + heatmaps)	37	33.14	no
				confidence - % cases very or extremely confident (grades)	appr. 72 <sup>¶</sup>	appr. 79 <sup>¶</sup>	yes
				confidence - % cases very or extremely confident (grades + heatmaps)	appr. 72 <sup>¶</sup>	appr. 81 <sup>¶</sup>	yes
Shimauchi	2010	histopathology	all participants together (6)	AUROC	0.8	0.84	yes
				sensitivity in %	83	88	yes
				specificity in %	50	53	no
				PPV in %	66	68	no
				NPV in %	75	83	yes
				PPV in % at 20% prevalence	39	41	NA
				PPV in % at 10% prevalence	28	29	NA
				NPV in % at 20% prevalence	92	95	NA
				NPV in % at 10% prevalence	96	98	NA
Sohns	2010	NA	1 attending	median time in s (early research)	7.47	6.8	NA
				median time in s (benign)	11.12	10.93	NA
				median time in s (malignant)	11.37	11.32	NA
			1 resident	median time in s (early research)	22.6	23.56	NA
				median time in s (benign)	26.53	28.24	NA
				median time in s (malignant)	27.54	30.43	NA
Steiner	2018	3 experienced pathologists (consensus)	all participants together (6)	sensitivity in % (micrometastasis)	83.3	91.2	yes
				sensitivity in % (macrometastasis)	appr. 96 <sup>¶</sup>	appr. 95 <sup>¶</sup>	no
				specificity in %	appr. 99 <sup>¶</sup>	100 <sup>¶</sup>	no
				time to decision in s (negative)	137	111	yes
				time to decision in s (isolated tumor cells)	145	124	no
				time to decision in s (micrometastasis)	117	61	yes
				time to decision in s (macrometastasis)	39	34	no
				subjective "obviousness score" (negative)	67.5	72	no
				subjective "obviousness score" (isolated tumor cells)	55.6	50.4	no
				subjective "obviousness score" (micrometastasis)	63.1	83.6	yes
Stoffel	2018	histopathology	1 radiologist (8y exp.)	AUROC	0.61	0.77	no
			1 radiologist (3y exp.)	AUROC	0.77	0.75	no
			1 radiologist (2y exp.)	AUROC	0.75	0.87	no
			1 radiology resident	AUROC	0.74	0.84	no



## Supplementary materials

Sun	2014	CT scan and successful therapy with Warfarin for 6 months or thrombus found during surgery	2 radiologists described as "senior"	Rad 1 - accuracy	0.883	0.997	NA
				Rad 1 - sensitivity	0.961	0.968	NA
				Rad 1 - specificity	0.859	0.98	NA
				Rad 1 - PPV	0.68	0.938	NA
				Rad 1 - NPV	0.986	0.99	NA
				Rad 1 - AUROC	0.854	0.943	yes
				Rad 2 - accuracy	0.874	0.973	NA
				Rad 2 - sensitivity	0.955	0.984	NA
				Rad 2 - specificity	0.848	0.97	NA
				Rad 2 - PPV	0.664	0.91	NA
				Rad 2 - NPV	0.984	0.995	NA
				Rad 2 - AUROC	0.848	0.942	yes
				Rad 3 - accuracy	0.865	0.969	NA
				Rad 3 - sensitivity	0.935	0.952	NA
				Rad 3 - specificity	0.842	0.975	NA
				Rad 3 - PPV	0.65	0.922	NA
				Rad 3 - NPV	0.977	0.985	NA
				Rad 3 - AUROC	0.827	0.936	NA
			2 radiologists described as "junior"	Rad 4 - accuracy	0.803	0.962	NA
				Rad 4 - sensitivity	0.916	0.935	NA
				Rad 4 - specificity	0.768	0.97	NA
				Rad 4 - PPV	0.553	0.906	NA
				Rad 4 - NPV	0.967	0.98	NA
				Rad 4 - AUROC	0.819	0.88	NA
				Rad 5 - accuracy	0.775	0.95	NA
				Rad 5 - sensitivity	0.897	0.935	NA
				Rad 5 - specificity	0.737	0.955	NA
				Rad 5 - PPV	0.517	0.866	NA
				Rad 5 - NPV	0.958	0.979	NA
				Rad 5 - AUROC	0.821	0.86	yes
			all participants together (5)	accuracy	0.84	0.966	yes
				sensitivity	0.933	0.955	yes
				specificity	0.811	0.97	yes
				PPV	0.613	0.908	yes
				NPV	0.974	0.986	yes
				AUROC	0.834	0.932	yes
Sunwoo	2017	2 experienced neuroradiologists with access to follow up studies (consensus)	2 board-certified neuroradiologists	sensitivity in % (patient)	87.3	88.7	no
				false positive per patient	0.25	0.25	NA
				reading time in s	121	57.3	NA
			2 radiology residents	sensitivity in % (patient)	67.9	76.1	yes
				false positive per patient	0.1	0.12	NA
				reading time in s	97.5	64.8	NA
			all participants together (4)	sensitivity in % (patient)	77.6	81.9	NA
				false positive per patient	0.18	0.18	NA
				reading time in s	114	72	NA
				FOM	0.87	0.9	yes

## Supplementary materials

				failure to detect at least one nodule, in % of positive cases	6.7	4.2	NA
				detection of at least one FP, in % of negative cases	5.0	4.2	NA
				accuracy in % (patient)	94.2	95.8	NA
Tang	2011	2 experienced radiologists with 10+ years experience (consensus)	2 radiologists	AUROC	0.998	0.999	NA
			2 radiology residents	AUROC	0.965	0.99	NA
			2 emergency physicians	AUROC	0.879	0.942	NA
Taylor	2018	2 experienced neurologists with access to follow-up data (local), PPMI core lab team (PPMI)	1 radiologist with 5+ years experience	sensitivity (local)	appr. 0.94 <sup>¶</sup>	appr. 0.94 <sup>¶</sup>	NA
				specificity (local)	appr. 0.91 <sup>¶</sup>	appr. 0.91 <sup>¶</sup>	NA
				accuracy (local)	appr. 0.93 <sup>¶</sup>	appr. 0.93 <sup>¶</sup>	NA
				sensitivity (PPMI)	appr. 0.90 <sup>¶</sup>	appr. 0.93 <sup>¶</sup>	NA
				specificity (PPMI)	appr. 0.85 <sup>¶</sup>	appr. 0.93 <sup>¶</sup>	NA
				accuracy (PPMI)	appr. 0.88 <sup>¶</sup>	appr. 0.93 <sup>¶</sup>	NA
			1 radiologist with 5+ years experience	sensitivity (local)	appr. 0.97 <sup>¶</sup>	appr. 0.94 <sup>¶</sup>	NA
				specificity (local)	appr. 0.82 <sup>¶</sup>	appr. 0.86 <sup>¶</sup>	NA
				accuracy (local)	appr. 0.91 <sup>¶</sup>	appr. 0.91 <sup>¶</sup>	NA
				sensitivity (PPMI)	appr. 0.85 <sup>¶</sup>	appr. 0.95 <sup>¶</sup>	NA
				specificity (PPMI)	appr. 0.90 <sup>¶</sup>	appr. 0.93 <sup>¶</sup>	NA
				accuracy (PPMI)	appr. 0.87 <sup>¶</sup>	appr. 0.94 <sup>¶</sup>	NA
			all participants together (2)	inter-observer reliability (local)	appr. 0.92 <sup>¶</sup>	appr. 0.96 <sup>¶</sup>	NA
				inter-observer reliability (PPMI)	appr. 0.91 <sup>¶</sup>	appr. 0.98 <sup>¶</sup>	yes
Vassallo	2018	2 experienced radiologists with access to follow up record if needed	all participants together (3)	sensitivity in % (nodule)	65	88	yes
				sensitivity in % (patient)	75	82	yes
				specificity in % (patient)	85	82	no
				reading time in s	296	329	yes
Wanatabe	2019	2 experts mammographers with access to biopsy results (consensus)	3 mammography fellowship trained radiologists	cancer detection rate in % (average)	58.5	63.75	NA
				number of false positive recall (average)	8	7.5	NA
				number of calcification recall (average)	8.5	10.25	NA
				number of discarded computer flag (calcification)	NA	6.75	NA
				number of mass recall (average)	44	47	NA
				number of discarded computer flag (mass)	NA	10.25	NA
			4 general radiologists	cancer detection rate in % (average)	40.3	59	NA
				number of false positive recall (average)	5.7	7	NA
				number of calcification recall (average)	6	11.7	NA
				number of discarded computer flag (calcification)	NA	5.3	NA
				number of mass recall (average)	30.7	41.7	NA
				number of discarded computer flag (mass)	NA	13	NA
			all participants together (7)	cancer detection rate in % (average)	51	62	NA
				number of false positive recall (average)	7	7.3	NA
				number of calcification recall (average)	7.4	10.8	NA
				Number discarded computer flag (calcification, average)	NA	6.1	NA
				number of mass recall (average)	38	44.4	NA
				number of discarded computer flag (mass, average)	NA	11	NA
				AUROC (combined readers, case scoring)	0.76	0.81	yes

## Supplementary materials

Way	2010	biopsy, other known metastatic diseases or 2 years follow up	all participants together (6)	AUROC	0.833	0.853	yes
				AUROC with true positive fraction > 0.9	0.39	0.456	yes
				AUROC (primary cancer VS benign)	0.823	0.848	yes
				AUROC with true positive fraction > 0.9 (primary cancer VS benign)	0.338	0.415	yes
				AUROC (metastatic cancer VS benign)	0.849	0.861	no
				AUROC with true positive fraction > 0.9 (metastatic cancer VS benign subset)	0.493	0.535	yes
				change in recommended action	NA	NA	no
Zhang	2016	biopsy or 6 months follow up	5 expert radiologists	AUROC	0.843	0.896	yes
				sensitivity in %	83.5	88.8	yes
				specificity in %	75.6	76	no
				inter-observer agreement, kappa	0.36	0.457	yes
				agreement on management recommendations, cases	34	45	NA
				number of correct recommendations	27	37	NA
				mean reading time in s	16.8	21.2	yes
			5 radiology residents	AUROC	0.705	0.822	yes
				sensitivity in %	63.2	80.7	yes
				specificity in %	62.6	72.9	yes
				inter-observer agreement, kappa	0.151	0.413	yes
				agreement on management recommendations, cases	20	41	NA
				number of correct recommendations	12	30	NA
				mean reading time in s	24.5	30.6	yes
			all participants together (10)	AUROC	0.774	0.859	yes
				sensitivity in %	73.3	84.7	yes
				specificity in %	69.1	74.5	yes
				agreement on management recommendations, cases	15	31	NA
				number of correct recommendations	10	28	NA
				inter-observer agreement, kappa	0.195	0.421	yes

**Suppl. Table III-3: complete list of the included studies' results for the primary outcome.** \*see supplementary table 8 for complete description of the study participants' level of experience, ¶estimated from graphics, †commercial name, NA = not available, AUROC = area under the receiver operating characteristic curve, OPS = operating point shift, PPV = positive predictive value, NPV = negative predictive value, ACL = anterior cruciate ligament, PE = pulmonary embolism, BI-RADS = breast imaging-reporting and data system, GS = Gleason score, JAFROC = jackknife free-response receiver operating characteristic, DCIS = ductal carcinoma in situ, FP = false positive, FOM = figure of merit, PPMI = Parkinson's Progression Markers Initiative

## Supplementary materials

First author	Year	Task to be performed	Description of the CDSS' support	Attempt to increase the interpretability of the CDSS' outputs	Level of experience of the study participants	Clinicians' familiarity with the system	Attempt to gather user feedback on the system
Aissa	2018	identification (lung nodules, ground glass opacities)	marking of suspicious lesions	NA	3 resident/board-certified radiologists with 5-6 years experience	NA	NA
Aslantas	2016	classification (metastasis, no metastasis)	hotspots marking with multiple colours scale 0 (no metastasis) / 1 (metastasis) classification	heatmap	1 non-specified doctor	NA	NA
Bargallo	2014	classification (normal, recallable)	marking of suspicious lesions, shape depending on lesion characteristics	NA	4 radiologists with and without breast unit experience	one-month familiarisation period	NA
Barinov	2019	classification (BI-RADS) and score assignment (likelihood of malignancy)	4 groups classification (benign, probably benign, suspicious, malign)	NA	1 ABR certified and breast fellowship trained radiologist with 20+ years experience and 1 ABR certified and breast fellowship trained radiologist with 10 years experience and 1 ABR certified radiologist with 5 years experience	30 min training + 10 practice cases with supervision	NA
Bartolotta	2018	classification (BI-RADS)	2 groups classification (possibly benign, possibly malign)	displaying of the input features extracted/used to generate the recommendation	2 radiologists with 20+ years experience in breast US and 2 4th/5th year resident radiologists	5h training session with 20 practice cases	NA
Bien	2018	classification (normal, abnormal; ACL intact, tear; meniscus intact, tear)	4 groups classification (normal, abnormal, ACL tear, meniscal tear) and probability score	heatmap of the important features	7 board-certified general radiologists and 2 orthopaedic surgeons with together 3-29 years experience	NA	NA
van den Biggelaar	2010	sketch of the lesion and classification (BI-RADS) and prescription (additional diagnostic test)	marking of suspicious lesions, shape depending on lesion characteristics	NA	2 radiologists with 5 and 20 years experience in mammography	instruction by manufacturer and 9 months optional use	Questionnaires about the added value of the CDSS diagnostic information
Blackmon	2011	identification (suspected PE)	marking of suspicious vessels	NA	2 first year resident radiologists with 9 months experience	NA	NA
Cha	2018	score assignment (likelihood of T0 disease, % response to treatment, grade of lesion conspicuity)	complete response likelihood score	display of the CDSS-T score distribution in a graphic	9 radiologists and 2 oncologists and 1 urologist with together 2-36 years experience	NA	NA

## Supplementary materials

Chabi	2012	classification (benign, malign) and classification (BI-RADS) and score assignment (malignancy score) and characterisation (lesion type)	5 groups classification (BI-RADS 1-5)	NA	1 radiologist with 20 years experience and 1 radiologist with 5 years experience and 1 radiologist with 1 years experience and 1 radiologist with 4 months experience	NA	NA
Cho	2017	classification (BI-RADS)	2 groups classification (possibly benign, possibly malign)	displaying of the input features extracted/used to generate the recommendation	2 radiologists with 7 and 1 years experience in breast imaging	NA	NA
Choi J.-H.	2018	classification (BI-RADS)	2 groups classification (possibly benign, possibly malign)	displaying of the input features extracted/used to generate the recommendation	2 radiologists with 5 years experience in breast imaging and 2 radiologists with 1 week of training in breast imaging	NA	NA
Choi J. S.	2019	classification (BI-RADS)	2 groups classification (possibly benign, possibly malign)	displaying of the input features extracted/used to generate the recommendation	2 radiologists with 11 and 3 years experience in breast imaging and 2 radiologists with <1 year experience in breast imaging	NA	NA
Cole	2014	classification (BI-RADS) and score assignment (DMIST probability of malignancy)	Image Checker: marking of suspicious lesions SecondLook: marking of suspicious lesions, shape depending on lesion characteristics	NA	Image Checker*: 15 radiologists with 6-40 years experience in mammography Secondlook*: 14 radiologists with 3.5-32 years experience in mammography	all selected participants had clinical experience using CAD	NA
Endo	2012	classification (benign, malign)	list of 4 similar cases with diagnosis	NA	1 radiologist with unknown experience and 2 lung specialists radiologists with 7 and 10 years experience	NA	The users were invited to rate the level of similarity of the most similar case on a 1-5 scale
Engelke	2010	score assignment (Mastora risk stratification) and identification (PE)	marking of suspicious vessels	NA	2 "inexperienced" radiologists and 2 "experienced" radiologists	NA	NA
Giannini	2017	characterisation (lesion) and localisation (lesion) and classification (PI-RADS) and classification (prostate carcinoma yes, no) and score assignment (self-confidence for malignancy)	coloured malignancy likelihood map (heatmap)	per voxel malignancy likelihood map	3 radiologists with 2-4 years experience in prostate MRI	NA	NA

