|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | INTERNATIONAL TELECOMMUNICATION UNION  **TELECOMMUNICATION STANDARDIZATION SECTOR**  STUDY PERIOD 2022-2024 | | FG-AI4H-Q-026-A01 | |
| **ITU-T Focus Group on AI for Health** | |
| **Original: English** | |
| **WG(s):** | | Plenary | Douala, 6-9 December 2022 | |
| **DOCUMENT** | | | | |
| **Source:** | | TG-MSK Topic Drivers | | |
| **Title:** | | Att.1 – TDD update (TG-MSK) | | |
| **Purpose:** | | Discussion | | |
| **Contact:** | | Peter Grinbergs EQL, UK | | E-mail: [tgmskorg@googlegroups.com](mailto:tgmskorg@googlegroups.com) |
| **Contact:** | | Yura Perov Independent contributor, UK | |

|  |  |
| --- | --- |
| **Abstract:** | 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 (and follows the template structure defined in document FGAI4H-J-105). The creation of this TDD is an ongoing iterative process until it is approved by the Focus Group on AI for Health (FG-AI4H) as deliverable DEL10.21. This draft will be a continuous input- and output document. |

|  |  |
| --- | --- |
| **Change notes:** | Version 7 (being submitted for meeting Q)   * It is the seventh version of the document.   Version 6 (was submitted for meeting P)   * It is the sixth version of the document.   Version 5 (was submitted for meeting O)   * It is the fifth version of the document.   Version 4 (submitted as FGAI4H-N-026-A01 to E-meeting N)   * It is the fourth version of the document.   Version 3 (submitted as FGAI4H-M-026-A01 to E-meeting M)   * It is the third version of the document.   Version 2 (submitted as FGAI4H-L-026-A01 to E-meeting L)   * It is the second version of the document.   Version 1 (submitted as FGAI4H-K-026-A01 to E-meeting K)   * It is the first version of the document. |

Contributors (topic group members, including ex-members) (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 |  |
| Emma Meehan [More info may be provided in the next version] |  |
| Yura Perov  Individual Contributor UK |  |
| Kate Ryan  EQL UK | E-mail: kate@eql.ai |
| Azadur Rahman Sarker [More info may be provided in the next version] |  |
| Ashwini Sathnur [More info may be provided in the next version] |  |
| 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 |
| --- |
| [1 Important General Note 7](#_Toc120562672)  [2 Introduction 7](#_Toc120562673)  [3 Relevance of the topic group 7](#_Toc120562674)  [4 Impact 8](#_Toc120562675)  [5 About the FG-AI4H topic group on MSK Medicine 8](#_Toc120562676)  [5.1 Documentation 9](#_Toc120562677)  [5.2 Status of this topic group 9](#_Toc120562678)  [5.2.1 Status update for meeting Q 9](#_Toc120562679)  [5.2.2 Status update for meeting P 10](#_Toc120562680)  [5.2.3 Status update for meeting O (meeting #14) 11](#_Toc120562681)  [5.2.4 Status update for meeting N (meeting #13) 11](#_Toc120562682)  [5.2.5 Status update for meeting M (meeting #12) 12](#_Toc120562683)  [5.2.6 Status update for meeting L (meeting #11) 13](#_Toc120562684)  [5.2.7 Status update for meeting K (meeting #10) 14](#_Toc120562685)  [5.2.8 Status of the topic group before meeting K (meeting #10) 16](#_Toc120562686)  [5.2.9 Status update for meeting [MEETING LETTER] 16](#_Toc120562687)  [5.2.10 Status update for meeting [MEETING LETTER] 16](#_Toc120562688)  [5.3 Topic group participation 16](#_Toc120562689)  [6 Topic description 17](#_Toc120562690)  [6.1 AI/ML Prediction for MSK Health 17](#_Toc120562691)  [6.1.1 Definition of and discussion regarding the AI task 17](#_Toc120562692)  [6.1.2 Current gold standard - Musculoskeletal Health 18](#_Toc120562693)  [6.1.3 Existing AI solutions - Ortho 19](#_Toc120562694)  [6.1.4 Existing AI solutions - MSK Physiotherapy 20](#_Toc120562695)  [6.1.5 Metrics (and related terms/notes) 22](#_Toc120562696)  [6.1.6 More information about Prognosis including some Case Studies 23](#_Toc120562697)  [6.2 Self-Management/Management/Treatment of MSK medicine/Physiotherapy conditions 25](#_Toc120562698)  [6.3 Subtopic [A] 26](#_Toc120562699)  [6.3.1 Definition of the AI task 26](#_Toc120562700)  [6.3.2 Current gold standard 26](#_Toc120562701)  [6.3.3 Relevance and impact of an AI solution 26](#_Toc120562702)  [6.3.4 Existing AI solutions 26](#_Toc120562703)  [6.4 Subtopic [B] 27](#_Toc120562704)  [7 Ethical considerations 27](#_Toc120562705)  [8 Existing work on benchmarking 29](#_Toc120562706)  [8.1 Subtopic [A] 29](#_Toc120562707)  [8.1.1 Publications on benchmarking systems 29](#_Toc120562708)  [8.1.2 Benchmarking by AI developers 29](#_Toc120562709)  [8.1.3 Relevant existing benchmarking frameworks 30](#_Toc120562710)  [8.2 Subtopic [B] 30](#_Toc120562711)  [9 Benchmarking by the topic group 30](#_Toc120562712)  [9.1 Subtopic [A] 31](#_Toc120562713)  [9.1.1 Benchmarking version [Y] 31](#_Toc120562714)  [9.1.2 Benchmarking version [X] 40](#_Toc120562715)  [10 Overall discussion of the benchmarking 40](#_Toc120562716)  [11 Regulatory considerations 41](#_Toc120562717)  [11.1 Existing applicable regulatory frameworks 41](#_Toc120562718)  [11.2 Regulatory features to be reported by benchmarking participants 42](#_Toc120562719)  [11.3 Regulatory requirements for the benchmarking systems 42](#_Toc120562720)  [11.4 Regulatory approach for the topic group 42](#_Toc120562721)  [12 References 43](#_Toc120562722)  [Annex A: Glossary 45](#_Toc120562723)  [Annex B: Information about members (including ex-members) & Declaration of conflict of interests 47](#_Toc120562724) |

