International Telecommunication Union

ITU-T FG-AI4H Deliverable

TELECOMMUNICATION
STANDARDIZATION SECTOR
OF ITU

15 September 2023

PRE-PUBLISHED VERSION

DEL10.21

FG-AI4H Topic Description Document for the Topic Group on musculoskeletal medicine (TG-MSK)

ITU-T

International Telecommunication Union

Summary

This topic description document (TDD) specifies a standardised benchmarking for AI-based Musculoskeletal Medicine applications. It covers scientific, technical, and administrative aspects relevant for setting up this benchmarking.

Keywords

Artificial intelligence; health; topic groups; overview; ethics; regulations; data quality; data audit; clinical relevance; topic description; Musculoskeletal medicine

Change Log

This document contains Version 1 of the Deliverable DEL10.21 on "*FG-AI4H Topic Description Document for the Topic Group on musculoskeletal medicine (TG-MSK)*" approved on 15 September 2023 via the online approval process for the ITU-T Focus Group on AI for Health (FG-AI4H).

Editor: Peter Grinbergs
EQL, UK
E-mail: tgmskorg@googlegroups.com

Dr Mark Elliott
University of Warwick, UK

Contributors: (in alphabetical order)

Danielle Chulan
Connect Health
UK

Nick Downing
Vita Health Group
UK

Mark Elliott
University of Warwick
UK

Peter Grinbergs
EQL
UK
E-mail: peter@eql.ai

Michael Guard
EQL
UK

Joseph LeMoine
prIME Assessments
Canada

Yura Perov
Individual Contributor
UK

Kate Ryan
EQL
UK
E-mail: kate@eql.ai

Christopher Tack
NHS
UK
(Christopher Tack stopped being a member on his request on the 12th of May 2021.)

Olalekan Uthman
University of Warwick
UK

CONTENTS

Page

Page

List of Tables

Page

List of Figures

Page

FG-AI4H Deliverable DEL10.21

FG-AI4H Topic Description Document for the Topic Group on musculoskeletal medicine (TG-MSK)

1 Important General Note

Important general note: This work has been a joint effort. Contributions to it have been made by different people. Contributions have not been necessarily checked, verified, peer reviewed, etc. There has not been necessarily an editor, and the topic driver(s) have not necessarily been editors. Participation and work by a member of the topic group (including topic drivers) does not necessarily mean an endorsement of or agreement with contribution(s).

2 Introduction

This topic description document specifies the standardised benchmarking for AI systems for MSK Medicine. It serves as deliverable DEL10.21 of the ITU/WHO Focus Group on AI for Health (FG-AI4H).

This topic group is dedicated to AI/ML applications for MSK medicine. It is dedicated to establishing a standardised benchmarking guidelines (including specifications of input data and outputs of AI systems for different AI tasks for MSK medicine) and potentially creating a prototype of a benchmarking platform for AI/ML application in Musculoskeletal medicine. The topic group focuses on prevention strategies, triage[1] (in particular identifying urgency), diagnosis, prognosis and treatment of musculoskeletal (MSK) conditions with the applications of artificial intelligence (AI) and machine learning (ML) approaches including computer vision (CV), augmented and virtual reality (AR/VR), natural language processing (NLP)/understanding and other approaches.

Primary prevention: early risk assessment, prognosis, risk detection of MSK trauma/deterioration and movement deficiencies using ML, CV, NLP to parse a patient's input, as well as to incorporate existing electronic health records (EHR) and data analysis (including data from wearables with the patients' consent).

Triage and diagnosis: assist in identifying the causes of a patient's signs and symptoms including pain, with the use of chatbots and similar approaches as for **primary prevention**.

Treatment: use of AI with CV and AR to enable self-management and, where clinician's guidance/oversight/involvement is required, to assist in such management. AR and CV technology provide more effective treatment and improve patient engagement and experience with the help of speech-to-text and text-to-speech capabilities (in combination with the use of common technology by showing exercise reminders for example).

3 Relevance of the topic group

Painful MSK conditions affect 20-33% of the world's population [1]. According to the WHO, "MSK conditions are the leading contributor to disability worldwide, with low back pain being the single leading cause of disability globally. ... MSK conditions significantly limit mobility and dexterity, leading to early retirement from work, reduced accumulated wealth and reduced ability to participate in social roles. The greatest proportion of non-cancer persistent pain conditions is accounted for by MSK conditions. ... MSK conditions are commonly linked with depression and increase the risk of developing other chronic health conditions" [1].

Up to 30% of consultations carried out by primary care doctors in the UK (as an example) are for MSK conditions [2]. Together with the worldwide shortage of health professionals (including

[1] Note that there are other definitions, in particular in relation to MSK medicine. One task of the topic group is to define and investigate this further.

doctors and physiotherapists) [3], it is clear there is a pressing need to introduce, support and grow the potential use of reliable, safe, accurate solutions powered by AI and ML which is evidence-informed and co-produced with lived experience. This need exists across the world and the solutions must be accessible and affordable in order to provide universal coverage. The latter is especially important in the light of existing inequalities: AI applications have the power to reduce them but it also should be ensured that they do not worsen any inequalities.

There have been several developments in the last few years that are particularly relevant for this area:

- The development of the next generation of CV and NLP techniques. (In particular, recent CV technology that allows fairly accurate pose recognition using just one camera e.g. a smartphone camera, without the need for special equipment.)
- The spread of mobile devices with high-resolution cameras and with powerful microprocessors.
- The spread of wearable technology and the resulting accumulated data.

4 Impact

Artificial intelligence and technology has the potential to enable more affordable, accessible and accurate diagnostics, prevention and care for people across the world who are either at risk of developing, or who have existing MSK conditions.

The use of AI for MSK conditions and physiotherapy (physical therapy) could provide (and is already doing so in limited, early settings) rapid access to the required prevention and care for the patients in need, especially those patients in some regions or countries who can't currently access such care. It also facilitates the work of clinicians, for example by identifying accelerated exercise-informed rehabilitation pathways and improving objective testing of patient movement abilities using CV and AR capabilities. In addition, it has the potential to reduce the burden on clinicians and healthcare systems by autonomously (or semi-autonomously in sync with clinicians) providing patients with triage, diagnosis, or treatment care where appropriate — allowing clinicians to focus on more complex or less typical presentations and other clinical work. This is especially important at present, because of the global shortage of health professionals [3].

It is vital to develop and maintain a set of diversified and robust benchmarks to ensure accurate, safe, scalable solutions that are applicable for different patient groups with varying needs, depending on their specific MSK conditions.

5 About the FG-AI4H topic group on MSK Medicine

The introduction highlights the potential of a standardised benchmarking of AI systems for Musculoskeletal Medicine to help solving important health issues and provide decision-makers with the necessary insight to successfully address these challenges.

To develop this benchmarking framework, FG-AI4H decided to create the TG-MSK at the meeting J (meeting #10), which was conducted online from the 30th of September to the 2nd of October 2020.

FG-AI4H assigns a *topic driver* to each topic group (similar to a moderator) who coordinates the collaboration of all topic group members on the TDD. During FG-AI4H meeting #10, which was conducted online from the 30th of September to the 2nd of October 2020, Yura Perov from EQL (UK) was nominated as topic driver for the TG-MSK. Since April 2021, Yura Perov was an individual contributor from the UK. Yura Perov stopped being a Topic Driver during Meeting R (March 2023). It was also suggested that Peter Grinbergs (EQL, UK) become a topic driver too so that Peter and Yura can effectively topic drive the topic group; Peter was provisionally working as a co-topic-driver. Since meeting L, Peter Grinbergs is also a topic driver of the topic group. Dr Mark Elliott a member of the Topic Group, took over as co-topic-driver with Peter Grinbergs in March 2023, following approval from the FG-AI4H committee.

5.1 Documentation

This document is the TDD for the TG-MSK. It introduces the health topics including the AI tasks, outlines their relevance and the potential impacts that the benchmarking will have on health systems and patient outcome, and provides an overview of the existing AI solutions for MSK Medicine. It describes the existing approaches for assessing the quality of AI-based MSK Medicine systems/approaches and provides the details that are likely relevant for setting up a new standardised benchmarking. We expect to specify the actual benchmarking methods for all subtopics at a level of detail that includes technological and operational implementation. There are individual subsections for all versions of the benchmarking. Finally, it summarises the results of the topic group's benchmarking initiative and benchmarking runs. In addition, the TDD addresses ethical and regulatory aspects.

The TDD will be developed cooperatively by all members of the topic group over time and updated TDD iterations are expected to be presented at each FG-AI4H meeting.

The final version of this TDD will be released as deliverable "DEL10.21 MSK Medicine (TG-MSK)." The topic group is expected to submit input documents reflecting updates to the work on this deliverable **(Table 1)** to each FG-AI4H meeting.

Table 1: Topic Group output documents

Number	Title
FGAI4H-S-026-A01	Latest update of the Topic Description Document of the TG-MSK
FGAI4H-S-026-A02	Latest update of the Call for Topic Group Participation (CfTGP)
FGAI4H-S-026-A03	Presentation slides for Meeting S

The working version of this document can be found in the official topic group SharePoint directory.

- https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/tg/SitePages/TG-MSK.aspx

5.2 Status of this topic group

The following subsections describe the update of the collaboration within the TG-MSK for the official focus group meetings.

5.2.1 Status update for meeting S

- The information was received that the FG-AI4H would end in September 2023 and become a Global Initiative.
- No further long term work was possible due to the limited time available, and the focus of the remaining time was spent on finalising the documents.