## Supplementary materials

Hwang	2019	classification (significant findings requiring treatment, no significant findings)	localisation probability for each disease (heatmap) and overall probability of abnormal findings	per-pixel disease probability per disease visualisation	5 specialised thoracic radiologists with 9-14 years experience and 5 board certified radiologists with 5-7 years experience and 5 non-radiologist physicians of unknown experience	NA	NA
Lindsey	2018	classification (fracture present, not present)	Fracture probability value and dense conditional probability map (heatmap)	per pixel confidence in fracture probability value	24 emergency physicians of unknown experience	NA	NA
Park	2019	classification (BI-RADS) and score assignment (malignancy)	2 groups classification (possibly benign, possibly malign)	displaying of the input features extracted/used to generate the recommendation	2 radiologists with 10 and 8 years experience 3 first year fellowship trainee radiologists	NA	NA
Rodríguez-Ruiz	2019	classification (BI-RADS) and score assignment (malignancy)	marking of suspicious lesions and level of suspicion score (0-100) and Transpara score (0-10)	NA	11 specialised breast radiologists and 3 general radiologists with together 3-25 years experience	45 practice cases	NA
Romero	2011	classification (normal, recallable)	marking of suspicious lesions	NA	2 specialised breast radiologists with 9 and 5 years experience	6 months familiarisation period and 3225 practice cases between the two participants	NA
Samulski	2010	score assignment (malignancy)	coloured-coded circle around the suspicious lesion if ROI is queried and malignancy score	NA	4 radiologists "certified in mammography" and 5 non-radiologists physicians "experienced in mammography"	10 to 60 practice cases per participant	informal feedback after the test
Sanchez Gomez	2011	classification (normal, recallable)	marking of suspicious lesions, shape depending on lesion characteristics	NA	2 general radiologists with 3-9 months experience in mammography and 4 specialised breast radiologists with 2-10 years experience in mammography	NA	NA
Sayres	2019	classification (DR grade) and score assignment (confidence in diagnosis)	Grades of evidence for each diabetic retinopathy category + heatmap in explanatory mode	heatmap highlighting image regions most contributing to the prediction	4 fellowship trained retina specialists and 1 retina fellow and 5 board certified ophthalmologists	briefing about the CDSS	NA

## Supplementary materials

Shimauchi	2010	score assignment (probability of malignancy) and prescription (recommended management)	contours of segmented lesion and graphical representation of estimated probability of malignancy and kinetic curves informing about signal intensity over time and display of most-enhancing regions within given lesion	details about features and probability distribution	2 breast imaging attending radiologists with 18 and 6 years experience and 4 breast imaging fellows radiologists	10 practice cases	NA
Sohns	2010	classification (BI-RADS) and classification (ACR types breast tissue)	marking of suspicious lesions	NA	1 attending physician of unknown specialty and experience and 1 resident physician of unknown specialty and experience	NA	NA
Steiner	2018	classification (negative, isolated tumour cells cluster, micrometastasis, macrometastasis)	heatmap highlighting suspicious regions of interest	NA	6 pathologists with 1-15 years experience	participation in pilot study and 5 practice cases	NA
Stoffel	2018	score assignment (confidence in diagnosis)	system "rating"	NA	1 board certified radiologist with 8 years experience and 1 board certified radiologist with 3 years experience and 1 board certified radiologist with 2 years experience and 1 3rd year radiology resident	NA	NA
Sun	2014	identification (thrombus)	Highlighting of suspicious regions and likelihood score for the presence of a thrombus	NA	3 "senior" radiologists and 2 "junior" radiologists	NA	NA
Sunwoo	2017	identification (metastasis candidates) and score assignment (confidence)	highlighting of suspicious regions and probability score	NA	2 board certified neuroradiologists with 7 years experience and 2 radiology residents with 4 and 2 years experience	NA	NA
Tang	2011	score assignment (confidence in the presence of abnormality)	highlighting of suspicious regions	NA	2 specialised radiologists with 9.5 years experience on average and 2 radiology residents with 6 years experience on average and 2 emergency physicians with 2.5 years experience on average	NA	NA
Taylor	2018	score assignment (confidence in normal findings)	5-points scale (probability of belonging to the disease class)	NA	2 radiologists with more than 5 years experience	NA	Interviews on human-CAD relationship and CAD effects on decision making

## Supplementary materials

Vassallo	2018	identification (lung metastasis)	highlighting of suspicious regions and nodule measurements	NA	3 radiologists with 3-35 years experience	NA	NA
Wanatabe	2019	classification (normal, recallable)	highlighting of suspicious regions and malignancy probability score	NA	3 mammography fellowship trained radiologists with 5-19 years experience and 4 general radiologists with 1-42 years experience	NA	NA
Way	2010	score assignment (likelihood of malignancy) and prescription (recommended management) and characterisation (features description)	0-10 scale (likelihood of malignancy) and class distribution curves	displays the fitted class distribution	6 fellowship-trained thoracic radiologists with 1-8 years post fellowship experience	one training session	NA
Zhang	2016	score assignment (estimated likelihood of malignancy) and prescription (recom. management)	likelihood of malignancy and 10 features distribution in context of the training set	gives details about features in context of the training set	5 expert radiologists with 12-21 years experience in sonography and 5 radiology residents with "limited experience"	30 practice cases	NA

**Table III-4: characteristics relevant to the human factors evaluation of the included studies.** \*commercial name, NA = not available, BI-RADS = breast imaging-reporting and data system, ACL = anterior cruciate ligament, PE = pulmonary embolism, DMIST = Digital Mammographic Imaging Screening Trial, PI-RADS = Prostate Imaging–Reporting and Data System, ROI = region of interest, ACR = American College of Radiology.



## Chapter IV

laboratory variable	missingness (%)	laboratory variable	missingness (%)
Albumin	72.1	Lactate	63.6
Alkaline phosphatase	72.7	Lymphocyte count	66.3
Alanine aminotransferase (ALT)	73.9	Mean corpuscular haemoglobin	66.3
Activated partial thromboplastin time	78.9	Mean corpuscular haemoglobin concentration	66.3
aspartate aminotransferase (AST)	99.9	Mean corpuscular volume	66.3
Basophil count	66.3	Methaemoglobin	66.8
Base excess	63.6	Monocyte count	66.3
Bicarbonate	63.2	Mean platelet volume	66.4
Calcium	63.6	Neutrophil count	66.3
Chloride	66.7	Osmolality	70.0
Creatinine	65.0	pH	63.6
C-reactive protein	80.5	Platelet count	66.3
Corected calcium	88.9	Potassium	49.2
Estimated glomerular filtration rate	66.4	Red blood cell count	66.3
Eosinophil count	66.3	Sodium	49.2
Glucose	52.6	Total bilirubin	73.4
Haematocrit	49.3	Troponine	99.5
Haemoglobin	49.2	Urea (plasma)	66.1
Prothrombin international normalized ratio (INR)	83.1	White blood cell count	66.3
Ketones	97.8		

**Suppl. Table IV-1: percentage of missing data for each of the laboratory values.** The rate of missing data was calculated after merging with point of care testing (POCT) results and after fetching any available value within the last 26 hours.

# Supplementary materials

a.

Layer	Neuron	Latent_4	Latent_8	Latent_16	Latent_32
2	64	0.5601	0.5230	0.4785	0.4161
	128	0.5548	0.5213	0.4711	0.3938
	256	0.5518	0.5295	0.4952	0.4154
	512	0.5606	0.5504	0.5431	0.4637
4	64	0.5601	0.5345	0.5122	0.4936
	128	0.5544	0.5320	0.5099	0.4746
	256	0.5610	0.5470	0.5350	0.5099
	512	0.5664	0.5660	0.5617	0.5528
8	64	0.5612	0.5542	0.5505	0.5488
	128	0.5653	0.5534	0.5515	0.5494
	256	0.5769	0.5634	0.5610	0.5589
	512	0.5879	0.5725	0.5738	0.5693
16	64	0.5985	0.5919	0.5845	0.5841
	128	0.6112	0.5835	0.5827	0.5871
	256	0.6298	0.6163	0.6134	0.6013
	512	0.6370	0.6370	0.6370	0.6370

b.

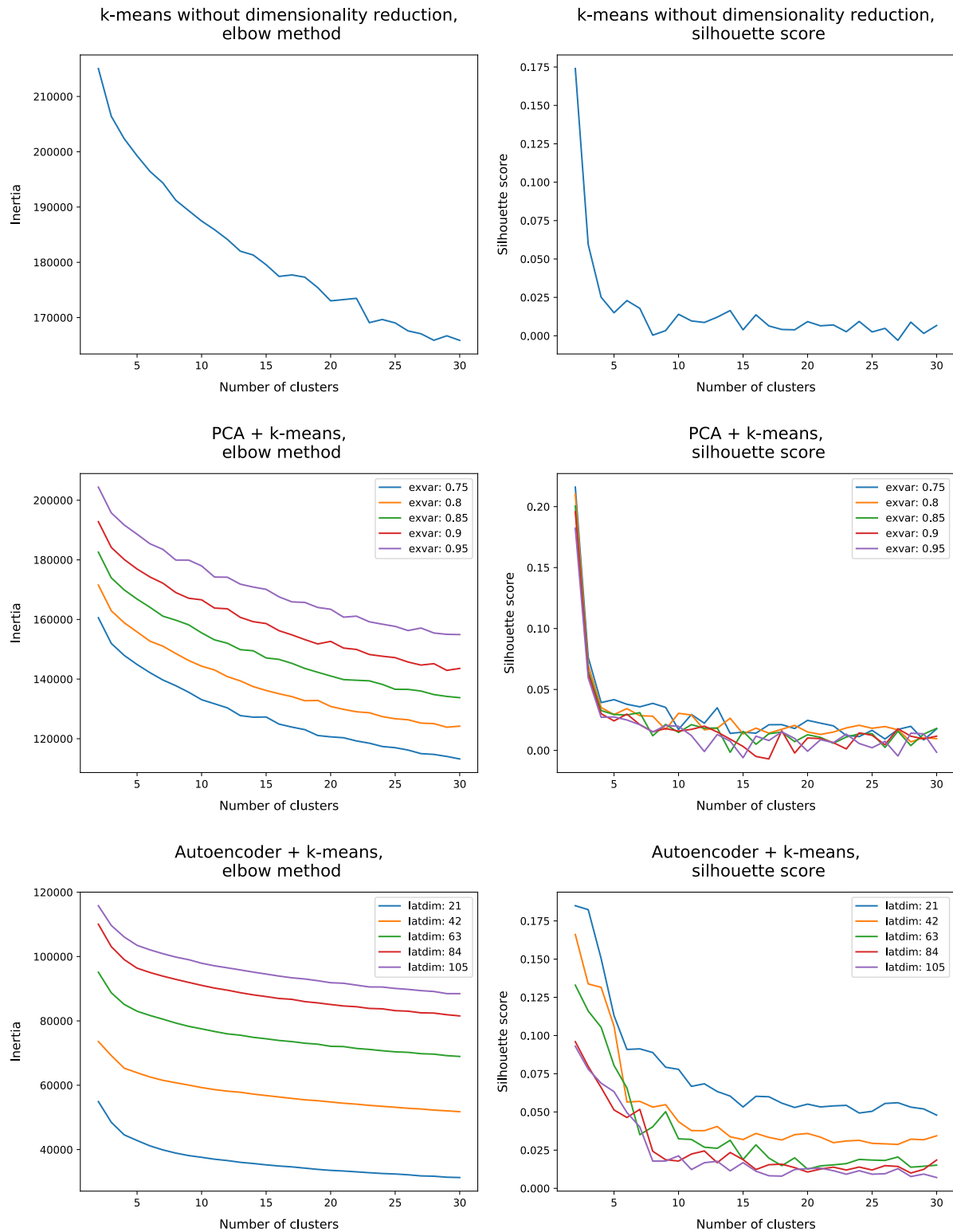
Layer	Neuron	Latent_4	Latent_8	Latent_16	Latent_32
2	64	0.5358	0.4889	0.4346	0.359
	128	0.5357	0.4904	0.4246	0.3347
	256	0.5353	0.4976	0.4443	0.3471
	512	0.5402	0.5197	0.4879	0.3858
4	64	0.5397	0.5032	0.4691	0.4544
	128	0.5334	0.5032	0.4722	0.4284
	256	0.5418	0.5213	0.4955	0.4623
	512	0.5583	0.5454	0.5249	0.5163
8	64	0.5464	0.5404	0.5355	0.5351
	128	0.5717	0.5407	0.5384	0.5351
	256	0.5659	0.5507	0.5436	0.5479
	512	0.5828	0.5702	0.5672	0.562
16	64	0.5973	0.5851	0.5868	0.5817
	128	0.6191	0.5831	0.5816	0.5809
	256	0.6207	0.6321	0.5824	0.6135
	512	0.6445	0.6444	0.6445	0.6444

C.

<i>Layer</i>	<i>Neuron</i>	<i>Latent_4</i>	<i>Latent_8</i>	<i>Latent_16</i>	<i>Latent_32</i>
<b>2</b>	<b>64</b>	0.4108	0.3226	0.185	0.0567
	<b>128</b>	0.399	0.3101	0.1816	0.0651
	<b>256</b>	0.3916	0.3153	0.1932	0.0755
	<b>512</b>	0.3984	0.3246	0.2236	0.0828
<b>4</b>	<b>64</b>	0.4189	0.3375	0.2562	0.1905
	<b>128</b>	0.4086	0.3291	0.2398	0.1683
	<b>256</b>	0.4086	0.3413	0.2776	0.2138
	<b>512</b>	0.4433	0.3686	0.3097	0.294
<b>8</b>	<b>64</b>	0.4576	0.4242	0.4168	0.4126
	<b>128</b>	0.4514	0.4205	0.4128	0.4093
	<b>256</b>	0.4542	0.4376	0.4317	0.4344
	<b>512</b>	0.4993	0.4739	0.4556	0.4429
<b>16</b>	<b>64</b>	0.5349	0.5052	0.505	0.5056
	<b>128</b>	0.5719	0.5168	0.5057	0.5069
	<b>256</b>	0.5876	0.5878	0.5906	0.5552
	<b>512</b>	0.6403	0.6403	0.6402	0.6402

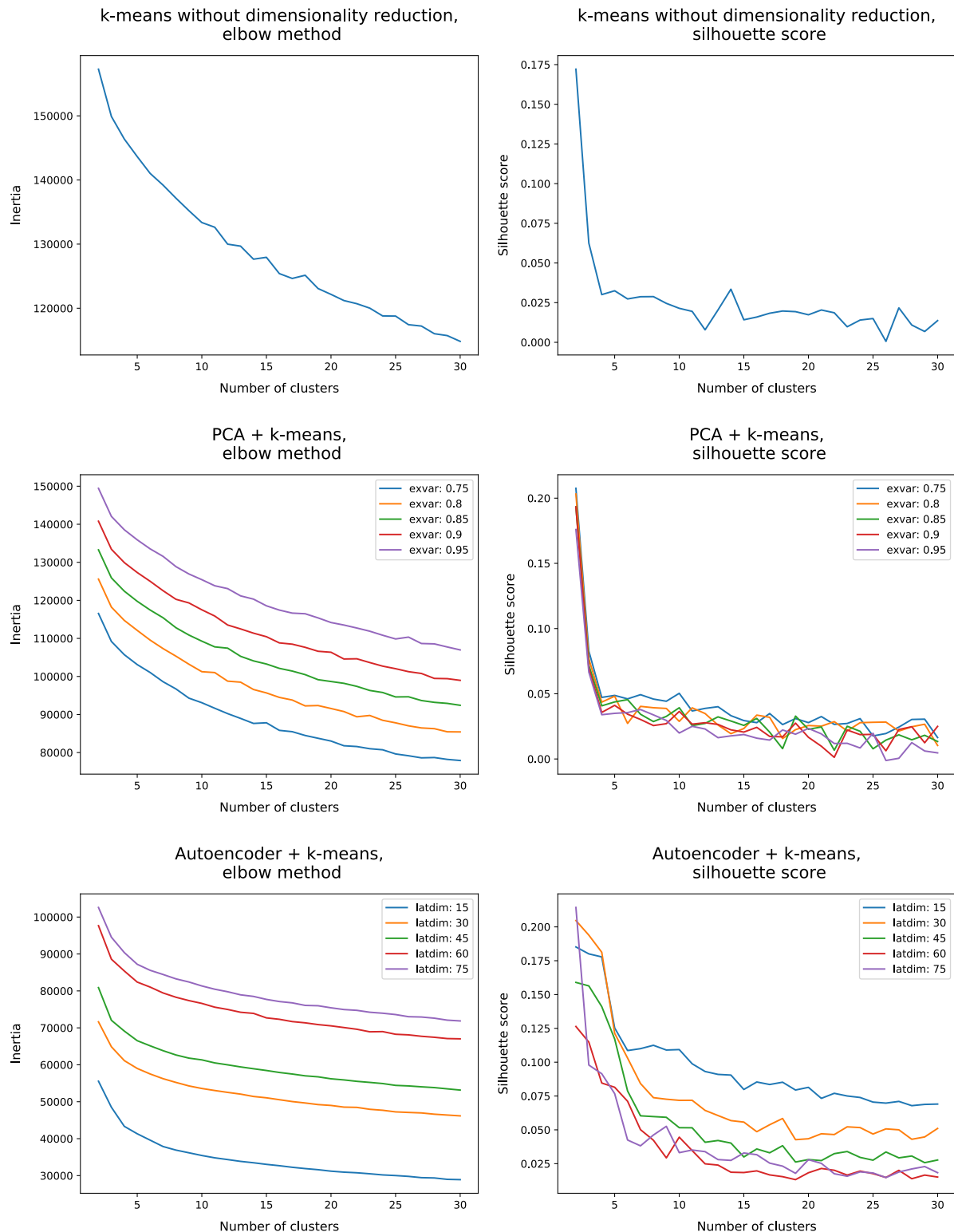
**Suppl. Table IV-2a-c: autoencoder optimisation results.** Mean absolute error (results displayed) was used as loss function, Adams as optimizer, reLu as activator, and the number of epochs was set at 100. The grid search optimisation performed on the number of layers, number of neurons and dimension of the latent space identified the following optimal parameters. (a.) for an input feature set composed of all variables: n layers = 2, n neurons = 128, and latent dimension = 32; (b.) for an input feature set composed of vital signs and common laboratory variables: n layers = 2, n neurons = 128, and latent dimension = 32; and (c.) for an input feature set composed of all variables: n layers = 2, n neurons = 64, and latent dimension = 32.