**List of Tables**

| Page |
| --- |
| [Table 1: Topic group output documents 9](#_Toc113892132)  [Table 4: Recent documents 12](#_Toc113892133)  [Table 2: Documents recently incorporated into the TDD 13](#_Toc113892134)  [Table 3: Overview of known AI MSK systems and their features 21](#_Toc113892135) |

**List of Figures**

| Page |
| --- |
| [Figure 2: Some of the data of the created synthetic cases. 12](#_Toc113892136)  [Figure 1: Illustration for AI sub-task “Self-Management/Management/Treatment of MSK medicine/Physiotherapy conditions” 25](#_Toc113892137) |

FG-AI4H Topic Description Document

Topic group – Musculoskeletal (MSK) Medicine

# 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).*

# Introduction

Please, note that this version of the document is a work in progress. There are many placeholders and template pieces from the topic description document template.

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]](#footnote-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).

# 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 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.

# 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.

# 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 is an individual contributor from the UK (and he remains to be a topic driver). 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.

## 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).”[[2]](#footnote-2) 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-Q-026-A01 | Latest update of the Topic Description Document of the TG-MSK |
| FGAI4H-Q-026-A02 | Latest update of the Call for Topic Group Participation (CfTGP) |
| N/A | No presentation for the topic group is planned for Meeting Q |

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>

## Status of this topic group

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

### 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.

### 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.

### 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>

### 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 4.
* In particular, synthetic cases were created to support the benchmark work. A screenshot that shows some of that data is provided as Figure 2. These synthetic cases can be used (e.g., if they are exported into a JSON file) for a benchmark prototype.

Table 4: 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#](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 |

Table

Description automatically generated

Figure 2: Some of the data of the created synthetic cases.