5.2.2 Status update for meeting R

- It is anticipated that Dr Mark Elliott (University of Warwick) will become a topic driver of this topic group. This is expected to be formalised at Meeting R (Dr Elliott has been provisionally working as a topic driver since January 2023).
- Yura Perov is going to stop being a topic driver of the topic group. This is expected to happen and be formalised at (or after) Meeting R of the Focus Group.
- We are working towards a sub-theme within the topic group, that will focus on benchmarking and standardisation of biomechanics data for AI applications.
- A list of papers extracted from a systematic review covering AI4MSK has been identified by TG member, Joseph LeMoine. These are being assessed by the TG. See: https://docs.google.com/spreadsheets/d/1T3jaFN-Ls2eTYNQZWqzu5lwgPBoaS8tB05os12NO72s/edit?usp=sharing

- 3 meetings have been held since Meeting Q:
  - 14th December 2022: https://docs.google.com/document/d/1wx6ANlPxlehoMHLWda3zRucVAU4isy0ajzoGUaQw9mo/edit?usp=sharing
  - 25th January 2023: https://docs.google.com/document/d/1lprfDTH0ef8EyTEZ--QarONTFTgaZuRCGUY-mBhRLK0/edit?usp=sharing
  - 28th February 2023: https://docs.google.com/document/d/13XU_3H9Q1zNAEe_V23Re4O1TR9uvEf_YCunND08dXZc/edit?usp=sharing

5.2.3 Status update for meeting Q

Some updates:

- Emma Meehan started being a member (as per an email from Emma dated the 20th of September 2022).
- There are 12 members in the topic group.
- "More information about Prognosis including some Case Studies" has been updated (based on an update in https://docs.google.com/document/d/1Z5AA2mhVYBFgORIT-WKLJSksRb5TSYZRSS9EFuUcZWY/edit).
- Work has been done during meetings and outside of them.
- An update, as provided by Peter Grinbergs:
  - We've taken the decision to investigate and identify an AI/ML solution against which we could conduct an abstract implementation of our proposed benchmarking solution. We feel this would effectively "stress test" and further define the required methodology including analysis, development of the required model and relevant variables against which the system will be evaluated. Secondary to this we proposed to review real world data in an attempt to identify the necessary variables against which this solution could be evaluated, inclusive of exploring ways in which to expand upon this real world data set e.g. creating further synthetic data to increase/enhance the breadth and depth of data. We have identified a collaborator who may be in a position to provide real world data
- The audit-related work is being paused, most likely. It may be continued when, e.g., there is at least one specific real solution (like a product/system) that the topic group is 'directly' 'working' with.
- Three topic group meetings took place:
  - 13th of September 2022: https://docs.google.com/document/d/1ISnfQl73RKDeDF1nKgjQ1VaSayOdn3N2VHiWWRKb7Bo/edit
  - 18th of October 2022: https://docs.google.com/document/d/1wjOxACySLZI0FrCo0FuocH5rad6n4-vQebtnU7-mgyg/edit
  - 22nd of November 2022: https://docs.google.com/document/d/1m3liYsoSCvetPsR3iaJAyMZr78y3VLJSIBKnYnKD-RQ/edit
- Yura Perov is planning to stop being one of the topic drivers. However, the transition has not finished, so Yura is going to remain being one of the topic drivers for some time.

5.2.4 Status update for meeting P

Some updates:

- Azadur Rahman Sarker and Ashwini Sathnur started being members (as per their emails dated the 15th of August 2022).
- There are 11 members in the topic group.
- Six topic group meetings took place:
    - 7th of June 2022: https://docs.google.com/document/d/1HnrsCn4rhraZdH8XTu0uJWoSArrB0UTLoz6bX1oDqeU/edit
    - 21st of June 2022: https://docs.google.com/document/d/1AtZdcq1ZSW6vdM6gXUh4pKhHrsoVt5ev5cdikRu75ao/edit
    - 5th of July 2022: https://docs.google.com/document/d/120HXQDD-SM0Rwa6yhKqbkqRoqlH7PiuMphxH4FJet90/edit
    - 26th of July 2022: https://docs.google.com/document/d/12WRkhd1e2FcmgC_pOIMk5Qw9dfTsBWtIPQZV21n7Vg4/edit
    - 16th of August 2022: https://docs.google.com/document/d/1Z5nV3kkIUt5b2YARRzb2sYcW1sr4nJfHFLbvGTOIXJQ/edit
    - 30th of August 2022: https://docs.google.com/document/d/1X2rZxujBuiJVMk62NdlbWKLq-rz1xd2v1rUJNiDeqWw/edit
- The document https://docs.google.com/spreadsheets/d/1qIZYut9DzAkuTQYqA9aQJ4e5oxM8LnEZ7XPCvBm2DSs/edit#gid=0 with synthetic cases has been updated:
    - Column (BE) "Summary Prevalence" (that is the new name of the column) was updated. *Nota bene* - Michael Guard: 'unable to get one single population or global data, due to heterogeneity of presenting conditions'. Some other cells were updated. Those updates were made by Michael Guard.
    - Column (BB) "Coding for Primary provisional Diagnosis - ICD11" added. Values for it were added by Joseph LeMoine.
- Received comments/documents from Ashwini Sathnur. Some of them were discussed and there can be further discussion(s) as part of topic group meeting(s).
- Some work for the audit has been done.
    - Document "[PUBLIC] Initial answers for the audit (July 2022)": https://docs.google.com/document/d/1uSeIb3vnmYIEazGDBMQDmGBrhNNO1lvDfsyzbIfWuR0/edit#heading=h.6z1l97e7aj1k . Some of the contributors (including via at least one topic group meeting): Peter Grinbergs, Joseph LeMoine, Yura Perov.

5.2.5 Status update for meeting O (meeting #14)

Updates:

- Michael Guard started being a member (as per Michael's request dated the 25th of May 2022).
- Four topic group meetings took place:
    - 22nd of February 2022: https://docs.google.com/document/d/1BtYGbuwXhJ6_TWi0Bqu3ds_vTENZ6Jibjq5GfubP79E/edit
    - 22nd of March 2022: https://docs.google.com/document/d/13SW-IbF72rljaUpZ82m26kYM0OGRyGiVbSbSehW5aA8/edit
        - Including presentation 'A snapshot of MSK disorders in the Global Burden of Disease' by Dr Lidia Sanchez-Riera, MD, PhD.

    - 3rd of May 2022: https://docs.google.com/document/d/1_dKDN7cqZTNow6l5W7ygfI8b7-MS10YhvdruiM-mATY/edit
    - 24th of May 2022: https://docs.google.com/document/d/1FYXB4060SPH-6QghRf8jY8ssa6BZ7Kd_mY2V5FuasLE/edit
- The synthetic cases have been updated:
    - The new version 1.11 is available at https://docs.google.com/spreadsheets/d/1qIZYut9DzAkuTQYqA9aQJ4e5oxM8LnEZ7XPCvBm2DSs/edit#gid=0
    - Added more cases (11 cases added to the original 8 cases), spanning extra pathologies (inclusive of hip, wrist, shoulder, foot, hallux, rheumatology examples), added data for incidence, prevalence, average disability weight, YLD Global (%) and DALY (%). This update (for version 1.1) has been done by Michael Guard.
- The first version of the prototype (demo) has been developed and is available at https://github.com/perov/fgai4h-tg-msk-prototype

5.2.6 Status update for meeting N (meeting #13)

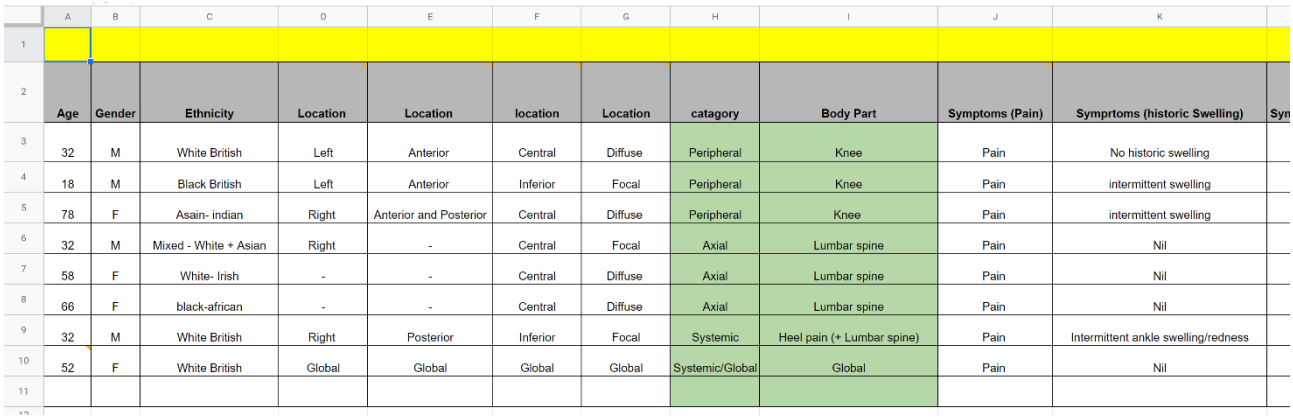
There are 8 members in the topic group.

Updates:

- Four topic group meetings took place (excluding cancelled one(s)):
    - 5th of October 2021: https://docs.google.com/document/d/1XNcv6QWpSF_mfhiiQ0xN_lYXqBqzKUKJH76yUR9kQtc/edit
    - 16th of November 2021: https://docs.google.com/document/d/1fqum98iim00GutiBHX1IwknjjPTud3yt6MbO6iCQwEg/edit
    - 7th of December 2021: https://docs.google.com/document/d/1fzUfFI00BjB5x5i8W_-EU6mComXXANp_UHzwwFkUjkU/edit
    - 1st of February 2022: https://docs.google.com/document/d/1iIJwMYRj8N-onxXpE_YiXioctbhu5h9d6xF1ytgd2j4/edit
- Several documents were created, some of the content of which was incorporated into the latest version of the topic description document. Some more details are found in Table 2.
- In particular, synthetic cases were created to support the benchmark work. A screenshot that shows some of that data is provided as Figure 1. These synthetic cases can be used (e.g., if they are exported into a JSON file) for a benchmark prototype.

Table 2: Recent documents

Document name	Link	The first version prepared by
Prognosis (2021)	https://docs.google.com/document/d/1Z5AA2mhVYBFgORIT-WKLJSksRb5TSYZRSS9EFuUcZWY/edit#heading=h.xoz8bub38lgv	Kate Ryan
Some info about case 'data format' (January 2022)	https://docs.google.com/document/d/1a0RoRrqRtgINpBxRtA7laBY31lWBImE3uPRDD2swg1g/edit#	Yura Perov
Synthetic cases	https://docs.google.com/spreadsheets/d/1ef8_v4H8uL9QGLAwoBMGC9N9wekQzDvUw8joT1CNNEE/edit#gid=574971709	Michael Guard and Kate Ryan



Age	Gender	Ethnicity	Location	Location	location	Location	catagory	Body Part	Symptoms (Pain)	Symptoms (historic Swelling)
32	M	White British	Left	Anterior	Central	Diffuse	Peripheral	Knee	Pain	No historic swelling
18	M	Black British	Left	Anterior	Inferior	Focal	Peripheral	Knee	Pain	intermittent swelling
78	F	Asain- indian	Right	Anterior and Posterior	Central	Diffuse	Peripheral	Knee	Pain	intermittent swelling
32	M	Mixed - White + Asian	Right	-	Central	Focal	Axial	Lumbar spine	Pain	Nil
58	F	White- Irish	-	-	Central	Diffuse	Axial	Lumbar spine	Pain	Nil
66	F	black-african	-	-	Central	Diffuse	Axial	Lumbar spine	Pain	Nil
32	M	White British	Right	Posterior	Inferior	Focal	Systemic	Heel pain (+ Lumbar spine)	Pain	Intermittent ankle swelling/redness
52	F	White British	Global	Global	Global	Global	Systemic/Global	Global	Pain	Nil

Figure 1: Some of the data of the created synthetic cases

5.2.7 Status update for meeting M (meeting #12)

There are 8 members in the topic group.