## Supplementary materials



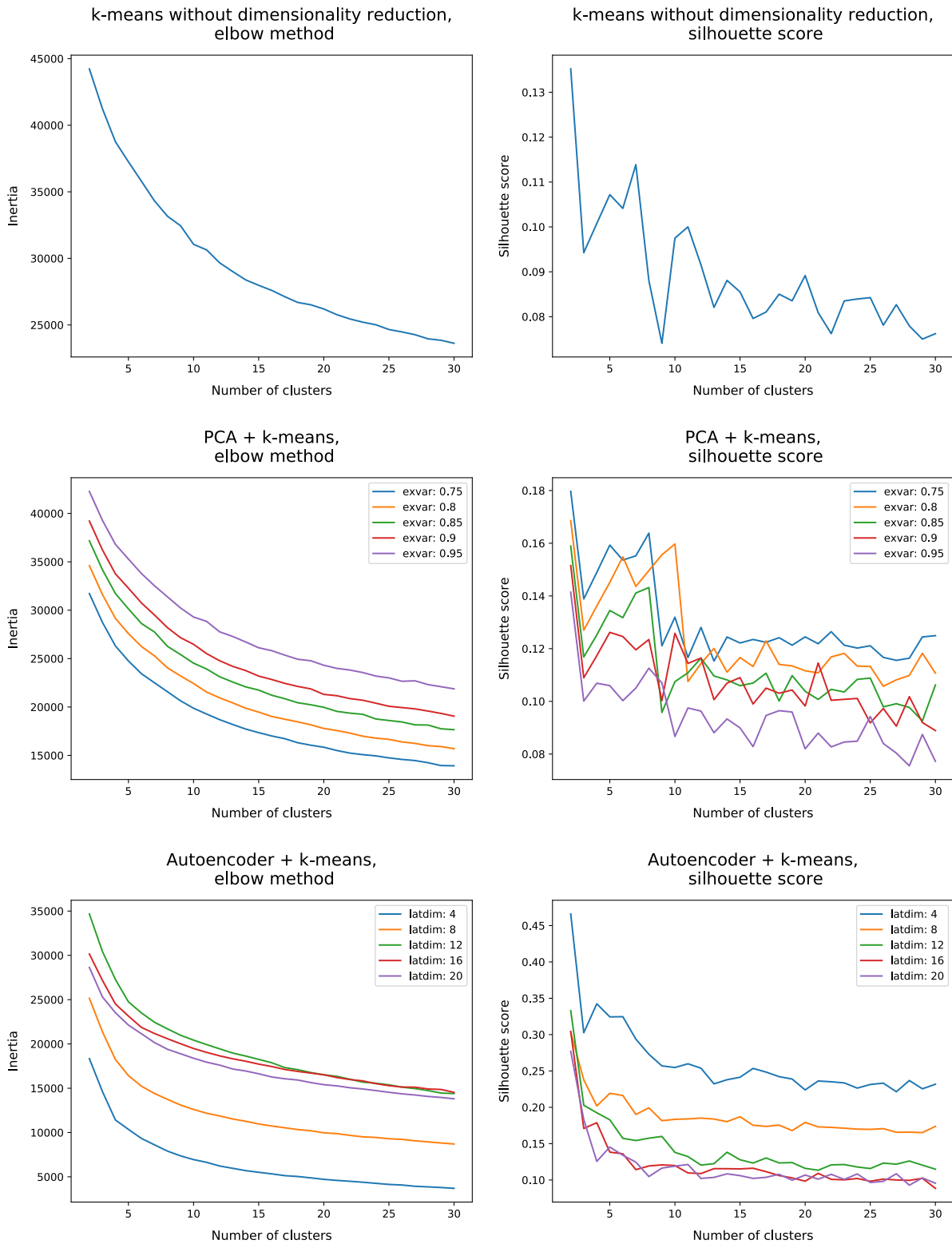
Suppl. Figure IV-1a: graphical representation of the elbow method and silhouette score values for cluster number optimisation after k-means clustering using all variables as input features ( $n=106$ ). PCA = principal component analysis; exvar = percentage of the variance explained by the number of selected components; latdim = dimensionality of the latent space.

## Supplementary materials



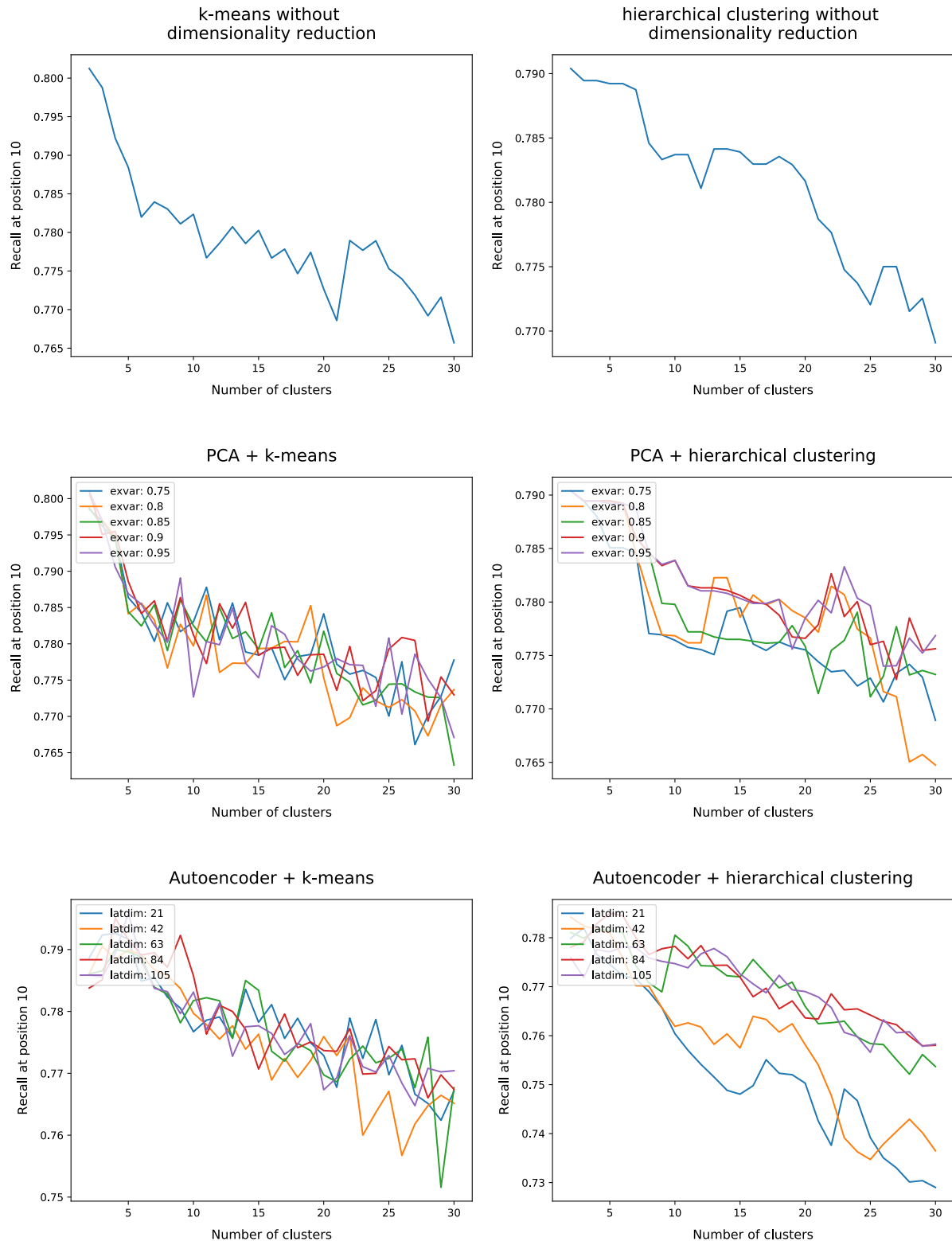
Suppl. Figure IV-1b: graphical representation of the elbow method and silhouette score values for cluster number optimisation after k-means clustering using vital signs and common laboratory variables as input features ( $n=78$ ). PCA = principal component analysis; exvar = percentage of the variance explained by the number of selected components; latdim = dimensionality of the latent space.

## Supplementary materials



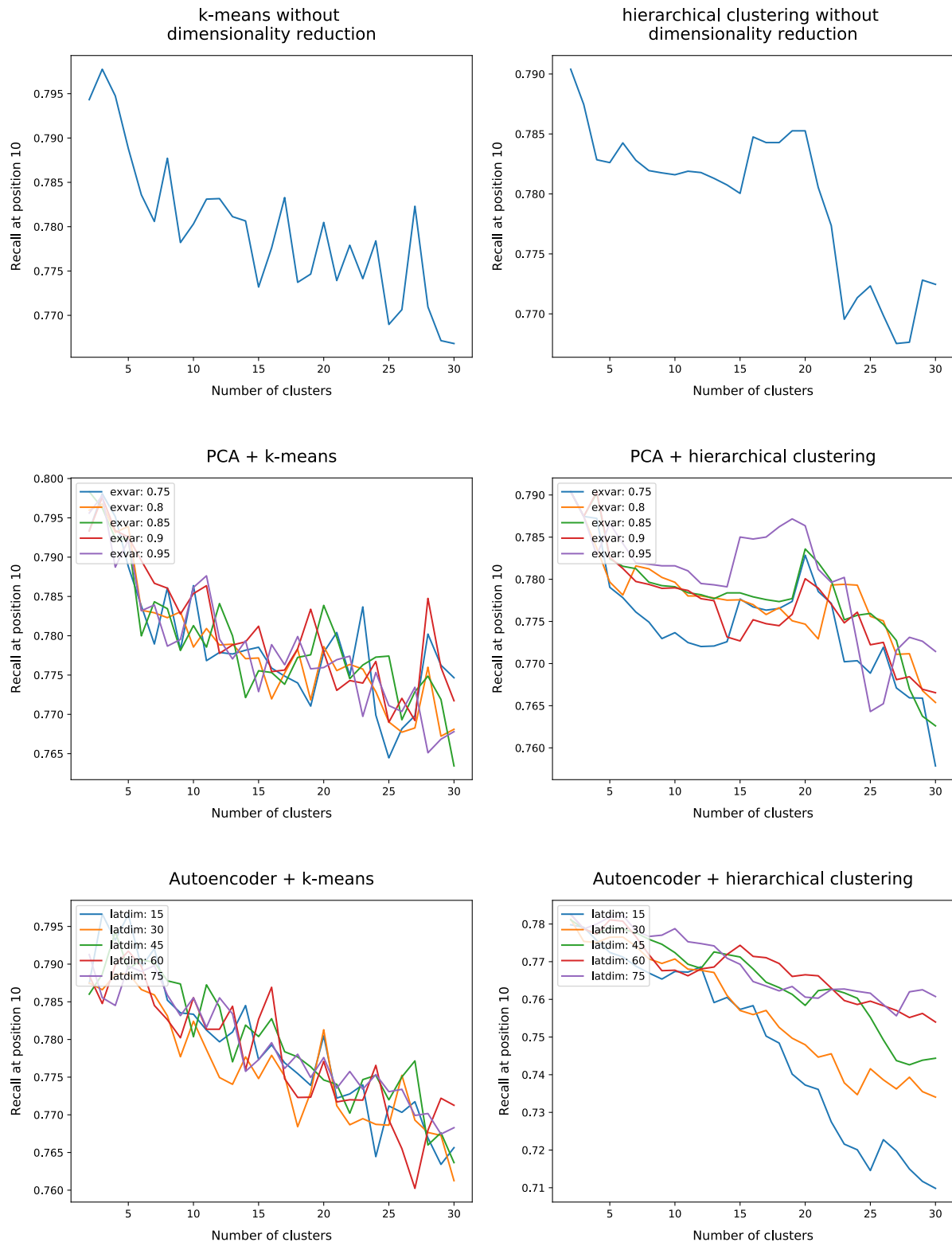
Suppl. Figure IV-1c: graphical representation of the elbow method and silhouette score values for cluster number optimisation after k-means clustering using only vital signs as input features (n=24). PCA = principal component analysis; exvar = percentage of the variance explained by the number of selected components; latdim = dimensionality of the latent space.

## Supplementary materials



**Suppl. Figure IV-2a: cluster number optimisation using cross validation (k=8) and recall at position ten on the validation set with all variables as input features (n=106).** Results for hierarchical clustering are displayed for comparison only (a method-specific approach was used for hierarchical cluster number optimisation). PCA = principal component analysis; exvar = percentage of the variance explained by the number of selected components; latdim = dimensionality of the latent space.

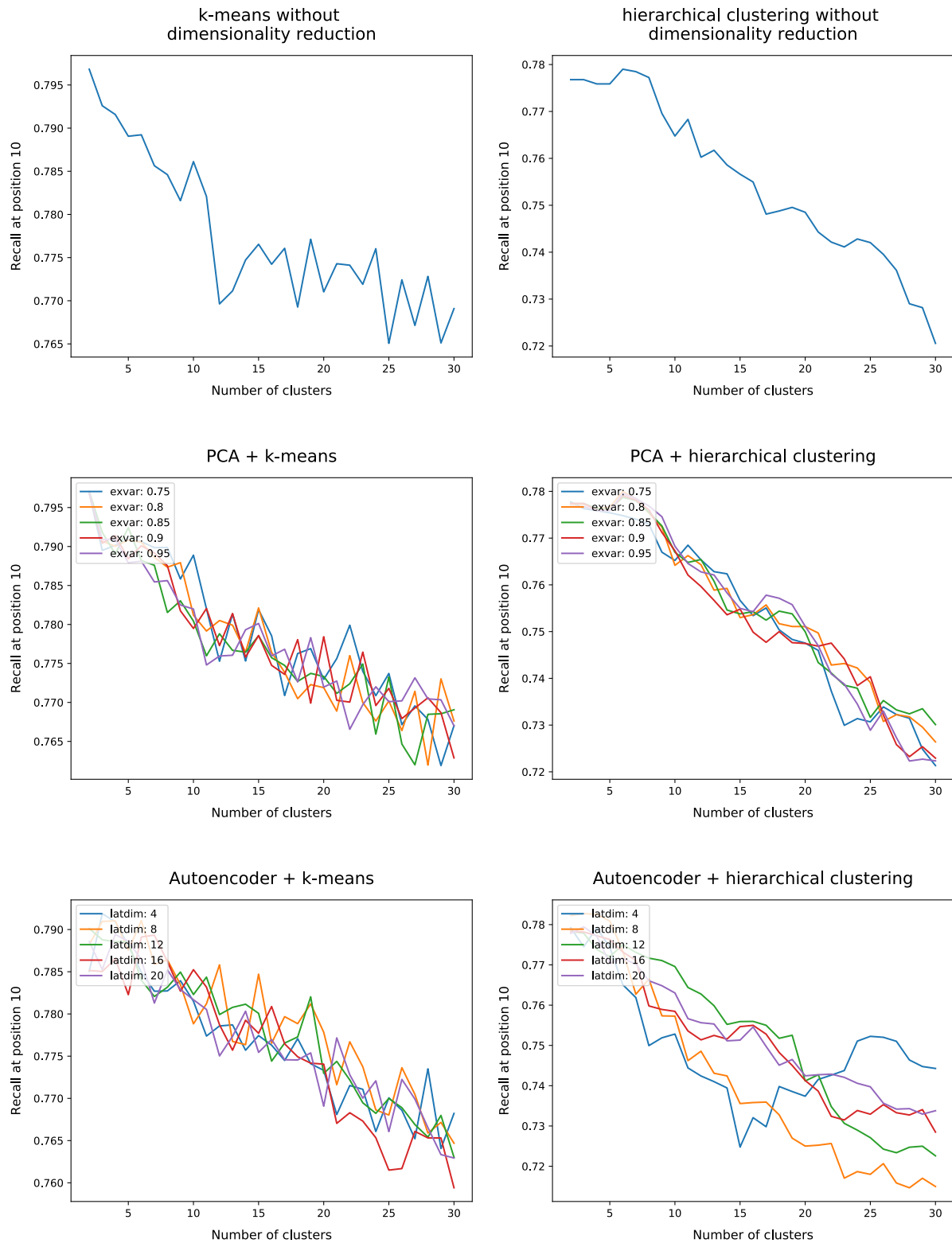
## Supplementary materials



Suppl. Figure IV-2b: cluster number optimisation using cross validation ( $k=8$ ) and recall at position ten on the validation set with vital signs and common laboratory variables as input features ( $n=78$ ). Results for hierarchical clustering are displayed for comparison only (a method-specific approach was used for hierarchical cluster number optimisation). PCA = principal component analysis; exvar = percentage of the variance explained by the number of selected components; latdim = dimensionality of the latent space.