### 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. 26th of May 2021: <https://docs.google.com/document/d/13gdDUCOs5NKBFd8B60plWqJ10UpdECD3rqO_A4slQjI/edit>
2. 10th of June 2021: <https://docs.google.com/document/d/1LxI_Ffly_RJ5d16exk03fb5n7tMyiUFSSuEM17mGHC0/edit>
3. 13th of July 2021: <https://docs.google.com/document/d/1UlOWRnlmJTVooNDGuBquYnUqwYieoOh3uxh6oxoTl9o/edit>
4. 29th of July 2021 (meeting with 1 participant): <https://docs.google.com/document/d/1svxb6lO9ETg_AirZnmXPCYwjX0iaagM82pOaGeNbbrI/edit>
5. 31st of August 2021: <https://docs.google.com/document/d/10Nq9_nAoJ5C2xbZO-CpEW6ixW9C7ZwvcBX5qbFgU7nk/edit>
6. 14th of September 2021: <https://docs.google.com/document/d/1hUJBxU9QgRVxon3WlyCmiTFhlMwis_pyFhJoVXLC6uw/edit>

* Raj Sengupta participated in the meeting on the 10th of June 2021. Robert Pawinski participated in the meetings on the 13th of July 2021, 31st of August 2021 and 14th of July 2021.
* Danielle Chulan started being a member on her request dated the 28th of 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 2.

Table 2: 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/1cd0NLO7F9llIH6Pu1CZ8ih68LouDRr5M1VKCYHUSu8Q/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 |

### 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.

### 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?

### 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.

### Status update for meeting [MEETING LETTER]

*Topic Driver: Please insert a one-page summary of the work since the last focus group meeting. This can include:*

* Work on this document
* Work on the benchmarking software
* Progress with data acquisition, annotation, etc.
* Overview of the online meetings including links to meeting minutes
* Relevant insights from interactions with other working groups or topic groups
* Partners joining the topic group
* List of current partners
* Relevant next steps

### Status update for meeting [MEETING LETTER]

[…]

## 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](mailto: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](mailto: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>

# 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.

## AI/ML Prediction for MSK Health

### 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.

### 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].

### 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].

### 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 3.

Table 3: 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 |
| 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 v.s. 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 ≥ 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 world wide.

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].

### 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”.

### 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.

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

Diagram

Description automatically generated

Figure 1: 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.

## Subtopic [A]

### Definition of the AI task

This section provides a detailed description of the specific task the AI systems of this TG are expected to solve. It is *not* about the benchmarking process (this will be discussed more detailed in chapter 4). This section corresponds to [DEL03](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7B7997F2C1-5A1D-4409-B2A0-CBC4E9CE8CDA%7D&file=DEL03.docx&action=default) *“AI requirements specifications*,” which describes the functional, behavioural, and operational aspects of an AI system.

* What is the AI doing?
* What kind of AI task is implemented (e.g., classification, prediction, clustering, or segmentation task)?
* Which input data are fed into the AI model?
* Which output is generated?

### Current gold standard

This section provides a description of the established gold standard of the addressed health topic.

* How is the task currently solved without AI?
* Do any issues occur with the current gold standard? Does it have limitations?
* Are there any numbers describing the performance of the current state of the art?

### Relevance and impact of an AI solution

This section addresses the relevance and impact of the AI solution (e.g., on the health system or the patient outcome) and describes how solving the task with AI improves a health issue.

* Why is solving the addressed task with AI relevant?
* Which impact of deploying such systems is expected (e.g., impact on the health system, overall health system cost, life expectancy, or gross domestic product)?
* Why is benchmarking for this topic important (e.g., does it provide stakeholders with numbers for decision-making; does it simplify regulation, build trust, or facilitate adoption)?

### Existing AI solutions

This section provides an overview of existing AI solutions for the same health topic that are already in operation. It should contain details of the operations, limitations, robustness, and the scope of the available AI solutions. The details on performance and existing benchmarking procedures will be covered in chapter 6.

* Description of the general status and the maturity of AI systems for the health topic of your TG (e.g., exclusively prototypes, applications, and validated medical devices)
* Which are the currently known AI systems and their inputs, outputs, key features, target user groups, and intended use (if not discussed before)? This can also be provided as a table.
* What are the common features found in most AI solutions that might be benchmarked?
* What are the relevant metadata dimensions characterizing the AI systems in this field and with relevance for reporting (e.g., systems supporting offline functions, availability in certain languages, and the capability to process data in a specific format)?
* Description of existing AI systems and their scope, robustness, and other dimensions.