Updates:

- There were 5 topic group meetings (excluding 1 more meeting that did not happen because there was only one participant). The meeting notes can be found below:
  1. 26 May 2021: https://docs.google.com/document/d/13gdDUCOs5NKBFd8B60plWqJ10UpdECD3rqO_A4slQjI/edit
  2. 10 June 2021: https://docs.google.com/document/d/1LxI_Ffly_RJ5d16exk03fb5n7tMyiUFSSuEM17mGHC0/edit
  3. 13 July 2021: https://docs.google.com/document/d/1UlOWRnlmJTVooNDGuBquYnUqwYieoOh3uxh6oxoTl9o/edit
  4. 29 July 2021 (meeting with 1 participant): https://docs.google.com/document/d/1svxb6lO9ETg_AirZnmXPCYwjX0iaagM82pOaGeNbbrI/edit
  5. 31 August 2021: https://docs.google.com/document/d/10Nq9_nAoJ5C2xbZO-CpEW6ixW9C7ZwvcBX5qbFgU7nk/edit

6. 14 September 2021: https://docs.google.com/document/d/1hUJBxU9QgRVxon3WlyCmiTFhlMwis_pyFhJoVXLC6uw/edit

- Raj Sengupta participated in the meeting on 10 June 2021. Robert Pawinski participated in the meetings on 13 July 2021, 31 August 2021 and 14 July 2021.
- Danielle Chulan started being a member on her request dated 28 May 2021.
- Several documents were created, some of the content of which was incorporated into the latest version of the topic description document. Some more details are found in Table 3.

Table 3: Documents recently incorporated into the TDD

Document name	Link	Original contributor(s) (to the first version)
Existing AI solutions - Prediction - Ortho	https://docs.google.com/document/d/1Sdf9zuBBnOKtj73LTOR7lktOGa0BpIjxMMRH8G7TLx0/edit#heading=h.xoz8bub38lgv	Joseph LeMoine
Current gold standard - Prediction Musculoskeletal Health	https://docs.google.com/document/d/1cd0NLO7F9llIH6Pu1CZ8ih68LouDRr5M1VKCYHSu8Q/edit#heading=h.xoz8bub38lgv	Joseph LeMoine
Metrics (and related terms/notes) for the Prediction task (work-in-progress) (July 2021)	https://docs.google.com/document/d/1h7OBCSzQ_k0aKLendz4ErPjkqLeEv2mFvNpKYDkj2hc/edit#heading=h.xoz8bub38lgv	Yura Perov
Existing AI solutions - Prediction MSK physiotherapy (August 2021)	https://docs.google.com/document/d/1odywCUsJT_gUVZ_AiSKopJaiA0Y0IMd-lB7WgMRyg48/edit#heading=h.xoz8bub38lgv	Kate Ryan
Ethical Considerations	https://docs.google.com/document/d/1jGArAAoIue6cOpxdnETT5Yrwo5Dx4hAzCHws--D0rfc/edit	Robert Pawinski

5.2.8 Status update for meeting L (meeting #11)

There are 7 members in the topic group.

Updates:

- There were 3 topic group meetings. The meeting notes can be found below:
  1. 2nd of February, 2021: https://docs.google.com/document/d/1Nup8ys5Uiz-uxQWhIGOcOimm1GhlLCWFi5bNinBkazU/edit
  2. 11th of March, 2021: https://docs.google.com/document/d/1j1d1BfNcGVu5Nx4Y41uuT4oE_hv5qyYlw9YhlpoG8BY/edit
  3. 15th of April, 2021: https://docs.google.com/document/d/1t868kUBmMQm4p94cfqc5D6fzmnfCUbhzXO4Jo1UtwP4/edit
- At the meeting on the 2nd of February 2021, there was a presentation "Example Application for Discussion: Fracture Risk Identification via Deep Learning" given by a topic group member Joseph LeMoine. There was a discussion following the talk. We discussed relevant AI tasks at the meetings on the 2nd of February and the 11th of March. Also, during the meetings (in particular, at the meeting on the 15th of April), we discussed ways to establish

partnerships for the topic group and attract more members and interested parties, as well as opportunities for organising relevant events and securing funding for the topic group work.
- Danielle Chulan contributed to the meeting on the 11th of March and Robert Pawinski contributed to the meeting on the 15th of April.
- There have been contacts with people from the industry in regard to the topic group (including healthcare providers, medical and technology companies and industry experts).
- Descriptions of two AI tasks for the applications of AI/ML for MSK medicine have been prepared and added to this document.
- A public shared folder has been created for the topic group: https://drive.google.com/drive/u/1/folders/1q7t_wJJzZnZdfOrRAZZnMVYVFztJZRq2
- Note: some edits were made to the section "Status update for meeting K (meeting #10)".
- Christopher Tack stopped being a member on his request on the 12th of May 2021.

5.2.9 Status update for meeting K (meeting #10)

At the focus group meeting J in September/October 2020, the topic group was approved and created. There were 8 members in the topic group.

There were 3 topic group meetings since the official creation of the topic group. The meeting notes can be found below:

- 17th of November 2020: https://docs.google.com/document/d/1Ni2lM83RattG9izL0ZlMTsQGKMVp2As708Si6TsqeSg/edit
- 18th of November 2020: https://docs.google.com/document/d/14qtY4ncduFyL4wTGZ410PrXL6TKwf2Le_R0vKYwhVTY/edit
- 17th of December 2020: https://docs.google.com/document/d/1iN4f5_Ai5N994FpmNhwQetYy6drFEfnRJcetmRV-cu8/edit

The meeting on the 17th of December 2020 included a talk by Dr Mark Elliott, a member of the topic group:

- Talk title: Sharing and integrating datasets for data driven research in osteoarthritis
- Given by Mark Elliott, Institute of Digital Healthcare, WMG, University of Warwick, UK; Theme Lead for Data Analysis, OATech Network+.
- Talk summary from Mark: "I will briefly introduce the OATech Network, an EPSRC funded network to investigate engineering and data driven approaches to osteoarthritis research. OA research is highly multidisciplinary, but research areas have remained siloed in their work and importantly, their data. To be able to apply data driven methods, such as machine learning, we need to combine datasets within and across disciplines to really benefit from these modern approaches. In this talk I will discuss some of the opportunities and challenges we found from a scoping study on data sharing. Finally, I'll discuss one of the projects we are leading in collaboration with the Alan Turing Institute, investigating the development of a large 3D motion capture dataset integrated from multiple smaller datasets (across institutions) for identifying biomechanical markers of OA progression."

Some of the outcomes from the meetings (including post-meeting analysis):

- MSK problems relate to bone, muscle and joint problems (by definition) but the subject is broader. It is not necessarily an injury but might be some dysfunction. Other elements are pain and mental health. Sometimes there is nothing to report in terms of injury but there is still pain, etc. There might be two parts that need to be benchmarked together: a section dealing with mechanical dysfunction and another section which is about psychosocial aspects.

- Benchmarks should contain objective and subjective measures.
- Benchmarks should include measures on how conditions/signs/symptoms affect a patient's life (and related improvement).
- Three areas for the topic group to focus on:
    - Self-management for and treatment of MSK conditions.
        - What is the assessment process that can be used to understand what intervention does a patient need for their MSK conditions? How to benchmark it?
    - Prediction and prevention of MSK conditions including risk identification and risk reduction (including new conditions, worsening or improvement of MSK condition states, etc.).
    - Motion capture, pose recognition, posture and gait analysis using computer vision and wearables based on video capture for analysis, treatment and prevention of MSK conditions.
        - Possible benchmarks:
            - Use of 3D motion capture to train and use ML/AI for pose/gait capture
            - Use of 2D cameras and align them with precise data.
            - Use of that data for gait analysis, movement deficiency detection, etc.
        - It was noted that one of the challenges is "getting out of the lab". The challenge for metrics (including for use in benchmarks) is measuring data outside of lab settings.
- Different benchmarks, or subtypes, might have to be developed for different conditions.
- Benchmark results should be stratified: by data from different agents (patients who are experiencing the symptoms and have conditions, whose life is or might be affected by MSK issues; and clinicians who are subject matter experts); different geographical regions; different MSK conditions.
- Next steps:
    - Start defining the applications in more details, then identifying metrics for benchmarks and weighting mechanisms for data points.
        - Also, identifying processes and guidelines for benchmarking.
    - Extend the reach of the group: find new members, collaborators and collect data. Letters to be drafted and sent to companies and research groups.
        - *TODO:* Write a letter to Google (FitBit), Apple, Samsung, London Marathon, other marathons, etc.
        - *TODO:* What other groups are doing/can be doing motion capture in general/for clinical applications. Contact them.
- A question: how to operate in the settings where interventions are "soft" and measures are "soft"? How to train and check AI in those settings? It is a challenge. How to formulate that problem from the AI/ML perspective for the benchmarking purpose? Can ideas from reinforcement learning and other similar machine learning approaches be used for benchmarking here?

5.2.10 Status of the topic group before meeting K (meeting #10)

The topic group proposal is document FG-AI4H-J-026-R01 which can be found here: https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/docs/Forms/200930.aspx. It also contains information about the preparatory meetings for this topic group work.

5.3 Topic Group participation

The participation in both, the Focus Group on AI for Health and in a TG is generally open to anyone (with a free ITU account). For this TG, the corresponding 'Call for TG participation' (CfTGP) can be found here:

– https://www.itu.int/en/ITU-T/focusgroups/ai4h/Documents/tg/CfP-TG-MSK.pdf

Each topic group also has a corresponding subpage on the ITU collaboration site. The subpage for this topic group can be found here:

– https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/tg/SitePages/TG-MSK.aspx

For participation in this topic group, interested parties can also join online meetings of the topic group. For all TGs, the default link will be the standard ITU-TG 'zoom' link:

– https://itu.zoom.us/my/fgai4h

unless a particular topic group meeting has its own invite link in the invite.