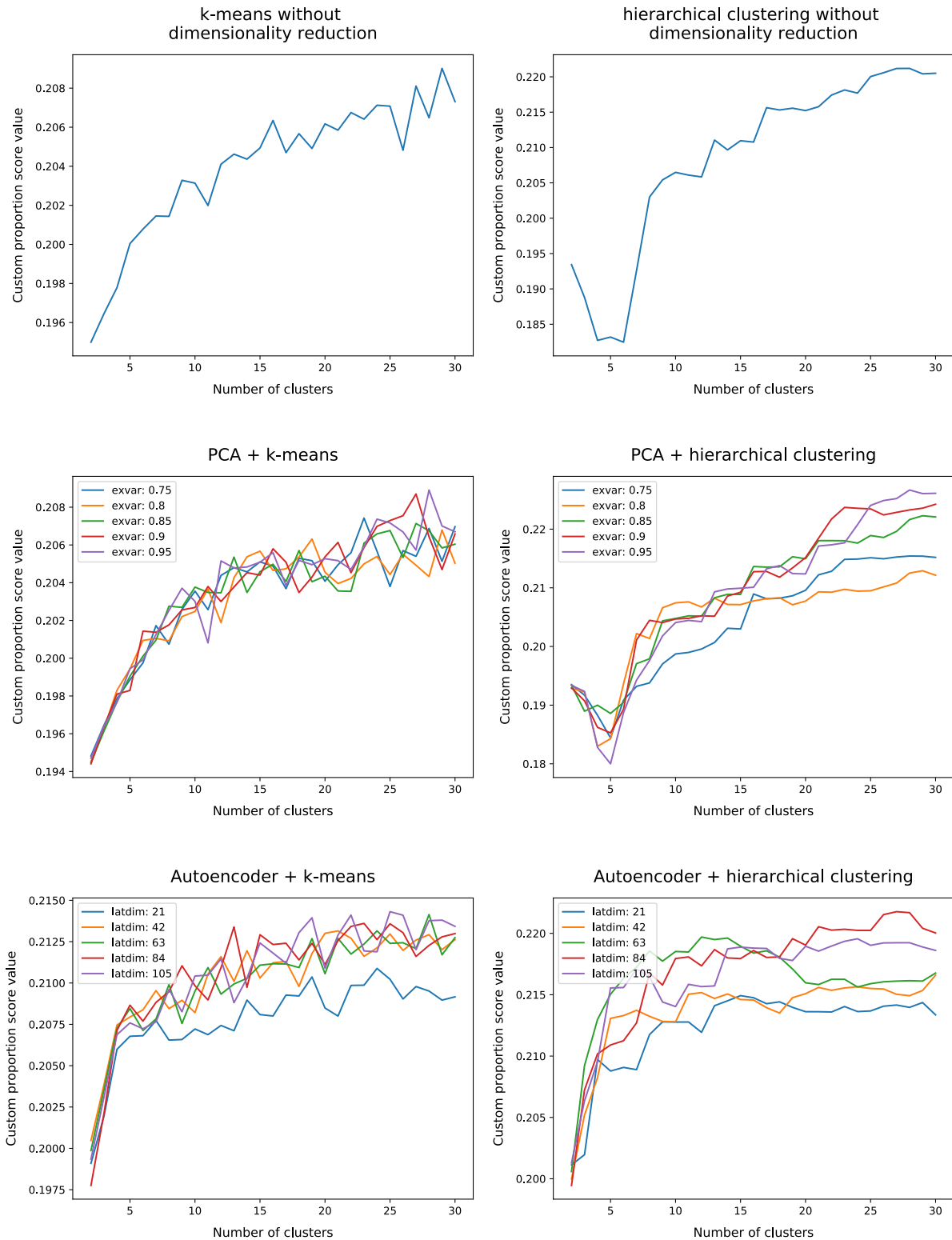


## Supplementary materials



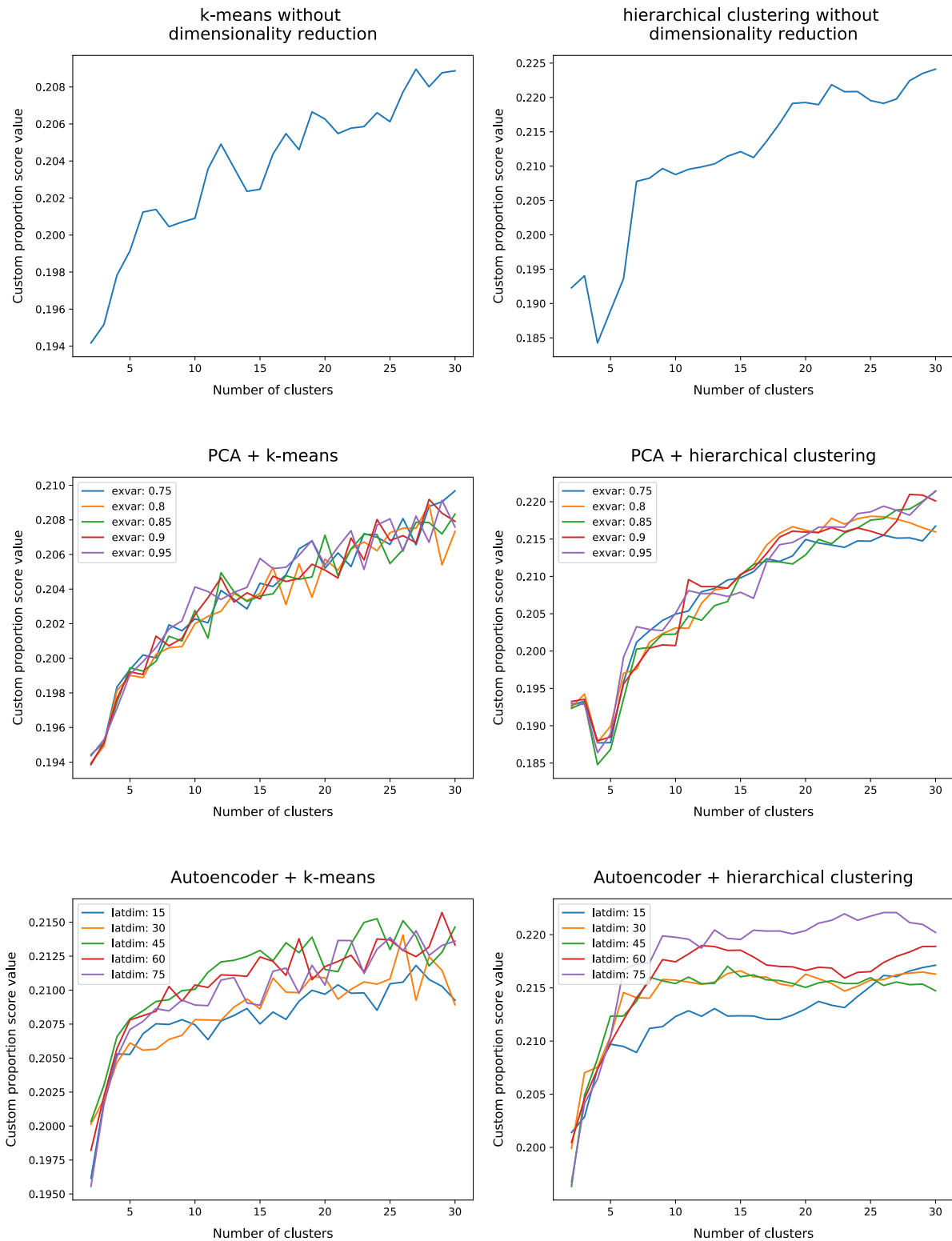
Suppl. Figure IV-2c: cluster number optimisation using cross validation ( $k=8$ ) and recall at position ten on the validation set with only vital signs as input features ( $n=24$ ). Results for hierarchical clustering are displayed for comparison only (a method-specific approach was used for hierarchical cluster number optimisation). PCA = principal component analysis; exvar = percentage of the variance explained by the number of selected components; latdim = dimensionality of the latent space.

## Supplementary materials



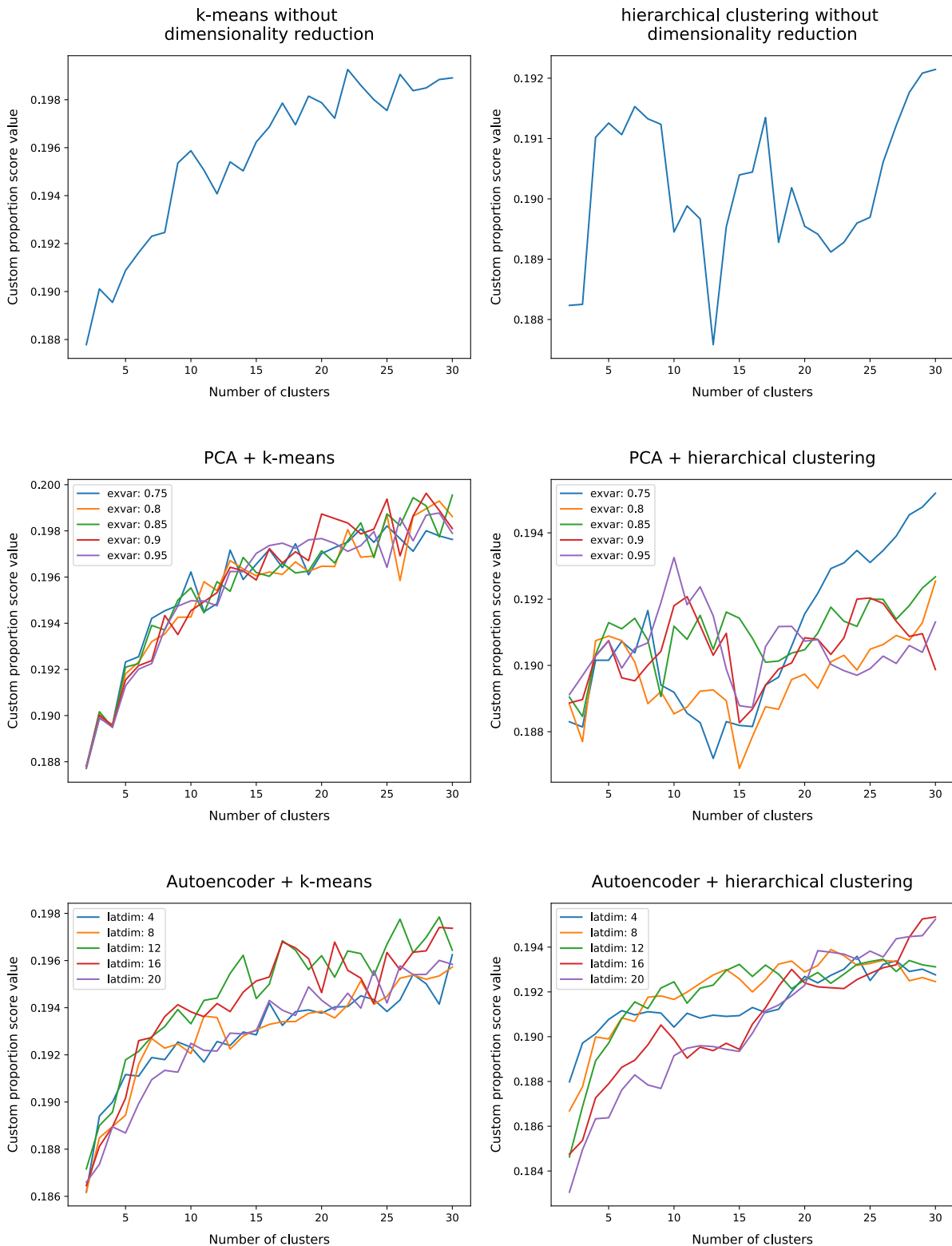
Suppl. Figure IV-3a: cluster number optimisation using cross validation ( $k=8$ ) and the custom proportion score on the validation set with all variables as input features ( $n=106$ ). Results for hierarchical clustering are displayed for comparison only (a method-specific approach was used for hierarchical cluster number optimisation). PCA = principal component analysis; exvar = percentage of the variance explained by the number of selected components; latdim = dimensionality of the latent space.

## Supplementary materials



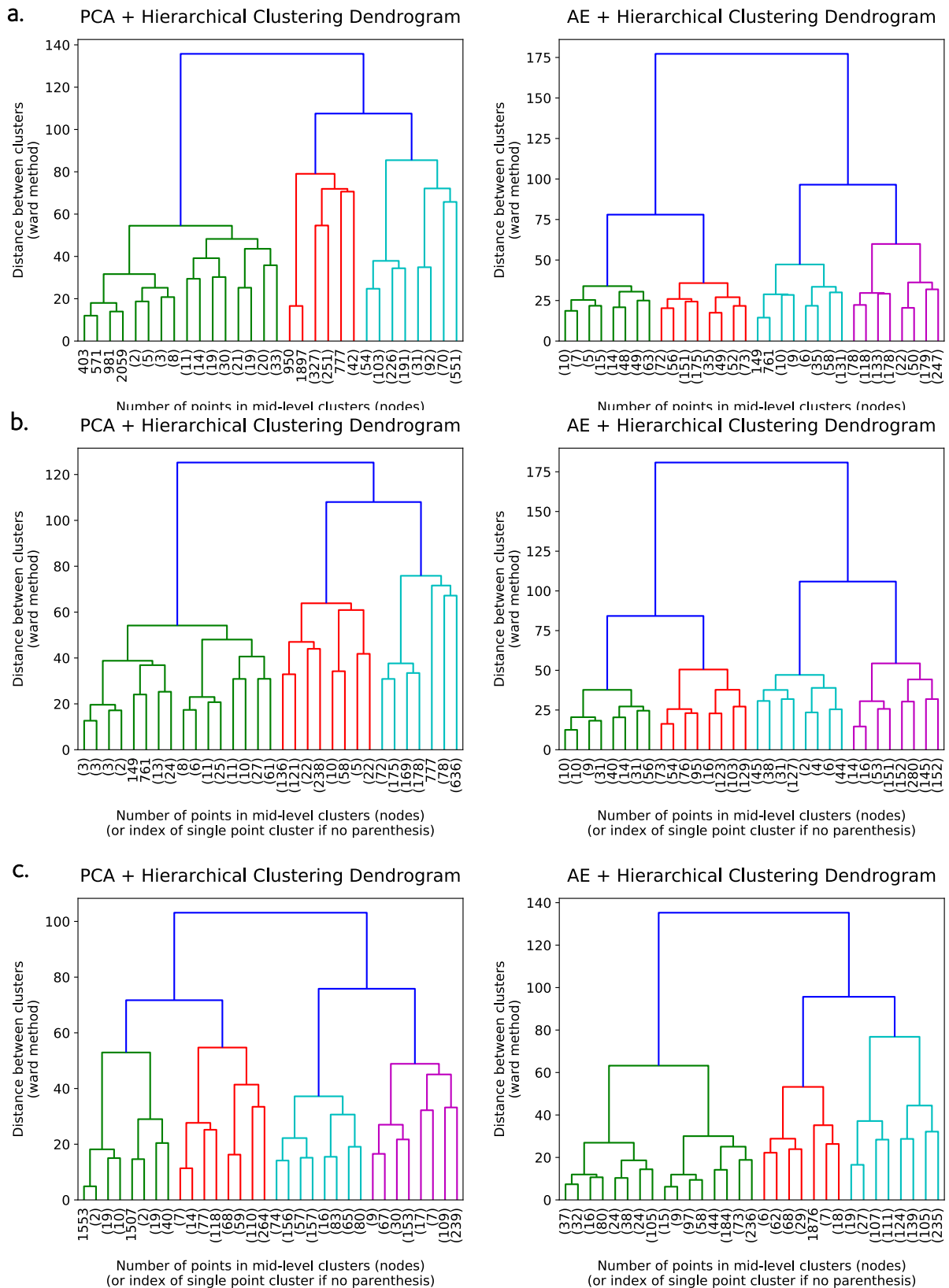
Suppl. Figure IV-3b: cluster number optimisation using cross validation ( $k=8$ ) and the custom proportion score on the validation set with vital signs and common laboratory variables as input features ( $n=78$ ). Results for hierarchical clustering are displayed for comparison only (a method-specific approach was used for hierarchical cluster number optimisation). PCA = principal component analysis; exvar = percentage of the variance explained by the number of selected components; latdim = dimensionality of the latent space.

## Supplementary materials



Suppl. Figure IV-3c: cluster number optimisation using cross validation ( $k=8$ ) and the custom proportion score on the validation set with only vital signs as input features ( $n=24$ ). Results for hierarchical clustering are displayed for comparison only (a method-specific approach was used for hierarchical cluster number optimisation). PCA = principal component analysis; exvar = percentage of the variance explained by the number of selected components; latdim = dimensionality of the latent space.