## Subtopic [B]

*Topic driver: If you have subtopics in your topic group, describe how the existing AI solutions in the second subtopic [B] deviate from the description in the previous section. Please use the same subsection structure as above for the first subtopic [A]. If there are no subtopics in your topic group, you can remove the “Subtopic” outline level, but - of course - you need to keep the subsections! In this case, please adapt the lower outline levels accordingly (section numbering).*

# 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.

From the TDD template:

The rapidly evolving field of AI and digital technology in the fields of medicine and public health raises a number of ethical, legal, and social concerns that have to be considered in this context. They are discussed in deliverable [DEL01](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7B0505B020-362C-45B2-94BF-215D2EBBD8F5%7D&file=DEL01.docx&action=default) “*AI4H ethics considerations,”* which was developed by the working group on “Ethical considerations on AI4H” (WG-Ethics). This section refers to DEL01 and should reflect the ethical considerations of the TG-MSK.

* What are the ethical implications of applying the AI model in real-world scenarios?
* What are the ethical implications of introducing benchmarking (having the benchmarking in place itself has some ethical risks; e.g., if the test data are not representative for a use case, the data might create the illusion of safety and put people at risk)?
* What are the ethical implications of collecting the data for benchmarking (e.g., how is misuse of data addressed, is there the need for an ethics board approval for clinical data, is there the need for consent management for sharing patient data, and what are the considerations about data ownership/data custodianship)?
* What risks face individuals and society if the benchmarking is wrong, biased, or inconsistent with reality on the ground?
* How is the privacy of personal health information protected (e.g., in light of longer data retention for documentation, data deletion requests from users, and the need for an informed consent of the patients to use data)?
* How is ensured that benchmarking data are representative and that an AI offers the same performance and fairness (e.g., can the same performance in high, low-, and middle-income countries be guaranteed; are differences in race, sex, and minority ethnic populations captured; are considerations about biases, when implementing the same AI application in a different context included; is there a review and clearance of ‘inclusion and exclusion criteria’ for test data)?
* What are your experiences and learnings from addressing ethics in your TG?

# 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.

## Subtopic [A]

*Topic driver: If there are subtopics in your topic group, describe the existing work on benchmarking for the first subtopic [A] in this section. If there are no sub-topics, you can remove the “Subtopic” outline level, but - of course - you need to keep the subsections below!*

### 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 DEL07 *“AI for health evaluation considerations,”* [DEL07\_1](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7B565EEC0A-D755-41C8-AC68-37B4C38C953F%7D&file=DEL07_1.docx&action=default) *“AI4H evaluation process description,”* [DEL07\_2](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7B58679341-C738-40F0-A822-3AC2B24DD09F%7D&file=DEL07_2.docx&action=default) *“AI technical test specification*,*”* [DEL07\_3](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7BA3088882-F82B-493B-B1C5-49CFF0EEEFA8%7D&file=DEL07_3.docx&action=default) *“Data and artificial intelligence assessment methods (DAISAM),”* and [DEL07\_4](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7BB846B260-373A-41FC-A892-EE5BBCFE3CF8%7D&file=DEL07_4.docx&action=default) *“Clinical Evaluation of AI for health”*.

* What is the most relevant peer-reviewed scientific publications on benchmarking or objectively measuring the performance of systems in your topic?
* State what are the most relevant approaches used in literature?
* Which scores and metrics have been used?
* How were test data collected?
* How did the AI system perform and how did it compare the current gold standard? Is the performance of the AI system equal across less represented groups? Can it be compared to other systems with a similar benchmarking performance and the same clinically meaningful endpoint (addressing comparative efficacy)?
* How can the utility of the AI system be evaluated in a real-life clinical environment (also considering specific requirements, e.g., in a low- and middle-income country setting)?
* Have there been clinical evaluation attempts (e.g., internal and external validation processes) and considerations about the use in trial settings?
* What are the most relevant gaps in the literature (what is missing concerning AI benchmarking)?

### 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.

* What are the most relevant learnings from the benchmarking by AI developers in this field (e.g., ask the members of your topic group what they want to share on their benchmarking experiences)?
* Which scores and metrics have been used?
* How did they approach the acquisition of test data?

### 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 [DEL07\_5](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7B8BFCFF21-3908-4BAD-AB9C-9814EB3F9B36%7D&file=DEL07_5.docx&action=default) *“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).