All relevant administrative information about FG-AI4H—like upcoming meetings or document deadlines—will be announced via the general FG-AI4H mailing list fgai4h@lists.itu.int.

All TG members should subscribe to this mailing list as part of the registration process for their ITU user account by following the instructions in the 'Call for Topic Group participation' and this link:

– https://itu.int/go/fgai4h/join

In addition to the general FG-AI4H mailing list, each topic group can create an *individual mailing list*. This topic group's mailing list is fgai4htgmsk@lists.itu.int. Instructions on how to register and subscribe to mailing lists are available here:

– https://www.itu.int/en/ITU-T/focusgroups/ai4h/Pages/reg2.aspx.

Regular FG-AI4H workshops and meetings proceed about every two months at changing locations around the globe or remotely. More information can be found on the official FG-AI4H website:

– https://itu.int/go/fgai4h

6 Topic description

This section contains a detailed description and background information of the specific health topic for the benchmarking of AI systems for Musculoskeletal Medicine and how this can help to solve a relevant 'real-world' problem.

Topic Groups summarise related benchmarking AI subjects to reduce redundancy, leverage synergies, and streamline FG-AI4H meetings. However, in some cases different subtopic groups can be established within one topic group to pursue different topic-specific fields of expertise. The TG-MSK currently has no subtopics. Future subtopics for different applications/approaches might be introduced.

6.1 AI/ML Prediction for MSK Health

6.1.1 Definition of and discussion regarding the AI task

Prediction models using multivariable regression are an integral part of medicine to estimate the probability of a diagnosis or a prognosis. AI/ML technology to assist in prediction of risks and/or outcomes is becoming a popular application in MSK Health. Prediction is one of the most developed uses of AI at present. Prediction has been proven useful in clinical practice and research in the healthcare domain. Scientific study traditionally observes and explains what is happening or what has happened and why. By leveraging statistics, computational power, and deep learning the AI models can extend this domain to what is likely to happen.

Identifying patients at risk for a condition, in the present or future, allows earlier preventative care strategies to eliminate or reduce the morbidity or mortality of disease. Using predictive models to supplement traditional modalities in health has the potential of offering prevention strategies to a greater population at less expense.

Using advanced prediction strategies in research allows focused study to target cohorts with predicted poorer outcomes for targeted improvement in management. For example, by predicting which subsets of individuals with open fractures are of greater risk of infection, research can identify this particular cohort and look at new innovations to improve outcomes for this subgroup.

These prediction models can also improve selection of treatment strategies for a given patient. Frequently the diagnosis of condition in MSK health is quite straightforward, the art and science is determining the optimal treatment for a given individual in a given situation. Many recent well-structured studies based on patient functional outcomes in orthopaedics have called into question some of the standard treatments for fractures in orthopaedics. The question remains: are there subsets of this cohort that could benefit from one treatment modality over another?

In order for predictive models to improve healthcare they need to be stable and validated. Reliability is essential when these tools are applied to patients. They must meet the same rigid standards of reporting, ethics, safety and reliability of other medical procedures, treatments and devices.

There are two major types of predictive models: regression with continuous output and classifications with binary or nominal output. If the algorithm produces a probability, usually a threshold is used to convert that to class outputs. Evaluation metrics can differ depending on the model.

Performance measures depend on the quality of data and labelling for training and testing. For example in using Computer Vision to determine and classify fractures, it will depend on the quality of imaging which can vary from one health unit to another and one technician to another. This "real world data" might not be available for the developer but should be for the benchmarking. Furthermore the classification labelling of fractures can vary depending on the classification system and the interpreter, such as generalist, specialist or speciality.

A prediction or classification model must have utility in research, or as a clinical tool. Before assessing a model by benchmarking its contribution to care must be identified. For example, identifying an incomplete or complete nondisplaced fracture of the femoral neck would at present not alter the treatment choice or the prognosis. The objective of the model's result can influence the choice of performance indicators. In the case of a clinical tool, once the performance indicators determine the possible utility of the model, the ultimate measure is done by controlled clinical trials measuring patient outcomes, with emphasis on improvement in function and quality of life.

For practical predictions the testing data should reflect "real world" data. The model should be able to accommodate missing data and extreme and possible erroneous outliers. A mechanism to identify and measure the success of the models handling of these situations would be desirable.

At present there is a set of recommendations for reporting of prediction models in medicine. The Transparent Reporting of a multivariable prediction model of Individual Prognosis Or Diagnosis (TRIPOD) was published in 2015 and includes a checklist of 22 items for reporting prediction models [4]. This is focused on regression analysis, but can apply to ML as well.

A new version of TRIPOD is in development for ML [5] and will address concerns regarding under and over prediction and overfitting of data and the need for robust validation of models using data that are of large scope and the developers do not have access to beforehand. It also aims to address comparing methods to simpler available models and the transparency of the model with a means to allow availability for independent evaluation and clinical implementation. This initiative [5] includes a call to participation from the AI/ML in the health community.

Performance of a model requires a benchmarking assessment. Many metrics can be used to report performance. AUROC, AUPRC, F1 score, accuracy are frequently used. Calibration is often cited as being underused, when predicting risk the reliability is important for safe practice by avoiding under and over treatment [6]. Calibration should be interpreted by the slope and the intercept together. Finally reporting should consider using terminology more familiar to clinicians. (For example the ROC is also termed the curve of true positive rate vs false positive rate or sensitivity vs 1-specificity and the PRC is Positive Predictive value versus the Sensitivity curve.)

Prediction modelling is likely to continue to be in the forefront of ML applications in musculoskeletal health in the near future. It has a proven utility in the past using regression models and this leads to a natural evolution towards deep learning models and other approaches such as generative modelling and causal inference for more complex models. These applications have the potential to contribute greatly to healthcare, but only if they are thoroughly tested and validated, first by benchmarking and then with clinical studies before safe adoption for use in healthcare.

6.1.2 Current gold standard - Musculoskeletal Health

Determining the present gold standard of prediction in healthcare is not always clearly defined, it is dependent on the question at hand and the development, performance and adoption of a solution.

Prediction is one of the cornerstones of clinical healthcare. It is a process of decision making based on probability with the goal of improving outcomes. Prediction in health is considered to be either diagnostic or prognostic. The famous physician educator William Osler once said "Medicine is a science of uncertainty and an art of probability". Traditionally a clinician considers the information available, from history taking, physical examination and imaging and laboratory studies to make a prediction. The decision is based on knowledge, training and intuition and the result is aptly named an opinion. Measurements of the specific history elements, physical and laboratory findings are assessed with sensitivity, specificity and accuracy metrics when used individually and when combined sequentially using Bayes' theorem and positive likelihood prediction. Unfortunately opinions are susceptible to heuristics which can lead to variation of prediction from patient to patient or between healthcare professionals.

In an era of information boom, the increase of data in both volume and complexity creates challenges for the physician to incorporate the information available into the decision process. The development of prediction modelling, which usually considers multiple variables and makes predictions based on logistic regression, improves clinical decisions by eliminating heuristics and stronger emphasis on evidence based medicine.

Predictive models require data for development, although prospectively collected data allows comprehensive data sets to specifically answer the question at hand they are often prohibitively time consuming and costly. More often retrospective data sets are used, these may be incomplete but it has been argued that they reflect more accurately the real world clinical scenario. The second element of a predictive model creation is selecting variables to use in prediction, traditionally these are variables with a pathophysiologic link between the data element and the outcome. Furthermore the number of variables is restricted to reduce costs and complexity of usage and avoid overfitting of the model.

There has been a surge in model development, often with multiple models addressing the same clinical question. This can reduce the adoption rate of model usage due to confusion. Other common reasons for non-adoption include excess complexity, lack of familiarity, transparency and utility, and finally clinical 'stubbornness', preferring personal judgement.

In the evolution of evidence based medicine, these prediction models are often incorporated into clinical decision processes such as pathways and guidelines. A given model's adoption into clinical decision algorithms and general adoption strengthens its argument for a gold standard.

For a particular clinical question, there can be more than one gold standard. A model that clearly outperforms another, when considering key metrics such as discrimination and calibration can be considered a statistical gold standard. However a model that has a higher adoption rate because of usability, practicality, transparency and incorporation into clinical decision algorithms could be considered a clinical gold standard.

In determining the present gold standard a given model must be assessed for its performance, accessibility, adoption and adherence to standards for reporting including methods, validation, metrics and bias as described in the TRIPOD and PROBAST statements.

References for this section: [7] and [8].

6.1.3 Existing AI solutions - Ortho

Artificial intelligence models in the domain of Orthopaedic Surgery are lagging compared to other health domains, in particular computer vision based modelling found in diagnostic imaging, dermatology and pathology. A recent review looked at 59 models found in search of the medical literature over the last 15 years applying to Orthopaedic Surgery. The vast majority of these studies (83%) have been published in the last five years and represent a trend of exponential growth.

Despite limited models in the area of surgery to the musculoskeletal system, a recent survey of surgeons indicates that they trust and are willing to use AI prediction models in clinical practice. However only just half would accept the model's prediction if it contradicted their present clinical judgement. And only 58% feel that AI prediction will have a significant role in decision making in the next five year. This indicates that there is enthusiasm for AI in Orthopaedic Surgery but full adoption can be limited. With greater emphasis in transparency of reporting, validation and benchmarking, and prospective clinical studies of the validated prediction tools, incorporation into practice can be improved.

Overall these studies have a focus on spine surgery, total joint replacement, hip fractures and tumours. (The list of publications is added in table form to the references). Popular models of prediction looked at domains of complications, patient reported outcome measures, health management, opioid consumption and mortality in the case of primary and metastatic tumours.

In this review only 3% used a prospectively collected data set. Furthermore only one half of the studies were registered in a national registry. Despite the publication of TRIPOD (Transparency in Reporting In Multivariate Prediction Model for Prognosis or Diagnosis) standards in 2014, only 20% of publications referred to this standard. Analysis with this tool showed only 53% median of completeness in TRIPOD reporting. Most notably, the model building procedure was grossly under reported limiting ability of external replication and validation.

In assessing the risk of bias in these studies using PROBAST (Prediction model Risk Of Bias ASsessment Tool) only 44% of the studies had a low risk of bias compared to 41% with high risk and a further 15% had insufficient information to determine the risk of bias.