## Supplementary materials



Suppl. Figure IV-4: cluster number optimisation after hierarchical clustering using (a.) all variables, (b.) vital signs and common laboratory variables, and (c.) only vital signs as input features. PCA = principal component analysis; exvar = percentage of the variance explained by the number of selected components; latdim = dimensionality of the latent space.

## Chapter V

### **Suppl. Note V-1: additional literature search about existing guidance for AI evaluation in clinical settings.**

1. ("machine learning" or "artificial intelligence" or "deep learning").ti,ab,kw.
2. ("quality assessment" or "quality evaluation" or appraisal or "risk of bias" or "reporting guidelines" or "reporting standards" or "regulatory requirements" or "strength of evidence" or "quality of evidence" or "methodological assessment" or "methodological quality" or ergonomic\* or "human factors" or usability).ti,ab,kw.
3. (clinic\* or hospital\* or "real life" or "live environment" or "everyday practice").ti,ab,kw.
4. 1 and 2 and 3
5. limit 4 to yr="2016 - 2021"

### **Suppl. Note V-2: DECIDE-AI Round 1 and 2 – participants' names and affiliations**

Aaron Y. Lee (Department of Ophthalmology, School of Medicine, University of Washington, Seattle, WA, USA); Alan G. Fraser (School of Medicine, Cardiff University, Cardiff, UK); Alastair K. Denniston (University Hospitals Birmingham NHS Foundation Trust, Birmingham, UK); Ali Connell (Google Health, London, UK); Alykhan Vira (Quantum Health, Johannesburg, SA); Andre Esteva (Artera Research, Artera, Mountain View, CA, USA); Andrew D. Althouse (University of Pittsburgh, Pittsburgh, PA, USA); Andrew L. Beam (Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA, USA); Anne de Hond (CAIRElab, Leiden University Medical Centre, Leiden, NL); Anne-Laure Boulesteix (Institute for Medical Information Processing, Biometry and Epidemiology, Ludwig Maximilian University, Munich, DE); Anthony Bradlow (Rheumatology Department, Royal Berkshire Hospital, Reading, UK); Ari Ercole (Cambridge Centre for AI in Medicine, University of Cambridge, Cambridge, UK); Arsenio Paez (Nuffield Department of Primary Care Health Sciences, University of Oxford, Oxford, UK); Athanasios Tsanas (Usher Institute, Edinburgh Medical School, University of Edinburgh, Edinburgh, UK); Baptiste Vasey (Nuffield Department of Surgical Sciences, University of Oxford, Oxford, UK); Barry Kirby (K Sharp, Llanelli, UK); Bart Geerts (Healthpus.ai BV, Amsterdam, NL); Ben Glocker (Department of Computing, Imperial College London, London, UK); Bilal A. Mateen (The Wellcome Trust, London, UK); Bruce Campbell (University of Exeter Medical School, Exeter, UK); Campbell Rogers (HeartFlow Inc., Redwood City, CA, USA); Carmelo Velardo (Sensyne Health, UK and Department of Engineering Science, University of Oxford, Oxford, UK); Chang Min Park (Seoul National University College of Medicine, Seoul, KR); Charisma Hehakaya (Division of Imaging & Oncology, University Medical Center Utrecht, Utrecht, NL); Chris Baber (University of Birmingham, Birmingham, UK); Chris Paton (Nuffield Department of Medicine, University of Oxford, Oxford, UK); Christian Johnner (Johnner Institute, Konstanz, DE); Christopher J. Kelly (Google Health, London, UK); Christopher J. Vincent (PDD Group Ltd, London, UK); Christopher Yau (University of Manchester,

## Supplementary materials

Manchester, UK); Clare McGenity (Pathology and Data Analytics, University of Leeds, Leeds, UK); Constantine Gatsonis (Department of Biostatistics, Brown University School of Public Health, Providence, RI, USA); Corinne Faivre-Finn (The Christie NHS Foundation Trust, Manchester, UK); Crispin Simon (London School of Economics, London, UK); Cyrus Espinoza (Patient representative); Daniel P. Jenkins (DCA Design, UK); Daniel S.W. Ting (Singapore National Eye Center, Singapore Eye Research Institute, Singapore, SG); Danielle Sent (Department of Medical Informatics, Amsterdam UMC, University of Amsterdam, Amsterdam, NL); Danilo Bzdok (Mila, Quebec Artificial Intelligence Institute, Montreal, CA); Darren Treanor (Leeds Teaching Hospitals NHS Trust, Leeds, UK); David A. Clifton (Institute of Biomedical Engineering, Department of Engineering Science, University of Oxford, Oxford, UK); David C. Wong (Department of Computer Science and Centre for Health Informatics, University of Manchester, Manchester, UK); David F. Steiner (Google Health, Palo Alto, USA); David Higgins (Berlin Institute of Health, Berlin, DE); Dawn Benson (Patient representative); Deborah Morrison (National Institute for Health and Care Excellence, UK); Declan P. O'Regan (MRC London Institute of Medical Sciences, Imperial College London, London, UK); Dominic Danks (Institute of Cancer and Genomic Sciences, University of Birmingham, Birmingham, UK); Emanuele Neri (University of Pisa, Pisa, IT); Evangelia Kyrimi (School of Electronic Engineering and Computer Science (EECS), Queen Mary University of London, London, UK); Falk Schwendicke (Charité Universitätsmedizin Berlin, Berlin, DE); Farah Magrabi (Australian Institute of Health Innovation, Macquarie University, Sydney, AU); Frances Ives (West Midlands Academic Health Science Network, Birmingham, UK); Frank E. Rademakers (Department Cardiovascular sciences, KU Leuven, Leuven, BE); Gary S. Collins (Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology & Musculoskeletal Sciences, University of Oxford, Oxford, UK); George E. Fowler (Bristol Centre for Surgical Research, Department of Population Health Sciences, Bristol Medical School, Bristol, UK); Giuseppe Frau (Deep Blue, Rome, IT); H. D. Jeffry Hogg (Population Health Science Institute, Newcastle University, Newcastle upon Tyne, UK); Hani J. Marcus (Department of Neurosurgery, National Hospital for Neurology and Neurosurgery, Queen Square, London, UK); Heang-Ping Chan (Department of Radiology, University of Michigan, Ann Arbor, MI, USA); Henry Xiang (The Abigail Wexner Research Institute, Nationwide Children's Hospital, The Ohio State University, Columbus, OH, USA); Hugh F. McIntyre (Department of Medicine, East Sussex Healthcare Trust, Hastings, UK); Hugh Harvey (Hardian Health, UK); Hyungjin Kim (Department of Radiology, Seoul National University Hospital, Seoul, KR); Ibrahim Habli (Department of Computer Science, University of York, York, UK); James C. Fackler (Department of Anesthesiology and Critical Care Medicine, The Johns Hopkins University School of Medicine, Baltimore, MD, USA); James Shaw (Joint Centre for Bioethics, University of Toronto, Toronto, CA); Janet Higham (University of Oxford, Oxford, UK); Jared M. Wohlgemut (Centre for Trauma Sciences, Blizzard Institute, Queen Mary University of London, London, UK); Jaron Chong (Medical Imaging, Western University, London, CA); Jean-Emmanuel Bibault (Radiation Oncology Department, Hôpital Européen Georges Pompidou, AP-HP, Paris, FR); Jérémie F. Cohen (Center of Research in Epidemiology and Statistics (Inserm 1153), Université de Paris, Paris, FR); Jesper Kers (Department of Pathology, Amsterdam UMC, University

## Supplementary materials

of Amsterdam, Amsterdam, NL); Jessica Morley (Oxford Internet Institute, University of Oxford, Oxford, UK); Joachim Krois (Oral Diagnostics & Digital Health & Health Services Research, Charité Universitätsmedizin Berlin, Berlin, Germany); Joao Monteiro (Nature Medicine, New York, NY, USA); Joel Horovitz (Department of Surgery, Maimonides Medical Center, Brooklyn, NY, USA); Johan Ordish (The Medicines and Healthcare products Regulatory Agency, London, UK); John Fletcher (The BMJ, London, UK); Jonathan Taylor (Nuclear Medicine / 3DLab, Sheffield Teaching Hospitals, Sheffield, UK); Jung Hyun Yoon (Department of Radiology, Severance Hospital, Yonsei University College of Medicine, Seoul, KR); Karandeep Singh (Department of Learning Health Sciences, University of Michigan Medical School, Ann Arbor, MI, USA); Karel G.M. Moons (Julius Center, UMC Utrecht, Utrecht University, Utrecht, NL); Kassandra Karpathakis (Harvard TH Chan School of Public Health, Boston, MA, USA); Ken Catchpole (Medical University of South Carolina, Charleston, SC, USA); Kerenza Hood (Centre for Trials Research, Cardiff University, Cardiff, UK); Konstantinos Balaskas (Moorfields Ophthalmic Reading Centre and Clinical AI Hub, Moorfields Eye Hospital, London, UK); Konstantinos Kamnitsas (School of Computer Science, University of Birmingham, Birmingham, UK); Laura Militello (Applied Decision Science LLC, USA); Laure Wynants (Department of Epidemiology, CAPHRI Care and Public Health Research Institute, Maastricht University, Maastricht, NL); Lauren Morgan (Morgan Human Systems Ltd, Shrewsbury, UK); Livia Faes (Moorfields Eye Hospital, London, UK); Luc J.M. Smits (Department of Epidemiology, Maastricht University, Maastricht, NL); Ludwig C. Hinske (Institute for Biomedical Information Processing, Biometry and Epidemiology, Ludwig-Maximilians-University Munich, Munich, DE); Luke Oakden-Rayner (Australian Institute for Machine Learning, University of Adelaide, Adelaide, AU); M. Khair ElZarrad (U.S. Food and Drug Administration)<sup>§</sup>; Maarten van Smeden (Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, Utrecht, NL); Mara Giavina-Bianchi (Hospital Israelita Albert Einstein, São Paulo, BR); Mark Daley (The University of Western Ontario, London, CA); Mark P. Sendak (Duke Institute for Health Innovation, Durham, USA); Mark Sujan (Human Factors Everywhere Ltd, Woking, UK); Maroeska Rovers (Department of Operating rooms, Radboudumc, Nijmegen, NL); Matthew DeCamp (University of Colorado, Boulder, CO, USA); Matthieu Komorowski (Dept of Surgery and Cancer, Imperial College London, London, UK); Max Marsden (Centre for Trauma Science, Blizzard Institute, Queen Mary University of London, London, UK); Maxine Mackintosh (Genomics England, Queen Mary University of London, London, United Kingdom); Melissa D. McCradden (The Hospital for Sick Children, Toronto, CA); Michael D. Abramoff (University of Iowa, Iowa City, IA, USA); Miguel Ángel Armengol de la Hoz (Big Data Department, FPS, Regional Ministry of Health of Southern Spain, ES); Myura Nagendran (UKRI Centre for Doctoral Training in AI for Healthcare, Imperial College London, London, UK); Neale Hambidge (National Hospital for Neurology and Neurosurgery, Queen Square, London, UK); Neil Daly (Skin Analytics, London, UK); Niels Peek (Division of Informatics, Imaging and Data Science, The University of Manchester, Manchester, UK); Oliver Redfern (Kadoorie Centre for Critical Care Research and Education, Nuffield Department of

---

<sup>§</sup> Participation represents personal views and perspectives that may not necessarily reflect the positions and opinions of the U.S. FDA.



## Supplementary materials

Clinical Neurosciences, University of Oxford, Oxford, UK); Omer F. Ahmad (Wellcome/EPSRC centre for Interventional & Surgical Sciences (WEISS), University College London, London, UK); Patrick M. Bossuyt (Amsterdam University Medical Centers, University of Amsterdam, Amsterdam, NL); Pearse A. Keane (Institute of Ophthalmology, University College London, London, UK); Pedro N.P. Ferreira (CENTEC - IST, University of Lisbon, Lisbon, PT); Peter McCulloch (Nuffield Department of Surgical Sciences, University of Oxford, Oxford, UK); Peter Watkinson (Nuffield Department of Clinical Neurosciences, University of Oxford, Oxford, UK); Peter Wheatstone (Patient representative); Petra Schnell-Inderst (Institut of Public Health, Medical Decision Making and HTA, UMIT - University for Health Sciences, Medical Informatics and Technology, Hall i. T., AT); Pietro Mascagni (Gastrointestinal Endoscopic Surgery, Fondazione Policlinico Universitario A. Gemelli IRCCS, Rome, IT); Piyush Mathur (Cleveland Clinic, Cleveland, Ohio, USA); Prokar Dasgupta (King's Health Partners Academic Surgery, King's College London, London, UK); Pujun Guan (Graduate School of Biomedical Sciences, University of Texas MD Anderson Cancer Center and UTHealth, Houston, USA); Rawen Kader (Division of Surgery and Interventional Sciences, University College London, London, UK); Reena Chopra (Google Health, London, UK); Ritse M. Mann (Department of medical imaging, Radboud University Medical Center, Nijmegen, NL); Rupa Sarkar (The Lancet Digital Health, The Lancet Group, London, UK); Saana M. Mäenpää (Department of Neurosurgery, Helsinki University Hospital, Helsinki, FI); Samuel G. Finlayson (Harvard Medical School, Boston, MA, USA); Sarah Vollam (Nuffield Department of Clinical Neurosciences, University of Oxford, Oxford, UK); Sean P. White (NHS England, UK); Sebastian J. Vollmer (Data Science and its Application, DFKI, Kaiserslautern, DE); Seong Ho Park (Department of Radiology, Asan Medical Center, Seoul, KR); Shakir Laher (University of York, York, UK); Shalmali Joshi (SEAS, Harvard University, Cambridge, MA, USA); Spiros Denaxas (Institute of Health Informatics, University College London, London, UK); Suchi Saria (Departments of Computer Sciences, Statistics, and health Policy, and Division of Informatics, Johns Hopkins University, Baltimore, MD, USA); Susan C. Shelmerdine (Department of Clinical Radiology, Great Ormond Street Hospital for Children NHS Foundation Trust, London, UK); Tom J.W. Stocker (PUBLIC Ltd, Oxford, UK); Valentina Giannini (University of Turin, Turin, IT); Valerie Keston-Hole (Patient representative); Vince I. Madai (QUEST Centre for Responsible Research, Berlin Institute of Health, Charité Universitätsmedizin Berlin, Berlin, Germany); Virginia Newcombe (University Division of Anaesthesia, Department of Medicine, University of Cambridge, Cambridge, UK); Wendy A. Rogers (Philosophy Department and School of Medicine, Macquarie University, Sydney, AU); William Ogallo (IBM Research Africa, Nairobi, KE); Wim Weber (The BMJ, London, UK); Xiaoxuan Liu (University Hospitals Birmingham NHS Foundation Trust, Birmingham, UK); Zane B. Perkins (Centre for Trauma Sciences, Queen Mary University of London, London, UK).

## Supplementary materials

Theme	Item n°	Recommendation
<b>Title and abstract</b>		
Title/abstract	1a	Identify the study as early stage or formative clinical evaluation of an artificial intelligence or machine learning based decision support system, mentioning the clinical problem addressed.
	1b	Provide a structured summary of the study, including: target clinical problem, intended use of the algorithm and integration in the clinical pathway, type of algorithm, study design, study setting, number of patients and users included, control group if applicable, primary and secondary outcomes, key safety endpoints, human factors aspects evaluated, main results, conclusions.
<b>Introduction</b>		
Target clinical problem and population	2	Describe the target clinical problem and medical condition, including the current state of the art practice, and the target patient population.
Intended use	3	Describe the intended use of the algorithm, its planned integration in the care pathway and the impact in terms of patient outcomes it intends to achieve.
Current stage of development	4	Describe the current stage of development of the algorithm (both from a scientific and a regulatory perspective). State if the algorithm is tested as a medical device and, if so, which regulatory approval is sought/was obtained.
Objectives	5	State the study objectives.
<b>Methods</b>		
Research governance	6a	Provide a reference to any study protocol, study registration number and ethics approval.
	6b	State what measures were taken to protect patient privacy and data security.
Study design	7	Describe the study design.
Participants	8a	Describe precisely how patients were recruited, stating the inclusion and exclusion criteria, and how the number of recruited patients was selected.
	8b	Describe precisely how users were recruited, stating the inclusion and exclusion criteria, and how the number of recruited users was selected. If applicable, describe the control group in sufficient detail to allow replication.
	8c	Describe any attempts to familiarise the users with the algorithm, including any training received.
Algorithm	9	Briefly describe the algorithm, including: the version number, the type of AI model used, the characteristics of the patient population on which it was trained and the expected performance from in silico study. Refer to any previous development work.
Implementation	10a	Describe precisely the environment in which the algorithm was tested, including the availability of the algorithm's input data and which additional clinical information (i.e. not provided by the algorithm) was accessible to the users to interpret or put into context the output of the algorithm.
	10b	Describe the clinical workflow/pathway in which the algorithm was deployed and who held the responsibility for the final clinical decision.
	10c	Describe precisely how the algorithm was used and the timing of the decision support.
	10d	Describe the technical details of the implementation, including the integration within the existing study site IT infrastructure, the software and hardware needed to run the algorithm and any algorithmic thresholds used.
	10e	Identify the data used as inputs. Describe how the data were acquired, the process needed to enter the input data, any pre-processing applied and how missing/low-quality data were handled.
	10f	Describe the algorithm outputs and how they were presented to the users.
Outcomes	11a	Specify the primary and secondary outcomes measured.
	11b	Describe how algorithm recommendation/output errors were defined and how they were identified.