* Which benchmarking platforms could be used for this topic group (e.g., EvalAI, AIcrowd, Kaggle, and CodaLab)?
* Are the benchmarking assessment platforms discussed, used, or endorsed by FG-AI4H an option?
* Are there important features in this topic group that require special attention?
* Is the reporting flexible enough to answer the questions stakeholders want to get answered by the benchmarking?
* What are the relative advantages and disadvantages of these diverse solutions?

## Subtopic [B]

*Topic driver: If there are subtopics in your topic group, describe the existing work on benchmarking for the second subtopic [B] in this section using the same subsection structure as above. (If there are no sub-topics, you can remove the “Subtopic” outline level.)*

# 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: [DEL05](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7B2012357A-941E-44BD-B965-370D7829F52C%7D&file=DEL05.docx&action=default) *“Data specification”* (introduction to deliverables 5.1-5.6), [DEL05\_1](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7B19830259-F63B-42D4-A408-48C854D6C124%7D&file=DEL05_1.docx&action=default)*“Data requirements”* (which lists acceptance criteria for data submitted to FG-AI4H and states the governing principles and rules), [DEL05\_2](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7B25141F77-E59A-45F1-B081-185C2194FE67%7D&file=DEL05_2.docx&action=default) *“Data acquisition”*, [DEL05\_3](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7B05D8938E-BC2A-4A62-BCB0-1FD46AA72235%7D&file=DEL05_3.docx&action=default) *“Data annotation specification”*, [DEL05\_4](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7BF267A95C-4C5B-4D63-A135-58AF487C3AD3%7D&file=DEL05_4.docx&action=default) *“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), [DEL05\_5](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7B71FE8B9D-ACB3-48CE-AA3F-136409B550A4%7D&file=DEL05_5.docx&action=default) *“Data handling”* (which outlines how data will be handled once they are accepted), [DEL05\_6](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7B5C95327E-96A5-4175-999E-3EDB3ED147C3%7D&file=DEL05_6.docx&action=default) *“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](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7BF5967277-90C8-4252-A0B9-43A5692F35E2%7D&file=DEL06.docx&action=default) *“AI training best practices specification”* (which reviews best practices for proper AI model training and guidelines for model reporting), [DEL07](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7B47E77197-F87B-49F4-80B3-2DD949A5F185%7D&file=DEL07.docx&action=default)*“AI for health evaluation considerations”* (which discusses the validation and evaluation of AI for health models, and considers requirements for a benchmarking platform), [DEL07\_1](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7B565EEC0A-D755-41C8-AC68-37B4C38C953F%7D&file=DEL07_1.docx&action=default) *“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), [DEL07\_2](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7B58679341-C738-40F0-A822-3AC2B24DD09F%7D&file=DEL07_2.docx&action=default) *“AI technical test specification”* (which specifies how an AI can and should be tested *in silico*), [DEL07\_3](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7BA3088882-F82B-493B-B1C5-49CFF0EEEFA8%7D&file=DEL07_3.docx&action=default) *“Data and artificial intelligence assessment methods (DAISAM)”* (which provides the reference collection of WG-DAISAM on assessment methods of data and AI quality evaluation), [DEL07\_4](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7BB846B260-373A-41FC-A892-EE5BBCFE3CF8%7D&file=DEL07_4.docx&action=default)*“Clinical Evaluation of AI for health”* (which outlines the current best practices and outstanding issues related to clinical evaluation of AI models for health), [DEL07\_5](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7B8BFCFF21-3908-4BAD-AB9C-9814EB3F9B36%7D&file=DEL07_5.docx&action=default) *“FG-AI4H assessment platform”* (which explores assessment platform options that can be used to evaluate AI for health for the different topic groups), [DEL09](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7B3E940987-8D75-44B8-85E4-F0E475964F15%7D&file=DEL09.docx&action=default) *“AI for health applications and platforms”* (which introduces specific considerations of the benchmarking of mobile- and cloud-based AI applications in health), [DEL09\_1](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7B1A2EC8D5-53CA-4C8C-9B09-B61CA6F428C5%7D&file=DEL09_1.docx&action=default) *“Mobile based AI applications,”* and [DEL09\_2](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7B3B5A31DE-D3B1-4EC1-A261-2C2E19F73810%7D&file=DEL09_2.docx&action=default) *“Cloud-based AI applications”* (which describe specific requirements for the development, testing and benchmarking of mobile- and cloud-based AI applications).