Most of the studies limited outcome reporting focused on discrimination with reporting of the AURUC. Few studies used Precision-Recall Curve despite the frequent presence of unbalanced data. There was little reporting of Calibration,or Decision Curves Analysis (DCA) which are important before clinical adoption. DCA determines the net clinical benefit across a full spectrum of prediction thresholds weighing the benefits of true positives for some to the harm of false positives for others.

Overall the reported models were based on small sets of retrospective data. Retrospective databases reflect the real world situation but can have incomplete and missing data. Smaller sample sizes can lead to overfitting of data. In these situations additional methods are required to improve the models and should be reported and described.

A second published review looked at external validation of the previous group of models in Orthopaedic Surgery. Of the published studies only 10 models were externally validated. Some multiple times, for a total of 18 external validation publications. Most of these validation involved models concerning tumour survivorship and total joint replacement surgery health management parameters such as length of stay and discharge disposition. In this group 17 of the 18 involved at least one author from the original model publication. In these validations there is good retention of discrimination but again poor reporting of other performance metrics, with calibration reported in only 7 of the 18 studies.

These reviews indicate that there is an emerging trend in adopting AI modelling in prediction in Orthopaedic Surgery for complications, outcomes and health management metrics. There is a great need for validation and benchmarking. This requires greater transparency in reporting of methods of model building and management of dataset limitations, and bias risk to assess the models. Furthermore, a range of metrics are required including Discrimination based on dataset, Calibration and DCA before clinical adoption. AI based prediction tools have great potential in Orthopaedic Surgery, however, improved reporting and benchmarking are required for their clinical adoption.

References for this section: [9], [10] and [11].

6.1.4 Existing AI solutions - MSK Physiotherapy

AI systems for prediction in the field of MSK physiotherapy are still in their infancy with the majority being prototypes. However, there is increasing interest in the potential of systems to aid with prediction of exercise performance [18], recovery outcome [19] and even development of certain pathologies, such as osteoarthritis [16].

An overview of known AI systems and their inputs, outputs, key features, target user groups, and intended uses is provided in Table 4.

Table 4: Overview of known AI MSK systems and their features

Ref #	Intended Use	Target Population	Type of AI used	Input	Performance
Burns et al. [13]	Prediction of successful exercise performance	Healthy adults	CNN k-	Inertial smart watch sensor	99.4% prediction accuracy
Fidalgo-Herrera et al. [14]	Prediction of the effect of rehabilitation in whiplash associated disorder	Patients with WAD	ANN	Kinematics recorded by the EBI 5 inc. normalized aROM, speed to peak and ROM coefficient of variation	Moderate correlation R=0.5 Error too large for use in practice MSE 290, (95% CI 308.07–272.75)
Kianifar et al. [17]	Prediction of knee injury risk based on SLS movement quality	Healthy adults	10-FCV	Inertial measurement unit (IMU)	95% prediction accuracy
Tschuggnall et al. [20]	Predict Rehabilitation Success based on Clinical and Patient-Reported Outcome Measures	Patients with ankle, knee or hip MSK injuries	Random Forest	PROMs and CROMs including TUG, joint ROM,VAS HAQ and WOMAC	65% Prediction accuracy

Ref #	Intended Use	Target Population	Type of AI used	Input	Performance
Al-Yousef et al, [12]	Predicting treatment outcome of spinal MSK pain	Patients with spinal MSK pain	ANN	Pre treatment variables including VAS, Serum Vit D and ferritin	85% prediction accuracy
Huber et al. [15]	Prediction of patient-reported outcomes following hip and knee replacement surgery	Patients following hip and knee replacement surgery	Extreme gradient boosting	EQ-5D-3L,(VAS) Oxford Hip and Knee Score (Q score).	VAS 87% hip and 86% knee Q score 70%

The common feature between all the AI systems is accuracy. Either when compared to the test data sets or with clinical opinion. Accuracy scores for the AI models detailed above range widely from 65% [20] to 99.4%. Reasons for this large variance in accuracy could be due to quality of data, data set size used for building the AI model, type of AI used and the type of input e.g. from a wearable device vs patient reported outcome measures. Standardised reporting of accuracy with errors is an important parameter to optimise along with an agreement of minimum acceptable accuracy for AI systems developed for prediction in the field of MSK physiotherapy.

Current AI systems developed for prediction in the field of MSK physiotherapy fall into two main categories. Prediction of injury or movement performance in a healthy population, and prediction of rehabilitation outcomes in patients who already have MSK conditions. Interestingly, the two studies on models for prediction of injury [17] or movement performance [13] have significantly higher accuracy $\geq 95\%$ than those predicting rehabilitation outcome (65-87%). This could be coincidence but it is encouraging as the ability to identify, intervene and hopefully prevent at risk people from developing MSK conditions is an important part of reducing the vast disease burden of MSK conditions worldwide.

Prediction of rehabilitation outcomes is also very important aspect as it allows resources to be allocated more efficiently and could ensure that those with poorer rehab prognosis may be quickly identified and given additional support or alternative care

As yet none are robust enough to be on the market as medical devices or used on a widespread clinical basis. Limitations of the AI systems include insufficient training/ testing data sets, in the case of neural networks a 'black box' AI model which is not transparent and cannot easily be adjusted, clinician and patient acceptability. Nevertheless there is significant scope for development and for AI prediction models to have a significant impact on the global MSK disease burden.

References for this section: from [12] to [20].

6.1.5 Metrics (and related terms/notes)

General assumptions: there is a classifier that given some information about a patient, predicts some information that relates to the patient's state of having now (diagnostic) or in the future (prognostic) some outcome or condition.

Classifier prediction (output): a classifier's output is a value, usually a scalar (one for each outcome/condition). That output is not necessarily a probability.

Meaning of a classifier's output: usually, the higher a classifier's output value, the more likely (in terms of the classifier's prediction that depends on the classifier's accuracy; not necessarily in terms of a real situation) the chance that a patient has/would have some condition. For example, a classifier might return a value from 0 to 1, where 0 means that it is "impossible" to have a condition, and 1 means that it is (almost) guaranteed for a patient to have a condition (again, based on the classifier's belief which might be not exactly correct). There also might be different outputs:

e.g., a raw output (e.g. a continuous value from 0 to 1) and a "final" output (i.e. e.g. only 0 or 1, which is produced e.g. by applying a threshold that is learnt as part of the training process and/or with some other considerations and processes).

Classifier certainty/uncertainty: a classifier might return not just a value, but a range of values (e.g. a confidence interval), or some other estimate of its certainty.

Uncertain classifier: some classifier might flag (or estimate a chance of a situation) that that classifier is "quite" uncertain about diagnosis/prognosis for a particular patient, e.g. because a patient's case is quite different from all cases, on which the classifiers were trained. In this case, the output of that classifier (generally) should not be used.

Probability: a classifier's output is not necessarily a probability. Some classifiers even might output only e.g. 0-s and 1-s without any further differentiation. Some classifiers are modelled to return a probability. For other classifiers, some transformations might be performed (if possible at all) to transform, approximately, a classifier's output to a probability value.

Training/testing datasets: a classifier is trained on some data. Usually, available data is separated into (at least) training and testing data so that a classifier is trained on some portion of available data and then it is "independently" tested on another portion of available data.

Cross-validation: a cross-validation might be performed. For example, data might be separated into N folds, and then N experiments are performed. For each experiment, (N-1) folds are used for training and the remaining fold (which is out of N folds and which is different for each experiment) is used for testing. Then, the results (from the testing fold in each experiment) are aggregated and analysed.

"Positives" and "negatives" in a dataset: these are the data points (e.g. patients' cases) are marked (e.g. by an expert) as belonging to a "positive" class or not. For example, a "positive" class might mean patients who have/would have some specific MSK condition in 1 year.

True (false) positives (aka tp (fp)): data points that belong (in terms of the "ground truth") to a "positive" class and that are identified by a classifier correctly (incorrectly).

True (false) negatives (aka tn (fn)): data points that belong (in terms of the "ground truth") to a non-"positive" class and that are identified by a classifier correctly (incorrectly).

Recall / sensitivity / true positive rate: tp / (tp + fn).

Precision / positive predictive value: tp / (tp + fp).

Specificity / true negative rate: tn / (tn + fp).

Accuracy: it is can be defined as (tp + tn) / (tp + tn + fp + fn) (as in e.g. https://en.wikipedia.org/wiki/Accuracy_and_precision). Note that there might be different ways in general to describe/analyse "accuracy".

6.1.6 More information about Prognosis including some Case Studies

Defining prognosis
Within medical circles, prognosis has been defined as 'the prospect of recovering from injury or disease, or a prediction or forecast of the course and outcome of a medical condition.' [21]

Within the field of MSK medicine this definition can be subdivided into two main areas:

1. *Development prognosis* - i.e. if a patient has a given a set of risk factors how likely are they to develop condition X.
2. *Recovery prognosis* - if a patient has a current diagnosis of condition Y what is their most likely recovery pathway taking into account the specifics of their condition plus other relevant comorbidities and lifestyle factors.

Prognosis Dependencies

One of the main reasons for dividing prognosis into the two categories outlined above, is that AI models require different dependencies to be able to make predictions.

AI models dealing with *developmental prognosis* require general patient demographics such as age and long term risk factors which help predict condition development such as a history of trauma or surgery to particular body areas, use of certain medications [22], jobs or hobbies with repetitive movements or sustained postures [18].

For example, Bonakdari et al. [23] have developed *developmental prognosis* machine learning (ML) model that 'bridges major OA risk factors (age and bone mass index (BMI)) and serum levels of adipokines/related inflammatory factors at baseline for early prediction of at-risk knee OA patient structural progressors over time.'

Whereas AI models dealing with *recovery prognosis* require further information about the severity, duration and impact of the condition on the patient's life and more in depth information about psychosocial factors [24] such as mental health status, self efficacy, support networks and even whether or not the patient self-referred into physiotherapy [25].

A good example of an MSK progression prognosis model is the one developed by Tschuggnall et al [20] which uses PROMs and CROMs inc TUG, joint ROM, VAS HAQ and WOMAC to predict rehabilitation success in patients with ankle, knee or hip MSK injuries.

Development Prognosis Case Studies

Development of adhesive capsulitis following a mastectomy. Other risk factors for developing adhesive capsulitis include being aged 40-60 [26], having diabetes or a Body Mass Index (BMI) of over 30 [27]. Identification of patients with these additional risk factors could trigger additional shoulder rehab / care post surgery.