## Supplementary materials

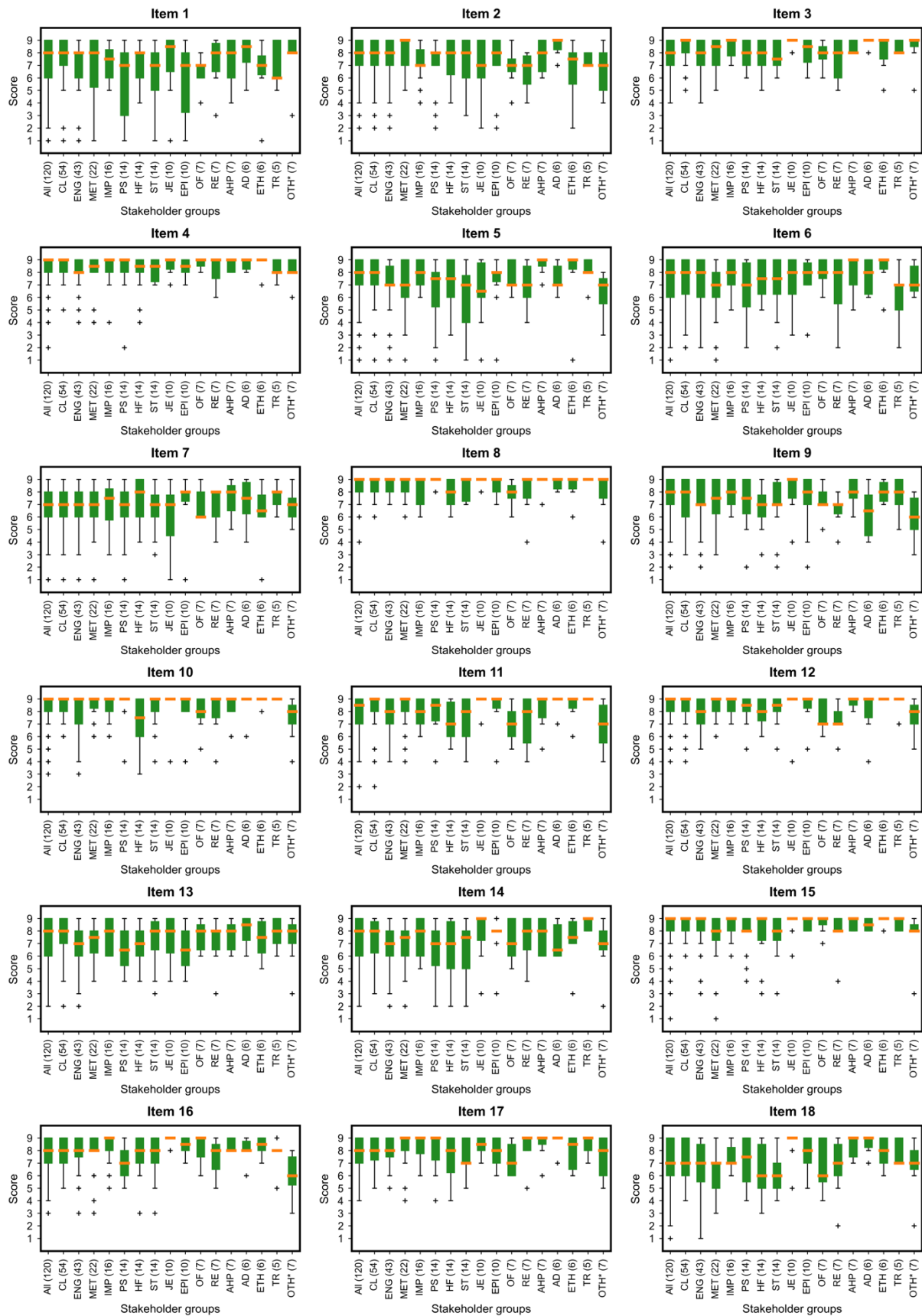
Analysis	12	Describe the pre-specified analysis plan for the primary and secondary outcomes as well as for any prespecified additional analyses, including subgroup analyses and their rationale.
Safety	13a	Define the algorithm safety requirements, how these were established preclinically, and how compliance to these requirements was evaluated during the study.
	13b	Describe the methodology used to detect any new, unexpected risks arising from the real-life clinical use of the algorithm.
Human factors	14	Describe the human factors tools, methods or frameworks used, the use cases considered and the users involved in the human factors evaluation.
Patient engagement	15	State whether patients were involved in any aspect of the study design, conduct or in the development of the research question or outcome measures.
Ethics consideration	16	Describe any ethics methodology, consultation or involvement during the design or implementation of the study.
<b>Results</b>		
Participants	17a	Describe the patient study group baseline characteristics (number, number of centres, age, sex, ethnicity if relevant, comorbidities, prevalence of the target conditions, etc.).
	17b	Describe the users study group baseline characteristics (number, number of centres, specialty, seniority, previous experience with digital support, etc.).
Implementation	18a	Report on the user exposure to the algorithm (implementation reach), on the number of instances the algorithm was used (implementation dose) and on the users' adherence to the intended implementation (implementation fidelity).
	18b	Report changes caused by the algorithm to the clinical workflow, if any.
Modifications	19	Report any changes made to the algorithm or its hardware platform between the prototype used at the beginning of the study and its final version. Report the timing of these modifications and the changes in outcomes observed after each of them.
Main results	20a	Report on the prespecified outcomes for the algorithm-assisted users (both overall and at an individual user level), including any variation over time.
	20b	Report on the prespecified outcomes for the stand-alone algorithm, if applicable.
	20c	Report on the prespecified outcomes for the control group, if applicable.
Safety and errors	21a	Report on the compliance with the specified safety requirements and any severe adverse events.
	21b	Report any additional risks identified from the real-life clinical use of the algorithm.
	21c	Report any algorithm malfunction or issues with hardware or software during the study.
	21d	Report any algorithm recommendation errors, detailing their rate of occurrence, causes, whether they were corrected and potential/actual impact on patient care.
	21e	Report any human errors, detailing their rate of occurrence, causes, whether they were corrected and potential/actual implication for patient care.
Subgroup analysis	22	Report on the difference in the main outcomes according to the specified subgroups.
Human factors	23a	Report on the user agreement with the algorithm. Describe any instances of and reasons for user deviation from the algorithm's recommendations and, if applicable, user changing their mind based on the algorithm recommendations.
	23b	Report on the evolution of users' trust in the algorithm.
	23c	Report on the usability evaluation, including time to task completion, display interface evaluation and user satisfaction.
	23d	Report on the user workload and learning curves evaluation.
	23e	Report on the user perception of the algorithm outputs' interpretability and clinical value.

## Supplementary materials

Discussion		
Support intended purpose	24	Discuss whether the obtained results support the intended purpose of the algorithm in real world clinical settings.
Safety and errors	25	Discuss what the results suggest about the safety profile of the algorithm. Discuss the algorithm's errors and, if appropriate, identify any underlying pattern or algorithmic bias, explain how these can be mitigated.
Human factors	26	Discuss the results of the human factors evaluation and the reasons for human deviation from the algorithm's recommendations or intended use.
Scale up	27	Discuss the scale-up feasibility and requirements, as well as the possible design of large-scale summative evaluation in light of the obtained results. Summarise the lessons learned from the study.
Strength and limitations	28	Discuss the strengths and limitations of the study, including any bias in the study design.
Statements		
Conflicts of interest	29	Disclose any relevant conflict of interest, including: the source of funding for the study, the role of funders, any other role played by commercial companies and authors' conflicts of interest.
Data Availability	30	Disclose if and how data and code (pre-processing and algorithm) are available.

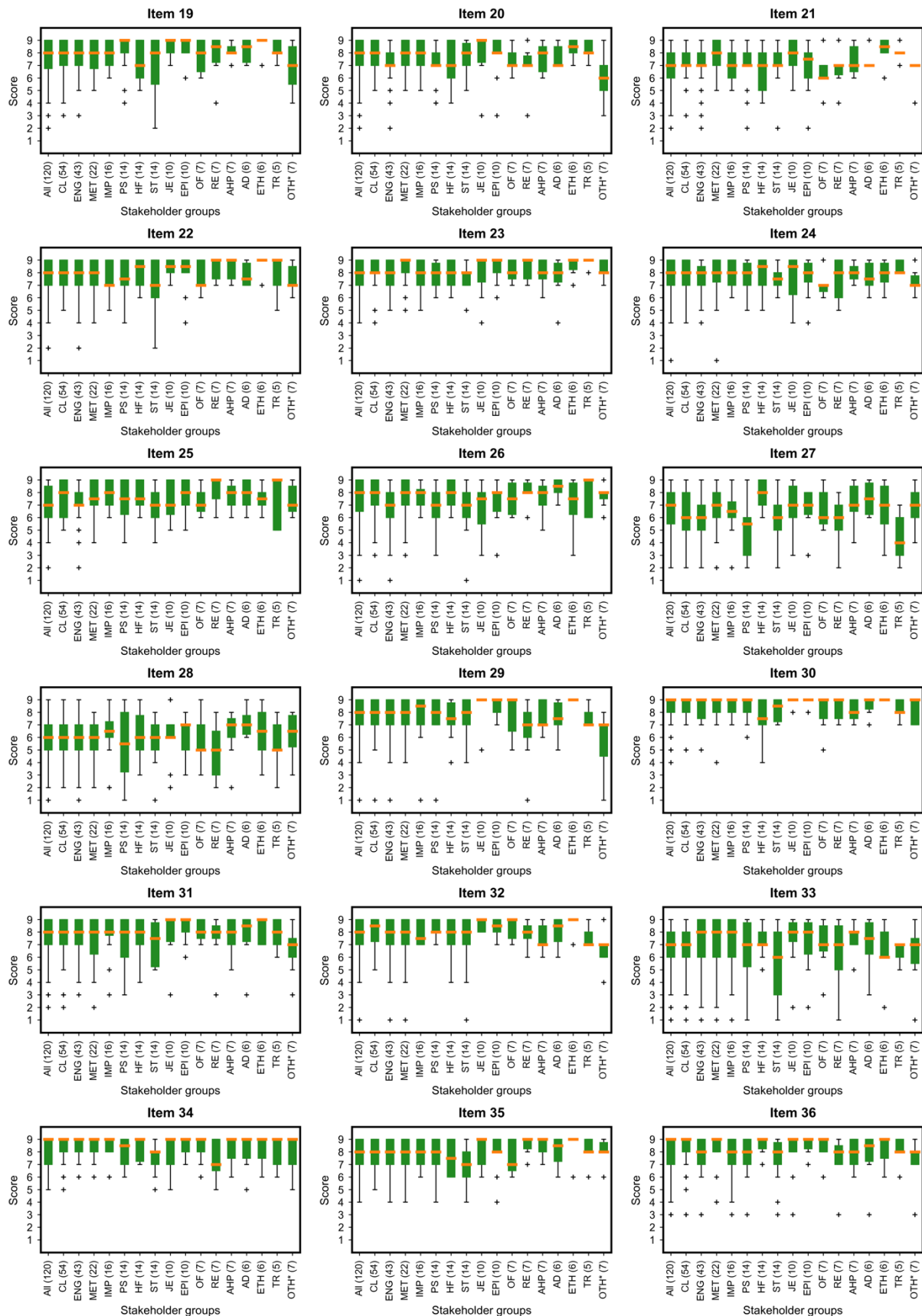
**Suppl. Table V-1: revised item list.**

## Supplementary materials



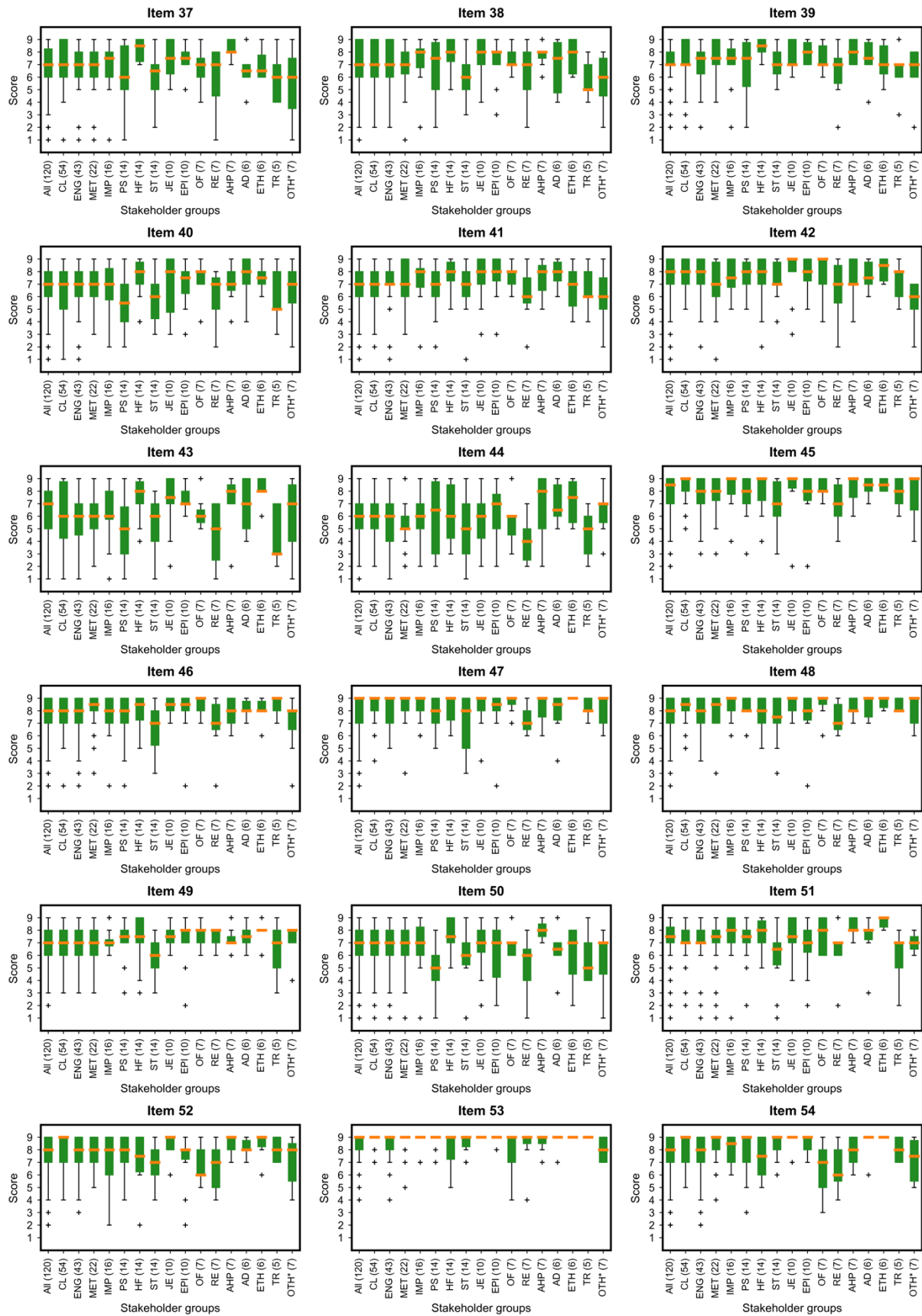
**Suppl. Figure V-1a: median scores and IQR of item 1 to 18 during the first round of Delphi, overall and by stakeholder group.** Whiskers are the last value comprised within 1.5 IQR on both side of the median. Crosses are outliers. A participant can self-affiliate to several stakeholder groups. \*Others: Funders, Patients representatives, Payers & Commissioners, and Psychologists. AD = Administrators; AHP = Allied Health Professionals; CL = Clinicians; ENG = Engineers & Computer Scientists; EPI = Epidemiologists; ETH = Ethicists; HF = Human Factors Specialists; IMP = Implementation Scientists; JE = Journal Editors; OF = Policy Makers & Official Institutions; PM = Policy Makers; PS = Entrepreneurs & Private Sector Representatives; RE = Regulators; TR = Trialists.

## Supplementary materials



**Suppl. Figure V-1b: median scores and IQR of item 19 to 36 during the first round of Delphi, overall and by stakeholder group.** Whiskers are the last value comprised within 1.5 IQR on both side of the median. Crosses are outliers. A participant can self-affiliate to several stakeholder groups. \*Others: Funders, Patients representatives, Payers & Commissioners, and Psychologists. AD = Administrators; AHP = Allied Health Professionals; CL = Clinicians; ENG = Engineers & Computer Scientists; EPI = Epidemiologists; ETH = Ethicists; HF = Human Factors Specialists; IMP = Implementation Scientists; JE = Journal Editors; OF = Policy Makers & Official Institutions; PM = Policy Makers; PS = Entrepreneurs & Private Sector Representatives; RE = Regulators; TR = Trialists.