## Subtopic [A]

*Topic driver: Please refer to the above comments concerning subtopics.*

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.

* Which benchmarking iterations have been implemented thus far?
* What important new features are introduced with each iteration?
* What are the next planned iterations and which features are they going to add?

### Benchmarking version [Y]

This section includes all technological and operational details of the benchmarking process for the benchmarking version [Y] (latest version, chronologically reversed order).

#### Overview

This section provides an overview of the key aspects of this benchmarking iteration, version [Y].

* What is the overall scope of this benchmarking iteration (e.g., performing a first benchmarking, adding benchmarking for multi-morbidity, or introducing synthetic-data-based robustness scoring)?
* What features have been added to the benchmarking in this iteration?

#### Benchmarking methods

This section provides details about the methods of the benchmarking version [Y]. It contains detailed information about the benchmarking system architecture, the dataflow and the software for the benchmarking process (e.g., test scenarios, data sources, and legalities).

##### Benchmarking system architecture

This section covers the architecture of the benchmarking system. For well-known systems, an overview and reference to the manufacturer of the platform is sufficient. If the platform was developed by the topic group, a more detailed description of the system architecture is required.

* How does the architecture look?
* What are the most relevant components and what are they doing?
* How do the components interact on a high level?
* What underlying technologies and frameworks have been used?
* How does the hosted AI model get the required environment to execute correctly? What is the technology used (e.g., Docker/Kubernetes)?

##### Benchmarking system dataflow

This section describes the dataflow throughout the benchmarking architecture.

* How do benchmarking data access the system?
* Where and how (data format) are the data, the responses, and reports of the system stored?
* How are the inputs and the expected outputs separated?
* How are the data sent to the AI systems?
* Are the data entries versioned?
* How does the lifecycle for the data look?

##### Safe and secure system operation and hosting

*From a technical point of view, the benchmarking process is not particularly complex. It is more about agreeing on something in the topic group with potentially many competitors and implementing the benchmarking in a way that cannot be compromised. This section describes how the benchmarking system, the benchmarking data, the results, and the reports are protected against manipulation, data leakage, or data loss. Topic groups that use ready-made software might be able to refer to the corresponding materials of the manufacturers of the benchmarking system.*

This section addresses security considerations about the storage and hosting of data (benchmarking results and reports) and safety precautions for data manipulation, data leakage, or data loss.

In the case of a manufactured data source (vs. self-generated data), it is possible to refer to the manufacturer’s prescriptions.

* Based on the architecture, where is the benchmarking vulnerable to risk and how have these risks been mitigated (e.g., did you use a threat modelling approach)? A discussion could include:
* Could someone access the benchmarking data before the actual benchmarking process to gain an advantage?
* What safety control measures were taken to manage risks to the operating environment?
* Could someone have changed the AI results stored in the database (your own and/or that of competitors)?
* Could someone attack the connection between the benchmarking and the AI (e.g., to make the benchmarking result look worse)?
* How is the hosting system itself protected against attacks?
* How are the data protected against data loss (e.g., what is the backup strategy)?
* What mechanisms are in place to ensure that proprietary AI models, algorithms and trade-secrets of benchmarking participants are fully protected?
* How is it ensured that the correct version of the benchmarking software and the AIs are tested?
* How are automatic updates conducted (e.g., of the operating system)?
* How and where is the benchmarking hosted and who has access to the system and the data (e.g., virtual machines, storage, and computing resources, configurational settings)?
* How is the system’s stability monitored during benchmarking and how are attacks or issues detected?
* How are issues (e.g., with a certain AI) documented or logged?
* In case of offline benchmarking, how are the submitted AIs protected against leakage of intellectual property?

##### Benchmarking process

This section describes how the benchmarking looks from the registration of participants, through the execution and resolution of conflicts, to the final publication of the results.