Development of De Quervain's Tendinopathy in patients with rheumatoid arthritis. Identification of patients at higher risk of De Quervain's Tendinopathy due to presence of multiple risk factors in the patient medical / social history such as those with hobbies/ jobs which involve repeated wrist flexion/extension, forearm rotation [28] or could allow early implementation of strategies to help avoid such as activity modification, prehab exercises etc.

Development of work related LBP. Professions with either very sedentary [29] requirements or those involving heavy lifting [30] are already at an increased risk of LBP. Giving employers a way to screen for employees most at risk by identifying co existing risk factors such as high BMI, smokers, those with a family history of LBP [29], low mood [30] or poor job satisfaction [31] could allow for early detection or preventative prehab where necessary.

Recovery Prognosis Case Studies

Prediction of patient-reported outcomes following hip and knee replacement surgery. Huber et al. [15] used an extreme gradient boosting to predict patient reported outcome measures of patients following hip and knee replacement surgery. The model inputs were EQ-5D-3L (VAS), Oxford Hip and Knee Score (Q score). After optimisation the model was able to predict hip and knee VAS, and Q score with accuracies of 87% 86% and 70% respectively.

Prediction of the effect of rehabilitation in whiplash associated disorder. Fidalgo-Herrera et al. [14] used an artificial neural network (ANN) to predict rehabilitation success in patients with WAD. The inputs were kinematics recorded by the EBI® 5 inc. normalized aROM, speed to peak and ROM coefficient of variation. The model was able to achieve a medium correlation (R=0.5) when predicting Neck Functional Holistic Analysis Scores (NFHAS).

Predicting treatment outcome of spinal MSK pain. Al-Yousef et al [12] also used an ANN to predict treatment outcome of patients with MSK spinal pain from pre-treatment variables inc VAS, Serum

Vit D and ferritin. Post-treatment endpoint follow-up (fourth week VAS) was selected as a good indicator of treatment outcome.

6.2 Self-Management/Management/Treatment of MSK medicine/Physiotherapy conditions

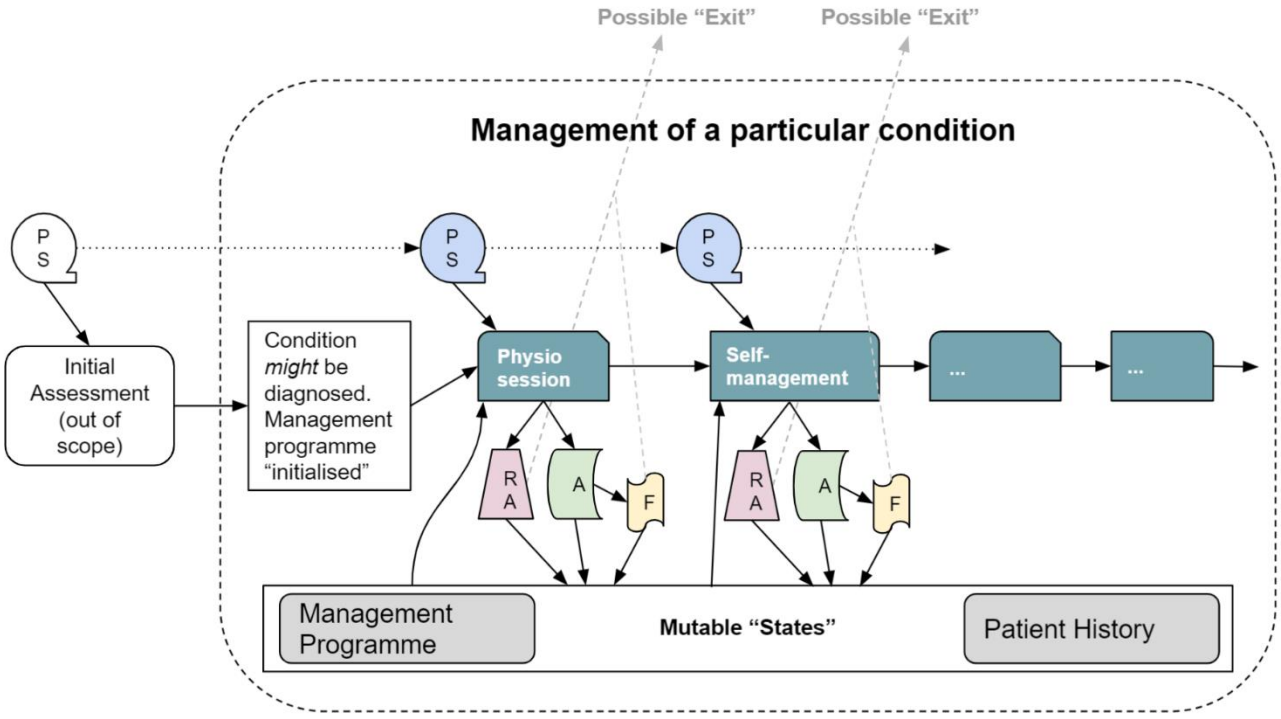


Figure 2: Illustration for AI sub-task "Self-Management/Management/Treatment of MSK medicine/Physiotherapy conditions"

This subtask relates to management/treatment of MSK conditions. The settings for this subtask are after an "initial" assessment (triage and/or some form of diagnostics) was already performed, and there has been prescribed a particular management/treatment programme/plan.

There are several points of management, including potentially sessions with a clinician as well as self-management sessions.

Each "session" depends on a particular patient state (PS) at that moment. Each "session" also depends on a version of the management programme up to date and on all preserved patient history (both of those "variables"/"states" are mutable since they are being updated inside/after each "session"). Before each session can start, there is a formal/informal "reassessment" (RA): e.g. if a patient feels not well or if a patient's health has deteriorated, that particular management programme can't be continued. If so, there happens an "exit" from the management programme; options for that "exit" include:

– Further, more detailed, reassessment of a patient's health (e.g. a new, more detailed, triage; (further) diagnostics; tests; etc.).

– A special intervention (e.g. an emergency healthcare call).

(Note that some patients will be "readmitted" back to the same management programme, with an updated patient history.)

If a reassessment has not identified any concerns that would require such an "exit", then a patient is advised to perform some "actions" (A) at that moment (e.g. exercises). As he/she performs them, objective and subjective feedback (F) is being accumulated, e.g.:

– How well can he/she perform them? (Including subjective (general assessment) and objective (e.g. angles) measures.)

– How does he/she feel whilst performing them and straight after that?

– What exercises can't be fully/partially performed?

(Note that an "action" (A) might be "empty" for some sessions, or, for some other sessions, it can just serve a purpose of collecting a patient's health state (i.e. no significant exercise for some sessions).)

(Note that the feedback can be accumulated by both the patient and by their clinician, if applicable.)

What parts of that process can be performed by AI/ML algorithms and can be measured:

– Predicting what exercises (A) are suggested by a clinician for a particular session, given a patient's health state and all other data up to this point.

– Predicting a patient's feedback (F) for a session.

– Predicting a need for an "exit" for a session.

7 Ethical considerations

AI considered in the following context:

(1) diagnosis, (2) patient morbidity or mortality risk assessment, (3) disease outbreak prediction and surveillance, and (4) health policy and planning.

Overall AI consideration includes:

(1) static (e.g. machine learning) (2) AI continuous release cycle and (3) AI and continuous-learning

Patient Considerations

– Differential access to health care, especially in resource constrained settings with low HCP: population ratio (health inequalities driven by lack of access to technology)

– AI may be tested on sub-populations and not validated / biased towards other populations

– Increased access to specialist solutions in previously underserved areas

– Potential to disrupt the patient / physician relationship

– Cybersecurity, GDPR

– Cost

– Informed consent and shared decision making (patient / HCP)

Health Care Professional Considerations

– Potential to disrupt the patient /physician relationship

– Accountability and responsibility of decision making

– AI may be tested on sub-populations and not validated / biased towards other populations

– Lack of buy in and commercial considerations / conflicts of interest (e.g. loss of patient cohorts)

– Informed consent and shared decision making (patient / HCP)

Public Health & Health System Owner / Health Care Provider Considerations

– Untested tools being implemented in an accelerated manner without appropriate validation

– Including in terms of reducing jobs (if some jobs are partially/fully replaced by AI)

– Potential Bias against some sub-populations

– Potential to increase health inequalities
– Potential lack of ethical review committee reviews of protocols used to develop or validate tools
– Accountability of decision making
– Cross border considerations
– Cost effective and cost benefit considerations, and potential conflict with financial incentives (e.g. compared to standard of care)

Developer / Owner

– Continuous positive benefit / risk over the life of the AI tool
– Cybersecurity, GDPR considerations
– Management of all aspects of bias

Conclusion

Global standards and guidelines are needed to inform the development and evaluate performance of AI tools in health settings. Potential ethical concerns require careful consideration in these settings. Patient advisors must be engaged at an early stage to ensure ethical considerations are at the forefront of AI tool development.

8 Existing work on benchmarking

This section focuses on the existing benchmarking processes in the context of AI systems and Musculoskeletal Medicine for quality assessment. It addresses different aspects of the existing work on benchmarking of AI systems (e.g., relevant scientific publications, benchmarking frameworks, scores and metrics, and clinical evaluation attempts). The goal is to collect all relevant learnings from previous benchmarking that could help to implement the benchmarking process in this topic group.

8.1 Publications on benchmarking systems

While a representative, comprehensive comparable benchmarking for AI systems for MSK Medicine does not yet exist, to the best of our knowledge, some work has been done in the scientific community assessing the performance of such systems. This section summarizes insights from the most relevant publications on this topic. It covers parts of the deliverable DEL7 *"AI for health evaluation considerations,"* DEL7.1 *"AI4H evaluation process description,"* DEL7.2 *"AI technical test specification,"* DEL7.3 *"Data and artificial intelligence assessment methods (DAISAM),"* and DEL7.4 *"Clinical Evaluation of AI for health"*.

8.2 Benchmarking by AI developers

All developers of AI solutions for MSK Medicine implemented internal benchmarking systems for assessing the performance. This section will outline the insights and learnings from this work of relevance for benchmarking in this topic group.

8.3 Relevant existing benchmarking frameworks

Triggered by the hype around AI, recent years have seen the development of a variety of benchmarking platforms where AIs can compete for the best performance on a determined dataset. Given the high complexity of implementing a new benchmarking platform, the preferred solution is to use an established one. This section reflects on the different existing options that are relevant for this topic group and includes considerations of using the assessment platform that is currently developed by FG-AI4H and presented by deliverable DEL7.5 *"FG-AI4H assessment platform"* (the

deliverable explores options for implementing an assessment platform that can be used to evaluate AI for health for the different topic groups).