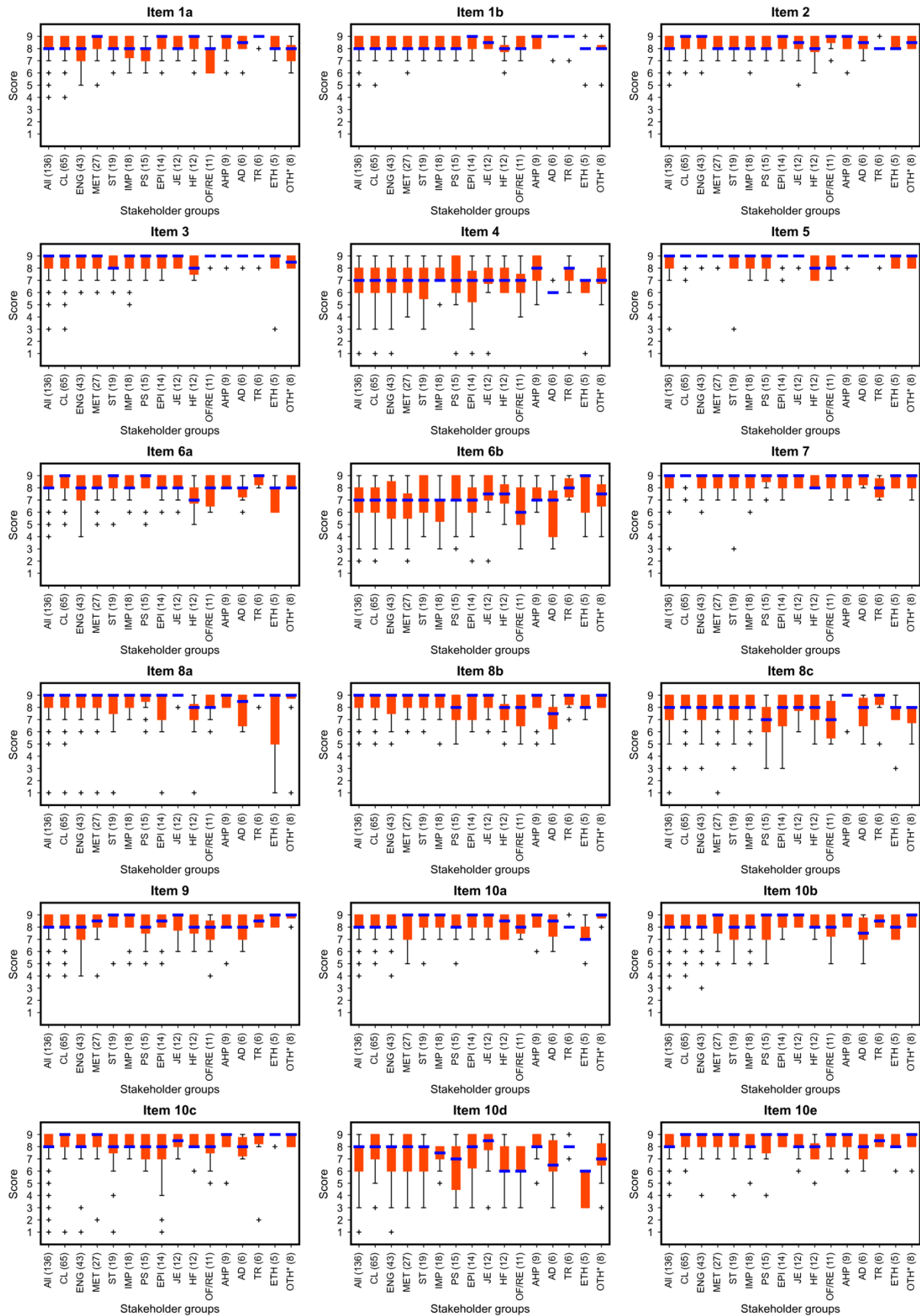
## Supplementary materials



**Suppl. Figure V-1c: median scores and IQR of item 37 to 54 during the first round of Delphi, overall and by stakeholder group.** Whiskers are the last value comprised within 1.5 IQR on both side of the median. Crosses are outliers. A participant can self-affiliate to several stakeholder groups. \*Others: Funders, Patients representatives, Payers & Commissioners, and Psychologists. AD = Administrators; AHP = Allied Health Professionals; CL = Clinicians; ENG = Engineers & Computer Scientists; EPI = Epidemiologists; ETH = Ethicists; HF = Human Factors Specialists; IMP = Implementation Scientists; JE = Journal Editors; OF = Policy Makers & Official Institutions; PM = Policy Makers; PS = Entrepreneurs & Private Sector Representatives; RE = Regulators; TR = Trialists.



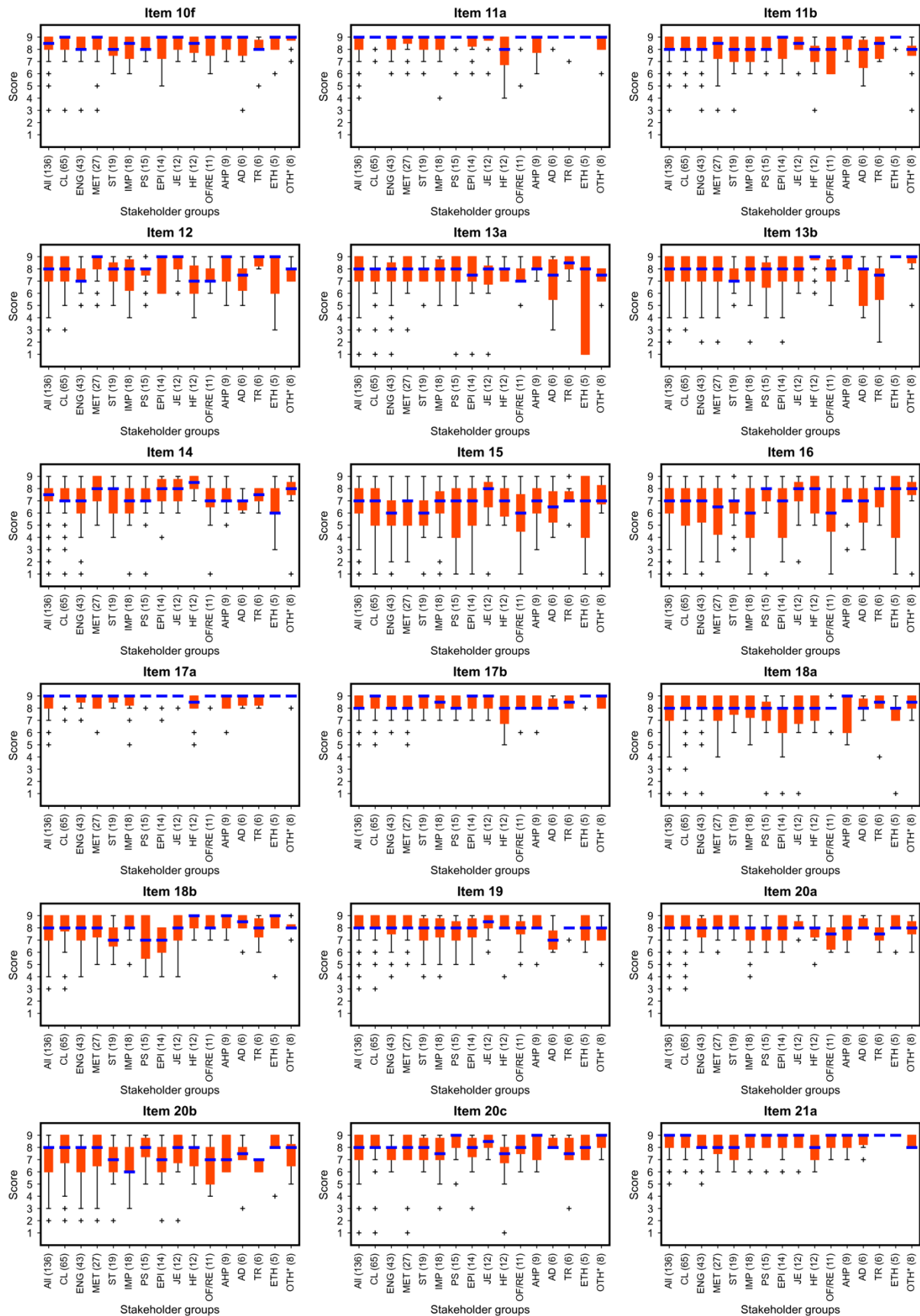
## Supplementary materials



**Suppl. Figure V-2a: median scores and IQR of item 1a to 10e during the second round of Delphi, overall and by stakeholder group.** Whiskers are the last value comprised within 1.5 IQR on both side of the median. Crosses are outliers. A participant can self-affiliate to several stakeholder groups. \*Others: Funders, Patients representatives, Payers & Commissioners, and Psychologists. AD = Administrators; AHP = Allied Health Professionals; CL = Clinicians; ENG = Engineers & Computer Scientists; EPI = Epidemiologists; ETH = Ethicists; HF = Human Factors Specialists; IMP = Implementation Scientists; JE = Journal Editors; OF = Policy Makers & Official Institutions; PM = Policy Makers; PS = Entrepreneurs & Private Sector Representatives; RE = Regulators; TR = Trialists.

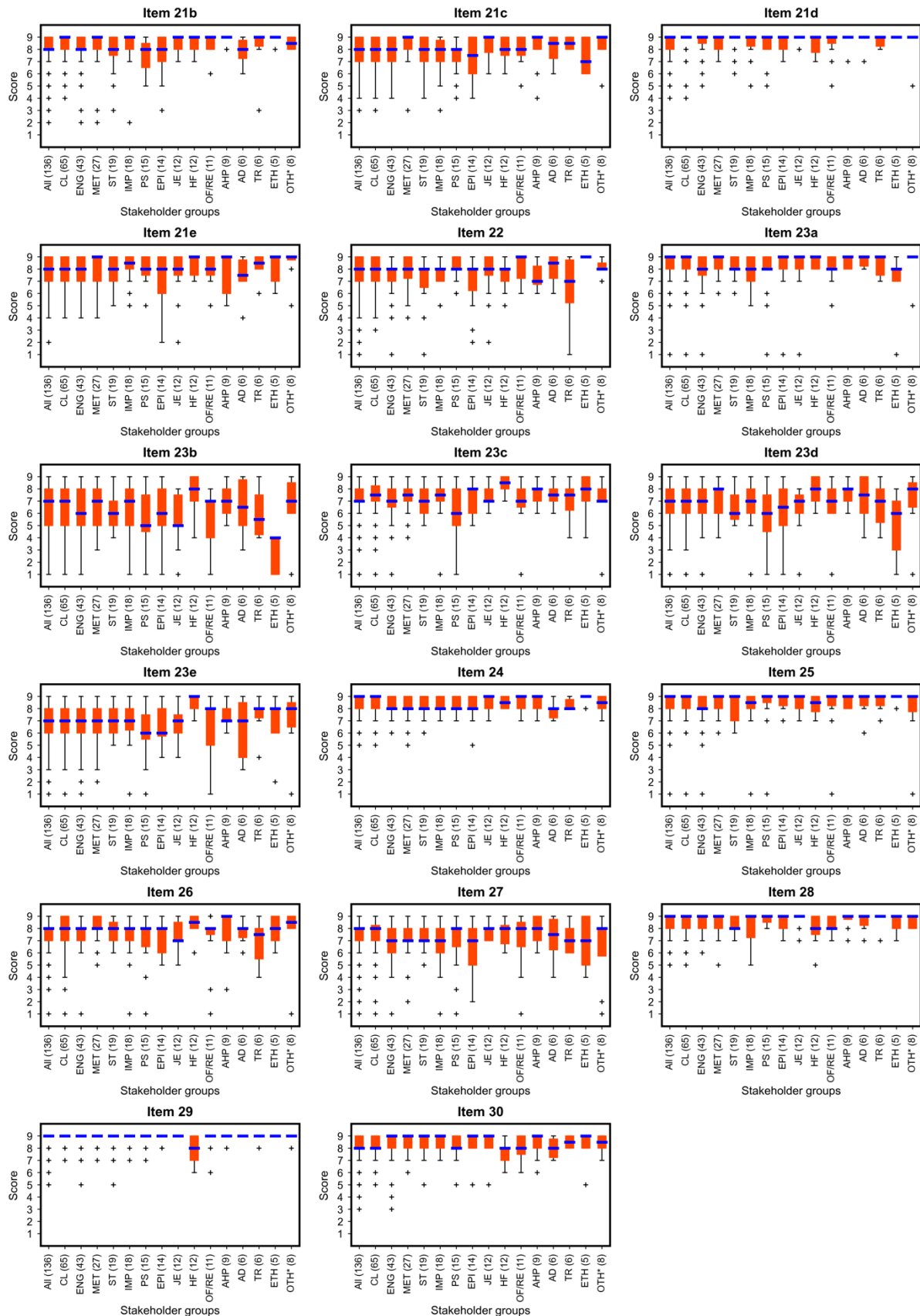


## Supplementary materials



**Suppl. Figure V-2b: median scores and IQR of item 10f to 21a during the second round of Delphi, overall and by stakeholder group.** Whiskers are the last value comprised within 1.5 IQR on both side of the median. Crosses are outliers. A participant can self-affiliate to several stakeholder groups. \*Others: Funders, Patients representatives, Payers & Commissioners, and Psychologists. AD = Administrators; AHP = Allied Health Professionals; CL = Clinicians; ENG = Engineers & Computer Scientists; EPI = Epidemiologists; ETH = Ethicists; HF = Human Factors Specialists; IMP = Implementation Scientists; JE = Journal Editors; OF = Policy Makers & Official Institutions; PM = Policy Makers; PS = Entrepreneurs & Private Sector Representatives; RE = Regulators; TR = Trialists.

## Supplementary materials



**Suppl. Figure V-2c: Median scores and IQR of item 21b to 30 during the second round of Delphi, overall and by stakeholder group.** Whiskers are the last value comprised within 1.5 IQR on both side of the median. Crosses are outliers. A participant can self-affiliate to several stakeholder groups. \*Others: Funders, Patients representatives, Payers & Commissioners, and Psychologists. AD = Administrators; AHP = Allied Health Professionals; CL = Clinicians; ENG = Engineers & Computer Scientists; EPI = Epidemiologists; ETH = Ethicists; HF = Human Factors Specialists; IMP = Implementation Scientists; JE = Journal Editors; OF = Policy Makers & Official Institutions; PM = Policy Makers; PS = Entrepreneurs & Private Sector Representatives; RE = Regulators; TR = Trialists.

## Chapter VI

### Suppl. Note VI-1: DECIDE-AI Consensus meeting – participants' names and affiliations

#### *With voting rights (n=16):*

Baptiste Vasey (Nuffield Department of Surgical Sciences, University of Oxford, Oxford, UK); Bart Geerts (Healthplus.ai BV, Amsterdam, NL); Bilal A. Mateen (The Wellcome Trust, London, UK); Campbell Rogers (HeartFlow Inc., Redwood City, CA, USA); Daniel S.W. Ting (Singapore National Eye Center, Singapore Eye Research Institute, Singapore, SG); Gary S. Collins (Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology & Musculoskeletal Sciences, University of Oxford, Oxford, UK); Johan Ordish (The Medicines and Healthcare products Regulatory Agency, London, UK); Lauren Morgan (Morgan Human Systems Ltd, Shrewsbury, UK); Melissa D. McCradden (The Hospital for Sick Children, Toronto, CA); Peter McCulloch (Nuffield Department of Surgical Sciences, University of Oxford, Oxford, UK); Peter Wheatstone (Patient representative); Piyush Mathur (Cleveland Clinic, Cleveland, Ohio, USA); Spiros Denaxas (Institute of Health Informatics, University College London, London, UK); Suchi Saria (Departments of Computer Sciences, Statistics, and health Policy, and Division of Informatics, Johns Hopkins University, Baltimore, MD, USA); Wim Weber (The BMJ, London, UK); Xiaoxuan Liu (University Hospitals Birmingham NHS Foundation Trust, Birmingham, UK);

#### *Chair (n=1):*

Bruce Campbell (University of Exeter Medical School, Exeter, UK)

#### *Observers (n=2):*

Myura Nagendran (Imperial College London, London, UK), Livia Faes (Moorfields Eye Hospital, London, UK)

### Suppl. Note VI-2: DECIDE-AI piloting – participants' names and affiliations

Ali Connell (Google Health, London, UK); Dinesh V. Gunasekaran (Singapore National Eye Center, Singapore Eye Research Institute, Singapore, SG); Falk Schwendicke (Charité Universitätsmedizin Berlin, Berlin, DE); Hani J. Marcus (Department of Neurosurgery, National Hospital for Neurology and Neurosurgery, Queen Square, London, UK); Jean-Emmanuel Bibault (Radiation Oncology Department, Hôpital Européen Georges Pompidou, AP-HP, Paris, FR); Laurence B. Lovat (University College London, London, UK); Lisa Baker (Mima, London, UK); Mara Giavina-Bianchi (Hospital Israelita Albert Einstein, São Paulo, BR); Matthew Woodward (THIS Institute, School of Clinical Medicine, University of Cambridge, Cambridge, UK); Oliver Redfern (Kadoorie Centre for Critical Care Research and Education, Nuffield Department of Clinical Neurosciences, University of Oxford, Oxford, UK); Omer F. Ahmad (Wellcome/EPSRC centre for Interventional & Surgical Sciences (WEISS), University College London, London, UK); Rachel Barnett (Healthplus.ai B.V., Amsterdam, NL); Siri L. van der Meijden (Healthplus.ai B.V., Amsterdam, NL and Leiden University Medical Center, Leiden, NL); Tien-En Tan (Singapore National Eye Center, Singapore Eye Research Institute, Singapore, SG); Wei Yan Ng (Singapore National Eye Center, Singapore Eye Research Institute, Singapore, SG); Yoonyoung Park (Center for Computational Health, IBM Research, Cambridge, MA USA).