* How are new benchmarking iterations scheduled (e.g., on demand or quarterly)?
* How do possible participants learn about an upcoming benchmarking?
* How can one apply for participation?
* What information and metadata do participants have to provide (e.g., AI autonomy level assignment (IMDRF), certifications, AI/machine learning technology used, company size, company location)?
* Are there any contracts or legal documents to be signed?
* Are there inclusion or exclusion criteria to be considered?
* How do participants learn about the interface they will implement for the benchmarking (e.g., input and output format specification and application program interface endpoint specification)?
* How can participants test their interface (e.g., is there a test dataset in case of file-based offline benchmarking or are there tools for dry runs with synthetic data cloud-hosted application program interface endpoints)?
* Who is going to execute the benchmarking and how is it ensured that there are no conflicts of interest?
* If there are problems with an AI, how are problems resolved (e.g., are participants informed offline that their AI fails to allow them to update their AI until it works? Or, for online benchmarking, is the benchmarking paused? Are there timeouts?)?
* How and when will the results be published (e.g., always or anonymized unless there is consent)? With or without seeing the results first? Is there an interactive drill-down tool or a static leader board? Is there a mechanism to only share the results with stakeholders approved by the AI provider as in a credit check scenario?
* In case of online benchmarking, are the benchmarking data published after the benchmarking? Is there a mechanism for collecting feedback or complaints about the data? Is there a mechanism of how the results are updated if an error was found in the benchmarking data?

#### AI input data structure for the benchmarking

This section describes the input data provided to the AI solutions as part of the benchmarking of MSK Medicine. It covers the details of the data format and coding at the level of detail needed to submit an AI for benchmarking. This is the only TDD section addressing this topic. Therefore, the description needs to be complete and precise. This section does *not* contain the encoding of the labels for the expected outcomes. It is only about the data the AI system will see as part of the benchmarking.

* What are the general data types that are fed in the AI model?
* How exactly are they encoded? For instance, discuss:
  + The exact data format with all fields and metadata (including examples or links to examples)
  + Ontologies and terminologies
  + Resolution and data value ranges (e.g., sizes, resolutions, and compressions)
  + Data size and data dimensionality

#### AI output data structure

Similar to the input data structure for the benchmarking, this section describes the output data the AI systems are expected to generate in response to the input data. It covers the details of the data format, coding, and error handling at the level of detail needed for an AI to participate in the benchmarking.

* What are the general data output types returned by the AI and what is the nature of the output (e.g., classification, detection, segmentation, or prediction)?
  + How exactly are they encoded? Discuss points like:
    - The exact data format with all fields and metadata (including examples or links to examples)
    - Ontologies and terminologies
* What types of errors should the AI generate if something is defective?

#### Test data label/annotation structure

*Topic driver: Please describe how the expected AI outputs are encoded in the benchmarking test data. Please note that it is essential that the AIs never access the expected outputs to prevent cheating. The topic group should carefully discuss whether more detailed labelling is needed. Depending on the topic, it might make sense to separate between the best possible output of the AI given the input data and the correct disease (that might be known but cannot be derived from the input data alone). Sometimes it is also helpful to encode acceptable other results or results that can be clearly ruled out given the evidence. This provides a much more detailed benchmarking with more fine-grained metrics and expressive reports than the often too simplistic leader boards of many AI competitions.*

While the AI systems can only receive the input data described in the previous sections, the benchmarking system needs to know the expected correct answer (sometimes called ‘labels’) for each element of the input data so that it can compare the expected AI output with the actual one. Since this is only needed for benchmarking, it is encoded separately. The details are described in the following section.

* What are the general label types (e.g., expected results, acceptable results, correct results, and impossible results)?
* How exactly are they encoded? Discuss points like:
  + The exact data format with all fields and metadata (including examples or links to examples)
  + Ontologies and terminologies
* How are additional metadata about labelling encoded (e.g., author, data, pre-reviewing details, dates, and tools)?
* How and where are the labels embedded in the input data set (including an example; e.g., are there separate files or is it an embedded section in the input data that is removed before sending to the AI)?

#### Scores and metrics

*Topic drivers: This section describes the scores and metrics that are used for benchmarking. It includes details about the testing of the AI model and its effectiveness, performance, transparency, etc. Please note that this is only the description of the scores and metrics actually used in* ***this*** *benchmarking iteration. A general description of the state of the art of scores and metrics and how they have been used in previous work is provided in section 3.*

Scores and metrics are at the core of the benchmarking. This section describes the scores and metrics used to measure the performance, robustness, and general characteristics of the submitted AI systems.