9 Benchmarking by the topic group

This section describes technical and operational details regarding the benchmarking process for the MSK Medicine AI tasks including subsections for each version of the benchmarking that is iteratively improved over time.

It reflects the considerations of various deliverables: DEL5 *"Data specification"* (introduction to deliverables 5.1-5.6), DEL5.1*"Data requirements"* (which lists acceptance criteria for data submitted to FG-AI4H and states the governing principles and rules), DEL5.2 *"Data acquisition"*, DEL5.3 *"Data annotation specification"*, DEL5.4 *"Training and test data specification"* (which provides a systematic way of preparing technical requirement specifications for datasets used in training and testing of AI models), DEL5.5 *"Data handling"* (which outlines how data will be handled once they are accepted), DEL5.6 *"Data sharing practices"* (which provides an overview of the existing best practices for sharing health-related data based on distributed and federated environments, including the requirement to enable secure data sharing and addressing issues of data governance), DEL06 *"AI training best practices specification"* (which reviews best practices for proper AI model training and guidelines for model reporting), DEL7*"AI for health evaluation considerations"* (which discusses the validation and evaluation of AI for health models, and considers requirements for a benchmarking platform), DEL7.1 *"AI4H evaluation process description"* (which provides an overview of the state of the art of AI evaluation principles and methods and serves as an initiator for the evaluation process of AI for health), DEL7.2 *"AI technical test specification"* (which specifies how an AI can and should be tested *in silico*), DEL7.3 *"Data and artificial intelligence assessment methods (DAISAM)"* (which provides the reference collection of WG-DAISAM on assessment methods of data and AI quality evaluation), DEL7.4*"Clinical Evaluation of AI for health"* (which outlines the current best practices and outstanding issues related to clinical evaluation of AI models for health), DEL7.5 *"FG-AI4H assessment platform"* (which explores assessment platform options that can be used to evaluate AI for health for the different topic groups), DEL9 *"AI for health applications and platforms"* (which introduces specific considerations of the benchmarking of mobile- and cloud-based AI applications in health), DEL9.1 *"Mobile based AI applications,"* and DEL9.2 *"Cloud-based AI applications"* (which describe specific requirements for the development, testing and benchmarking of mobile- and cloud-based AI applications).

The benchmarking of MSK Medicine is going to be developed and improved continuously to reflect new features of AI systems or changed requirements for benchmarking. This section outlines all benchmarking versions that have been implemented thus far and the rationale behind them. It serves as an introduction to the subsequent sections, where the actual benchmarking methodology for each version will be described.

10 Regulatory considerations

For AI-based technologies in healthcare, regulation is not only crucial to ensure the safety of patients and users, but also to accomplish market acceptance of these devices. This is challenging because there is a lack of universally accepted regulatory policies and guidelines for AI-based medical devices. To ensure that the benchmarking procedures and validation principles of FG-AI4H are secure and relevant for regulators and other stakeholders, the working group on *"Regulatory considerations on AI for health" (WG-RC)* compiled the requirements that consider these challenges.

The deliverables with relevance for regulatory considerations are DEL2 *"AI4H regulatory considerations"* (which provides an educational overview of some key regulatory considerations), DEL2.1 *"Mapping of IMDRF essential principles to AI for health software"*, and DEL2.2

"Guidelines for AI based medical device (AI-MD): Regulatory requirements" (which provides a checklist to understand expectations of regulators, promotes step-by-step implementation of safety and effectiveness of AI-based medical devices, and compensates for the lack of a harmonized standard). DEL4 identifies standards and best practices that are relevant for the "*AI software lifecycle specification.*" The following sections discuss how the different regulatory aspects relate to the TG-MSK.

10.1 Existing applicable regulatory frameworks

Most of the AI systems that are part of the FG-AI4H benchmarking process can be classified as *software as medical device* (SaMD) and eligible for a multitude of regulatory frameworks that are already in place. In addition, these AI systems often process sensitive personal health information that is controlled by another set of regulatory frameworks. The following section summarizes the most important aspects that AI manufacturers need to address if they are developing AI systems for MSK Medicine.

10.2 Regulatory features to be reported by benchmarking participants

In most countries, benchmarked AI solutions can only be used legally if they comply with the respective regulatory frameworks for the application context. This section outlines the compliance features and certifications that the benchmarking participants need to provide as part of the metadata. It facilitates a screening of the AI benchmarking results for special requirements (e.g., the prediction of prediabetes in a certain subpopulation in a country compliant to the particular regional regulatory requirements).

10.3 Regulatory requirements for the benchmarking systems

The benchmarking system itself needs to comply with regulatory frameworks (e.g., some regulatory frameworks explicitly require that all tools in the quality management are also implemented with a quality management system in place). This section outlines the regulatory requirements for software used for benchmarking in this topic group.

10.4 Regulatory approach for the topic group

Building on the outlined regulatory requirements, this section describes how the topic group plans to address the relevant points in order to be compliant. The discussion here focuses on the guidance and best practice provided by the DEL2 *"AI4H regulatory considerations."*

References

[1] "Musculoskeletal conditions" on WHO website. https://www.who.int/news-room/fact-sheets/detail/musculoskeletal-conditions. Accessed on the 24th of June 2020.

[2] "Musculoskeletal" page on NHS England website. https://www.england.nhs.uk/elective-care-transformation/best-practice-solutions/musculoskeletal/. Accessed on the 24th of June 2020.

[3] "Health workforce requirements for universal health coverage and the Sustainable Development Goals", Human Resources for Health Observer, Issue No. 17. https://www.who.int/hrh/resources/health-observer17/en/

[4] "TRIPOD Checklist: Prediction Model Development". https://www.tripod-statement.org/wp-content/uploads/2020/01/Tripod-Checlist-Prediction-Model-Development.pdf. Accessed on the 10th of May 2021.

[5] "Reporting of artificial intelligence prediction models", by Collins, G.S. and Moons, K.G., 2019. https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(19)30037-6/fulltext. Accessed on the 10th of May 2021.

[6] "Calibration: the Achilles heel of predictive analytics", Van Calster, B., McLernon, D.J., van Smeden, M., Wynants, L. and Steyerberg, E.W. (on behalf of Topic Group 'Evaluating diagnostic tests and prediction models' of the STRATOS initiative); BMC Medicine, 17, 230, 2019. https://bmcmedicine.biomedcentral.com/articles/10.1186/s12916-019-1466-7. Accessed on the 10th of May 2021.

[7] "Developing prediction models for clinical use using logistic regression: an overview", Shipe, M. E., Deppen, S. A., Farjah, F., & Grogan, E. L., 2019. Journal of thoracic disease, 11 (Suppl 4), S574.

[8] "Clinical prediction rules in practice: review of clinical guidelines and survey of GPs", Plüddemann, A., Wallace, E., Bankhead, C., Keogh, C., Van der Windt, D., Lasserson, D., Galvin, R., Moschetti, I., Kearley, K., O'Brien, K. and Sanders, S., 2014. British Journal of General Practice, 64 (621), e233-e242.

[9] "Machine learning prediction models in orthopedic surgery: A systematic review in transparent reporting", Groot, O.Q., Ogink, P.T., Lans, A., Twining, P.K., Kapoor, N.D., DiGiovanni, W., Bindels, B.J., Bongers, M.E., Oosterhoff, J.H., Karhade, A.V. and Oner, F.C., 2021. Journal of Orthopaedic Research (2021).

[10] "Availability and reporting quality of external validations of machine-learning prediction models with orthopedic surgical outcomes: a systematic review", Groot, O.Q., Bindels, B.J., Ogink, P.T., Kapoor, N.D., Twining, P.K., Collins, A.K., Bongers, M.E., Lans, A., Oosterhoff, J.H., Karhade, A.V. and Verlaan, J.J., 2021. Acta Orthopaedica (2021): 1-9.

[11] "Machine Learning Driven Tools in Orthopaedics and Spine Surgery: Hype or Reality? Applications and Perception of 31 Physician Opinions", Lans, A., Oosterhoff, J.H., Groot, O.Q. and Fourman, M.S., 2021. Seminars in Spine Surgery.

[12] "Predicting treatment outcome of spinal musculoskeletal pain using artificial neural networks: a pilot study", Al-Yousef, A., Eloqayli, H., Obiedat, M. and Almoustafa, A., 2021. International Journal of Medical Engineering and Informatics, 13(3), pp.237-253.

[13] "Shoulder physiotherapy exercise recognition: machine learning the inertial signals from a smartwatch", Burns, D.M., Leung, N., Hardisty, M., Whyne, C.M., Henry, P. and McLachlin, S., 2018. Physiological measurement, 39(7), p.075007.

[14] "Artificial intelligence prediction of the effect of rehabilitation in whiplash associated disorder", Fidalgo-Herrera, A.J., Martínez-Beltrán, M.J., de la Torre-Montero, J.C., Moreno-Ruiz, J.A. and Barton, G., 2020. Plos one, 15(12), p.e0243816.

[15] "Predicting patient-reported outcomes following hip and knee replacement surgery using supervised machine learning", Huber, M., Kurz, C. and Leidl, R., 2019. BMC medical informatics and decision making, 19(1), pp.1-13.

[16] "Machine-learning-based patient-specific prediction models for knee osteoarthritis", Jamshidi, A., Pelletier, J.P. and Martel-Pelletier, J., 2019. Nature Reviews Rheumatology, 15(1), pp.49-60.

[17] "Automated assessment of dynamic knee valgus and risk of knee injury during the single leg squat", Kianifar, R., Lee, A., Raina, S. and Kulić, D., 2017. IEEE journal of translational engineering in health and medicine, 5, pp.1-13.

[18] "Artificial intelligence and machine learning applications in musculoskeletal physiotherapy", Tack, C., 2019. Musculoskeletal Science and Practice, 39, pp.164-169.

[19] "Artificial intelligence to improve back pain outcomes and lessons learnt from clinical classification approaches: three systematic reviews", Tagliaferri, S.D., Angelova, M., Zhao, X., Owen, P.J., Miller, C.T., Wilkin, T. and Belavy, D.L., 2020. NPJ digital medicine, 3(1), pp.1-16.

[20] "Machine Learning Approaches to Predict Rehabilitation Success based on Clinical and Patient-Reported Outcome Measures", Tschuggnall, M., Grote, V., Pirchl, M., Holzner, B., Rumpold, G. and Fischer, M.J., 2021. Informatics in Medicine Unlocked, p.100598.