# Supplementary materials

	included	vote	Participation	main	main (%)	suppl.	suppl. (%)	blank	blank (%)	Results list
Item 1a	yes	14	93.3	13	93.33	1	6.67	0	0	main
Item 1b	yes	14	100.0	4	28.57	10	71.43	0	0	supplementary
Item 2	yes	14	93.3	10	71.43	4	28.57	0	0	main
Item 3	yes	14	93.3	12	85.71	2	14.29	0	0	main
Item 4	no	NA	NA	NA	NA	NA	NA	NA	NA	NA
Item 5	yes	14	93.3	6	42.86	8	57.14	0	0	supplementary
Item 6a	yes	15	100.0	3	20	12	80	0	0	supplementary
Item 6b	no	NA	NA	NA	NA	NA	NA	NA	NA	NA
Item 7	no	NA	NA	NA	NA	NA	NA	NA	NA	NA
Item 8a	yes	14	100.0	3	21.43	10	71.43	1	7.14	supplementary
Item 8b	yes	14	100.0	5	35.71	9	64.29	0	0	supplementary
Item 8c	yes	13	92.9	12	92.31	1	7.69	0	0	main
Item 9	yes	14	100.0	12	85.71	2	14.29	0	0	main
Item 10a	yes	15	100.0	13	86.7	2	13.3	0		main
Item 10b+c	yes	15	100.0	15	100	0	0	0	0	main
Item 10d	no	NA	NA	NA	NA	NA	NA	NA	NA	NA
Item 10e	yes	15	100.0	13	86.67	2	13.33	0	0	main
Item 10f	yes	14	93.3	14	100	0	0	0	0	main
Item 11a	yes	14	93.3	5	35.71	9	64.29	0	0	supplementary
Item 11b	yes	13	92.9	12	92.31	1	7.69	0	0	main
Item 12	yes	14	100.0	9	64.29	5	35.71	0	0	main
Item 13a+b	yes	13	92.9	12	92.31	1	7.69	0	0	main
Item 13a+b - revote	yes	14	100.0	11	78.57	3	21.43	0	0	main
Item 14	yes	14	100.0	12	85.71	2	14.29	0	0	main
Item 15	yes	15	100.0	1	6.67	14	93.33	0	0	supplementary
Item 16	yes	14	93.3	10	71.43	4	28.57	0	0	main
Item 17a	yes	14	100.0	3	21.43	11	78.57	0	0	supplementary
Item 17b	yes	13	92.9	4	30.77	9	69.23	0	0	supplementary
Item 18a	yes	15	100.0	12	80	3	20	0	0	main
Item 18b	no	NA	NA	NA	NA	NA	NA	NA	NA	NA
Item 18b - revote	yes	14	100.0	14	100	0	0	0	0	main
Item 19	yes	14	93.3	13	92.86	1	7.14	0	0	main
Item 20a+c	yes	14	100.0	6	42.86	8	57.14	0	0	supplementary
Item 20b	no	NA	NA	NA	NA	NA	NA	NA	NA	NA
Item 21a+b	yes	13	92.9	8	61.54	5	38.46	0	0	main
Item 21c	no	Na	NA	NA	NA	NA	NA	NA	NA	NA
Item 21c+d+e	yes	12	85.7	12	100	0	0	0	0	main
Item 22	yes	14	100.0	7	50	7	50	0	0	draw*
Item 23a	yes	14	100.0	14	100	0	0	0	0	main
Item 23b	no	NA	NA	NA	NA	NA	NA	NA	NA	NA

## Supplementary materials

<b>Item 23c</b>	yes	14	100.0	13	92.86	1	7.14	0	0	main
<b>Item 23d</b>	yes	14	100.0	12	85.71	2	14.29	0	0	main
<b>Item 23e</b>	no	NA	NA	NA	NA	NA	NA	NA	NA	NA
<b>Item 24</b>	yes	14	100.0	12	85.71	2	14.29	0	0	main
<b>Item 25</b>	yes	14	100.0	12	85.71	2	14.29	0	0	main
<b>Item 26</b>	no	NA	NA	NA	NA	NA	NA	NA	NA	NA
<b>Item 27</b>	no	NA	NA	NA	NA	NA	NA	NA	NA	NA
<b>Item 28</b>	yes	14	100.0	2	14.29	12	85.71	0	0	supplementary
<b>Item 29</b>	yes	14	100.0	1	7.14	13	92.86	0	0	supplementary
<b>Item 30</b>	yes	14	100.0	10	71.43	4	28.57	0	0	main

**Suppl. Table VI-1: list attribution votes.** No list attribution votes were held for items which were not included (NA). Suppl. = supplementary list (later renamed good research practice list); \* was later attributed to the supplementary list.

## Supplementary materials

Theme	Item n°	Recommendation
<b>Title and abstract</b>		
Title and abstract	1	Identify the study as early clinical evaluation of a decision support system based on artificial intelligence or machine learning, specifying the problem addressed.
<b>Introduction</b>		
Target problem and population	2	Describe the target problem and medical condition, including the current standard practice, and the target patient population.
Intended use	3	Describe the intended use of the algorithm, its planned integration in the care pathway and the potential impact, including patient outcomes, it intends to achieve.
<b>Methods</b>		
Participants	4	a) Describe how patients were recruited, stating the inclusion and exclusion criteria at both patient and data level, and how the number of recruited patients was decided.
		b) Describe how users were recruited, stating the inclusion and exclusion criteria, and how the intended number of recruited users was decided.
		c) Describe steps taken to familiarise the users with the algorithm, including any training received prior to the study.
Algorithm	5	a) Briefly describe the algorithm, specifying the version and type of model used. Describe, or provide a direct reference to, the characteristics of the patient population on which the algorithm was trained and its expected performance from development/validation studies.
		b) Identify the data used as inputs. Describe how the data were acquired, the process needed to enter the input data, any pre-processing applied and how missing/low-quality data were handled.
		c) Describe the algorithm outputs and how they were presented to the users (an image may be useful).
Implementation	6	a) Describe the settings in which the algorithm was evaluated, including which additional clinical information (i.e. not provided by the algorithm) was accessible to the users to interpret or put into context the output of the algorithm.
		b) Describe the clinical workflow/pathway in which the algorithm was evaluated, the timing of its use, how the final decision was reached and by whom.
Safety and errors	7	a) Provide a description of how significant errors/malfunctions were defined and identified.
		b) Describe how any risks to patient safety or instances of harm were identified, analysed, and minimised.
Human factors	8	Describe the human factors tools, methods or frameworks used, the use cases considered, and the users involved.
Ethics consideration	9	Describe whether specific methodology was utilised to fulfil an ethics-related goal (such as algorithmic fairness) and its rationale.
<b>Results</b>		
Participants	10	a) Describe the baseline characteristics of the patients included in the study, and report on input data availability.
		b) Describe the baseline characteristics of the users included in the study.
Implementation	11	a) Report on the user exposure to the algorithm, on the number of instances the algorithm was used and on the users' adherence to the intended implementation.
		b) Report any significant changes to the clinical workflow or patient pathway caused by the algorithm.
Modifications	12	Report any changes made to the algorithm or its hardware platform during the study. Report the timing of these modifications, the rationale for them, and the change in outcomes observed after each of them.

## Supplementary materials

Human-computer agreement	13	Report on the user agreement with the algorithm. Describe any instances of and reasons for user variation from the algorithm’s recommendations and, if applicable, users changing their mind based on the algorithm recommendations.
Safety and errors	14	a) List any significant errors/malfunctions related to: algorithm recommendations, supporting software/hardware, or users. Include details of: (i) rate of occurrence, (ii) apparent causes, (iii) whether they could be corrected, and (iv) any significant potential impacts on patient care.
		b) Report on any risks to patient safety or observed instances of harm (including indirect harm) identified during the study.
Human factors	15	a) Report on the usability evaluation, according to recognised standards or frameworks.
		b) Report on the user learning curves evaluation.
Discussion		
Support intended purpose	16	Discuss whether the results obtained support the intended use of the algorithm in clinical settings.
Safety and errors	17	Discuss what the results suggest about the safety profile of the algorithm. Discuss the observed errors/malfunctions and instances of harm, their implications for patients and whether/how they can be mitigated.
Statements		
Data Availability	18	Disclose if and how data and relevant code are available.

Suppl. Table VI-2: consensus list – AI-specific items.

Theme	Item n°	Recommendation
<b>Title and abstract</b>		
Title and abstract	I	Provide a structured summary of the study. Consider including: target condition and problem, intended use of the algorithm, type of algorithm, study setting, number of patients and users included, primary and secondary outcomes, key safety endpoints, human factors evaluated, main results, conclusions.
<b>Introduction</b>		
Objectives	II	State the study objectives.
<b>Methods</b>		
Research governance	III	Provide a reference to any study protocol, study registration number and ethics approval.
Outcomes	IV	Specify the primary and secondary outcomes measured.
Analysis	V	Describe the statistical methods by which the primary and secondary outcomes were analysed, as well as any prespecified additional analyses, including subgroup analyses and their rationale.
Patient engagement	VI	State how patients were involved in any aspect of the study design, conduct or in the development of the research question.
<b>Results</b>		
Main results	VII	Report on the prespecified outcomes, including outcomes for any comparison group if applicable.
Subgroup analysis	VIII	Report on the difference in the main outcomes according to the prespecified subgroups.
<b>Discussion</b>		
Strength and limitations	IX	Discuss the strengths and limitations of the study.
<b>Statements</b>		
Conflicts of interest	X	Disclose any relevant conflict of interest, including: the source of funding for the study, the role of funders, any other role played by commercial companies and authors' conflicts of interest.

Suppl. Table VI-3: consensus list – good research practice items.



## Supplementary materials

<b>AI system</b>	Decision support system incorporating AI and consisting of: (i) the artificial intelligence or machine learning algorithm; (ii) the supporting software platform; and (iii) the supporting hardware platform.
<b>AI system version</b>	Unique reference for the form of the AI system and the state of its components at a single point in time. Allows for tracking changes to the AI system over time and comparing between different versions.
<b>Algorithm</b>	Mathematical model responsible for learning from data and producing an output.
<b>Artificial intelligence (AI)</b>	"Science of developing computer systems which can perform tasks normally requiring human intelligence" <sup>26</sup> .
<b>Bias</b>	"Systematic difference in treatment of certain objects, people, or groups in comparison to others." <sup>59</sup>
<b>Care pathway</b>	Series of interactions, investigations, decision-making and treatments experienced by patients in the course of their contact with a healthcare system for a defined reason.
<b>Clinical</b>	Relating to the observation and treatment of actual patients rather than <i>in silico</i> or scenario-based simulations.
<b>Clinical evaluation</b>	Set of ongoing activities, analysing clinical data and using scientific methods, to evaluate the clinical performance, effectiveness and/or safety of an AI system, when used as intended <sup>35</sup> .
<b>Clinical investigation</b>	Study performed on one or more human subjects to evaluate the clinical performance, effectiveness and/or safety of an AI system <sup>60</sup> . This can be performed in any setting (e.g. community, primary care, hospital).
<b>Clinical workflow</b>	Series of tasks performed by healthcare professionals in the exercise of their clinical duties.
<b>Decision support system</b>	System designed to support human decision-making by providing person- and situation-specific information or recommendations, to improve care or enhance health.
<b>Exposure</b>	State of being in contact with, and having used, an AI system or similar digital technology.
<b>Human-computer interaction</b>	Bidirectional influence between human users and digital systems through a physical and conceptual interface.
<b>Human factors</b>	Also called ergonomics. "The scientific discipline concerned with the understanding of interactions among humans and other elements of a system, and the profession that applies theory, principles, data and methods to design in order to optimise human well-being and overall system performance." (International Ergonomics Association)
<b>Indication for use</b>	Situation and reason (medical condition, problem and patient group) where the AI system should be used.
<b><i>In silico</i> evaluation</b>	Evaluation performed via computer simulation outside the clinical settings.
<b>Intended use</b>	Use for which an AI system is intended, as stated by its developers, and which serves as the basis for its regulatory classification. The intended use includes aspects of: the targeted medical condition, patient population, user population, use environment, mode of action.
<b>Learning curves</b>	Graphical plotting of user performance against experience <sup>61</sup> . By extension, analysis of the evolution of user performance with a task as exposure to the task increases. The measure of performance often uses other context-specific metrics as a proxy.
<b>Live evaluation</b>	Evaluation under actual clinical conditions, in which the decisions made have a direct impact on patient care. As opposed to "offline" or "shadow mode" evaluation where the decisions do not have a direct impact on patient care.
<b>Machine learning</b>	"Field of computer science concerned with the development of models/algorithms that can solve specific tasks by learning patterns from data, rather than by following explicit rules. It is seen as an approach within the field of AI" <sup>26</sup> .
<b>Participant</b>	Subject of a research study, on which data will be collected and from whom consent is obtained (or waived). The DECIDE-AI guideline considers that both patients and users can be participants.



## Supplementary materials

<b>Patient</b>	Person (or the digital representation of this person) receiving healthcare attention or using health services, and who is the subject of the decision made with the support of the AI system. <i>NB: DECIDE-AI uses the term “patient” pragmatically to simplify the reading of the guideline. Strictly speaking, a person with no health conditions who is the subject of a decision made about them by an AI-based decision support tool to improve their health and wellbeing or for a preventative purpose is not necessarily a “patient” per se.</i>
<b>Patient Involvement in research</b>	Research carried out ‘with’ or ‘by’ patients or members of the public rather than ‘to’, ‘about’ or ‘for’ them. (Adapted from the INVOLVE definition of “Public Involvement”)
<b>Standard practice</b>	Usual care currently received by the intended patient population for the targeted medical condition and problem. This may not necessarily be synonymous with the state-of-the-art practice.
<b>Usability</b>	“Extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency, and satisfaction in a specified context of use” <sup>62</sup> .
<b>User</b>	Person interacting with the AI system to inform their decision making. This person could be a healthcare professional or a patient.

**Suppl. Table VI-4: glossary of terms.** The definitions given pertain to the specific context of DECIDE-AI and the use of the terms in the guideline. They are not necessarily generally accepted definitions and might not always be fully applicable to other areas of research.

## Supplementary Files

---

- Suppl. File III-1: Systematic review: full extraction table with conflicts resolved
- Suppl. File IV-1: A patient similarity-based approach to postoperative complications – full code (Python Notebook and HTML)
- Suppl. File IV-2: Data extraction and pre-processing – full code (SQL)
- Suppl. File IV-3: Prescriptions and procedures grouping and inclusion

## List of Annexes

---

Annex II-1:	Study on clinician cognitive process and desired computerised support needs – study protocol
Annex II-2:	Study on clinician cognitive process and desired computerised support needs – participant information sheet
Annex II-3:	Study on clinician cognitive process and desired computerised support needs – staff consent form
Annex III-1:	Systematic review – study protocol
Annex V-1:	DECIDE-AI presentation document – invited experts
Annex V-2:	DECIDE-AI presentation document – patient representatives
Annex V-3:	DECIDE-AI participant information sheet: Delphi round 1 and 2
Annex V-4:	DECIDE-AI Round 1 – questionnaire
Annex V-5:	DECIDE-AI Round 1 – executive summary
Annex V-6:	DECIDE-AI Round 1 – per item summary
Annex V-7:	DECIDE-AI Round 1 – thematic analysis
Annex V-8:	DECIDE-AI Round 1 – summary of proposed new items
Annex V-9:	DECIDE-AI Round 2 – questionnaire
Annex V-10:	DECIDE-AI Round 2 – per item summary
Annex VI-1:	DECIDE-AI piloting form with results
Annex VI-2:	DECIDE-AI Consensus meeting minutes
Annex VI-3:	Post piloting modifications to the item list
Annex VI-4:	DECIDE-AI checklist
Annex VI-5:	DECIDE-AI E&E document