[21] "How to use an article about prognosis", Hansebout, R.R., Cornacchi, S.D., Haines, T. and Goldsmith, C.H., 2009. Canadian Journal of Surgery, 52(4), p.328.

[22] "Association between tendon ruptures and use of fluoroquinolone, and other oral antibiotics: a 10-year retrospective study of 1 million US senior Medicare beneficiaries", Baik, S., Lau, J., Huser, V. and McDonald, C.J., 2020. BMJ open, 10(12), p.e034844

[23] "A warning machine learning algorithm for early knee osteoarthritis structural progressor patient screening", Bonakdari, H., Jamshidi, A., Pelletier, J.P., Abram, F., Tardif, G. and Martel-Pelletier, J., 2021. Therapeutic Advances in Musculoskeletal Disease, 13, p.1759720X21993254.

[24] "Do therapist effects determine outcome in patients with shoulder pain in a primary care physiotherapy setting?", Kooijman, M.K., Buining, E.M., Swinkels, I.C., Koes, B.W. and Veenhof, C., 2020. Physiotherapy, 107, pp.111-117.

[25] "Characteristics of patients with knee and ankle symptoms accessing physiotherapy: self-referral vs general practitioner's referral", Lankhorst, N.E., Barten, J.A., Meerhof, R., Bierma-Zeinstra, S.M.A. and van Middelkoop, M., 2020. Physiotherapy, 108, pp.112-119.

[26] "A profile of patients with adhesive capsulitis", Boyle-Walker, K.L., Gabard, D.L., Bietsch, E., Masek-VanArsdale, D.M. and Robinson, B.L., 1997. Journal of hand therapy, 10(3), pp.222-228.

[27] "Shoulder adhesive capsulitis: epidemiology and predictors of surgery", Kingston, K., Curry, E.J., Galvin, J.W. and Li, X., 2018. Journal of shoulder and elbow surgery, 27(8), pp.1437-1443.

[28] "Risk factors for de Quervain's disease in a French working population", Le Manac'h, A.P., Roquelaure, Y., Ha, C., Bodin, J., Meyer, G., Bigot, F., Veaudor, M., Descatha, A., Goldberg, M. and Imbernon, E., 2011. Scandinavian journal of work, environment & health, pp.394-401.

[29] "Non-specific low back pain", Balagué, F., Mannion, A.F., Pellisé, F. and Cedraschi, C., 2012. The Lancet, 379(9814), pp.482-491.

[30] "Non-specific low back pain", Maher, C., Underwood, M. and Buchbinder, R., 2017. The Lancet, 389(10070), pp.736-747.

[31] "Risk and prognostic factors for non-specific musculoskeletal pain: a synthesis of evidence from systematic reviews classified into ICF dimensions", Lakke, S.E., Soer, R., Takken, T. and Reneman, M.F., 2009. PAIN®, 147(1-3), pp.153-164.

Annex A:
Glossary

This section lists all the relevant abbreviations, acronyms and uncommon terms used in the document.

Acronym/Term	Expansion	Comment
AI	Artificial intelligence	
AI4H	Artificial intelligence for health	
AI-MD	AI based medical device	
API	Application programming interface	
CfTGP	Call for topic group participation	
DEL	Deliverable	
FDA	Food and Drug administration	
FGAI4H	Focus Group on AI for Health	
GDP	Gross domestic product	
GDPR	General Data Protection Regulation	
IMDRF	International Medical Device Regulators Forum	
IP	Intellectual property	
ISO	International Standardization Organization	
ITU	International Telecommunication Union	
LMIC	Low-and middle-income countries	
MDR	Medical Device Regulation	
PII	Personal identifiable information	
SaMD	Software as a medical device	
TDD	Topic Description Document	Document specifying the standardized benchmarking for a topic on which the FG AI4H Topic Group works. This document is the TDD for the Topic Group MSK
TG	Topic Group	
WG	Working Group	
WHO	World Health Organization	

Annex B:
Information about members (including ex-members) & Declaration of conflict of interests

Important: the information in this section is (or may be) generally provided as it is provided by the members of the topic group (potentially, with some exceptions, e.g., in regard to formatting). The information is not necessarily checked, verified, etc.

In alphabetical order:

Danielle Chulan, Connect Health, UK

Within my current role with Connect Health, I chair our internal and external digital MSK framework meetings, including the partnership meetings with EQL. My special interests are digital innovation and big data analysis to inform clinical and operational development. I have a wide network across the Country as part of my National role that I think can add value from a UK perspective to this topic group. I am also an MSK clinician by background and believe that this provides invaluable insights into evidence based clinical care and patient diversity.

Nick Downing, Vita Health Group, UK

The NHS Head of Transformation for the Vita Health Group. As transformation lead I have both a personal and business interest in AI technologies for health and implementing these into our MSK and MH business. I understand healthcare systems and both the drive and potential benefit of digital in transforming healthcare.

Mark Elliott, University of Warwick, UK

I am currently Associate Professor at the Institute of Digital Healthcare, WMG, University of Warwick, UK. My research interests focus on measuring health, wellbeing and behaviour through data-driven approaches, often working in partnership with commercial, public health and NHS organisations. I lead the WMG Motion Capture Lab at Warwick and my work currently focusses on analysing and modelling data from wearable and mobile devices for orthopaedic applications and also on physical activity behaviour change using smartphone data. I am also the theme lead for data analysis on the OATech EPSRC Network+ for osteoarthritis research.

Peter Grinbergs, EQL, UK

A Co-founder and the Chief Medical Officer at EQL. Before EQL, he founded two medical companies (including a nationwide physiotherapy chain) and was CMO for a large medical reporting agency. Peter is a Member of the Chartered Society of Physiotherapy, where he sits on the Digital and Informatics Physiotherapy Group. He is also on the Health and Care Professions Council. Under his direction, his company, Physio 1st, grew from a single site to a team of over 50 people across 35 locations in 20 major cities, delivering in excess of 50,000 physiotherapy treatments a year. Earlier in his career, Peter was Birmingham City FC team's physiotherapist for two years (a season in the Championship, followed by a season in the Premier League). EQL is a digital health-tech organisation based in London, UK, which focuses on MSK conditions and physiotherapy. EQL's product, Phio Access, provides a conversational AI-enabled digital solution to support triage for MSK conditions. EQL is currently working on its next-generation products, with the extended application of AI and ML techniques for MSK medicine and physiotherapy.

Michael Guard, EQL, UK

I am Michael Guard, a clinical specialist chartered MSK Physiotherapist with 10 years postgraduate experience across multiple sectors, working currently as the head of clinical services at EQL, UK (alongside Peter Grinbergs). I have a keen interest in data-science, digital transformation and digital healthcare. I am currently (Cohort 3) a Topol digital fellow [link: https://topol.hee.nhs.uk/digital-fellowships/fellows/michael-guard/] and have led successful service-level (an example - [link: https://www.physiotherapyjournal.com/article/S0031-9406(21)00638-6/fulltext]) clinical data projects within the UK NHS. I am keen to learn and contribute to progressing the understanding of best MSK management, at a population level.

Joseph LeMoine, prIME Assessments, Canada

I am an orthopaedic surgeon. Involved as director of prIME Assessments. Interest in using NLP, CV, ML, OCR, AI/ML in structuring and extracting data from medical charts for diagnosis and treatment validation with correlation with outcomes. I am not a professional or trained data scientist, but am a strong supporter of the discipline and my interest is applications for data structure and extraction and for predictive analysis in the MSK medicine domain. (applicable in the private insurance sector, in practice auditing and metaanalysis based academic research). Expertise in orthopaedics (medical and surgical) with keen interest in diagnosis criteria, treatment guidelines and metaanalysis of outcomes and incorporating AI algorithms into these subjects. Standardized guidelines backed by a benchmarking data set is a great step forward in developing and introducing the technology to practical applications.

Emma Meehan

[No info provided.]

Yura Perov, Individual contributor, UK

Yura is a Chartered Scientist, Chartered Mathematician, Member of the Institute of Mathematics and its Applications, and Professional Member of the British Computer Society. He studied and carried out research in Computer Science, AI and Mathematics at the University of Oxford, MIT, EPFL and Siberian Federal University. Yura was previously a senior research scientist at Babylon Health, co-leading the development of the AI-triage/diagnostics product for primary care which was utilised by Babylon, Samsung and Prudential worldwide. He later was Head of AI and Data Science at EQL. Yura is now a Principal Research Scientist at Babylon Health. Yura has been a member of the Symptom Assessment topic group of the ITU/WHO focus group AI for Health.

Kate Ryan, EQL, UK

MSK Data Science Clinical Expert at EQL. Kate is a chemistry academic, turned MSK physiotherapist. She studied and conducted research at the University of Southampton, the University of Oxford, Argonne National Laboratory and King's College London. Over the course of her doctoral and postdoctoral work, Kate has co-authored numerous highly-cited research papers and several successful grant proposals.

Christopher Tack, NHS, UK

I am a clinical specialist musculoskeletal physiotherapist by background. I am also one of the inaugural Topol Digital Health Fellows at Health Education England, the digital lead for AHPs at my host organisation (GSTT), and co-chair of the London AHP Informatics and Digital Network.

(Christopher Tack stopped being a member on his request on the 12th of May 2021.)

Olalekan Uthman, University of Warwick, UK

Prof Ola Uthman is a seasoned clinical epidemiologist, currently employed as a Professor in Global Health Informatics at Warwick Centre for Applied Research and Delivery, University of Warwick, where I am primarily involved developing and help managing a portfolio of research relevant to Global Health Informatics for Improving Quality of Healthcare including: (1) Application of innovative machine learning algorithms for identifying the opportunities for prevention and treatment of diseases; (2) natural language processing big data for public health surveillance; (3) mobile health and clinical decision support system; and (4) using natural experiments to evaluate population health interventions and translating evidence into practice, implementation research science and evaluating health service effectiveness. He is proficient in mathematical modelling and focuses on the use of mathematical models to understand the epidemiology and control of diseases of public health importance and utilize epidemiologic and surveillance data to assess the impact of interventions and to set programmatic priorities. In addition, to advanced evidence synthesis such as network meta-analysis; and he is proficient in modern machine learning algorithms, including directly applying the advancements in NLP to biomedical text mining, BioBERT (Bidirectional Encoder Representations from Transformers for Biomedical Text Mining). Prof Ola is working on clinical AI technology to analyse clinically curated, anonymised patient data to solve serious unmet medical needs across a wide range of therapeutic areas, enabling a new approach to clinical trial design and post-marketing surveillance.