* Who are the stakeholders and what decisions should be supported by the scores and metrics of the benchmarking?
* What general criteria have been applied for selecting scores and metrics?
* What scores and metrics have been chosen/defined for robustness?
* What scores and metrics have been chosen/defined for medical performance?
* What scores and metrics have been chosen/defined for non-medical performance?
  + Metrics for technical performance tracking (e.g., monitoring and reporting when the performance accuracy of the model drops below a predefined threshold level as a function of time; computational efficiency rating, response times, memory consumption)
* What scores and metrics have been chosen/defined for model explainability?
* Describe for each aspect
  + The exact definition/formula of the score based on the labels and the AI output data structures defined in the previous sections and how they are aggregated/accumulated over the whole dataset (e.g., for a single test set entry, the result might be the probability of the expected correct class which is then aggregated to the average probability of the correct class)
  + Does it use some kind of approach for correcting dataset bias (e.g., the test dataset usually has a different distribution compared to the distribution of a condition in a real-world scenario. For estimating the real-world performance, metrics need to compensate this difference.)
  + What are the origins of these scores and metrics?
  + Why were they chosen?
  + What are the known advantages and disadvantages?
  + How easily can the results be compared between or among AI solutions?
  + Can the results from benchmarking iterations be easily compared or does it depend too much on the dataset (e.g., how reproducible are the results)?
* How does this consider the general guidance of WG-DAISAM in [DEL07\_3](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7BA3088882-F82B-493B-B1C5-49CFF0EEEFA8%7D&file=DEL07_3.docx&action=default) “Data and artificial intelligence assessment methods (DAISAM)”?
* Have there been any relevant changes compared to previous benchmarking iterations? If so, why?

#### Test dataset acquisition

Test dataset acquisition includes a detailed description of the test dataset for the AI model and, in particular, its benchmarking procedure including quality control of the dataset, control mechanisms, data sources, and storage.

* How does the overall dataset acquisition and annotation process look?
* How have the data been collected/generated (e.g., external sources vs. a process organized by the TG)?
* Have the design goals for the benchmarking dataset been reached (e.g., please provide a discussion of the necessary size of the test dataset for relevant benchmarking results, statistical significance, and representativeness)?
* How was the dataset documented and which metadata were collected?
  + Where were the data acquired?
  + Were they collected in an ethical-conform way?
  + Which legal status exists (e.g., intellectual property, licenses, copyright, privacy laws, patient consent, and confidentiality)?
  + Do the data contain ‘sensitive information’ (e.g., socially, politically, or culturally sensitive information; personal identifiable information)? Are the data sufficiently anonymized?
  + What kind of data anonymization or deidentification has been applied?
  + Are the data self-contained (i.e., independent from externally linked datasets)?
  + How is the bias of the dataset documented (e.g., sampling or measurement bias, representation bias, or practitioner/labelling bias)?
  + What addition metadata were collected (e.g., for a subsequent detailed analysis that compares the performance on old cases with new cases)? How was the risk of benchmarking participants accessing the data?
* Have any scores, metrics, or tests been used to assess the quality of the dataset (e.g., quality control mechanisms in terms of data integrity, data completeness, and data bias)?
* Which inclusion and exclusion criteria for a given dataset have been applied (e.g., comprehensiveness, coverage of target demographic setting, or size of the dataset)?
* How was the data submission, collection, and handling organized from the technical and operational point of view (e.g., folder structures, file formats, technical metadata encoding, compression, encryption, and password exchange)?
* Specific data governance derived by the general data governance document (currently [F-103](https://www.itu.int/en/ITU-T/focusgroups/ai4h/Documents/FGAI4H-F-103-DataPolicy.pdf) and the deliverables beginning with [DEL05](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7B2012357A-941E-44BD-B965-370D7829F52C%7D&file=DEL05.docx&action=default))
* How was the overall quality, coverage, and bias of the accumulated dataset assessed (e.g., if several datasets from several hospitals were merged with the goal to have better coverage of all regions and ethnicities)?
* Was any kind of post-processing applied to the data (e.g., data transformations, repackaging, or merging)?
* How was the annotation organized?
  + How many annotators/peer reviewers were engaged?
  + Which scores, metrics, and thresholds were used to assess the label quality and the need for an arbitration process?
  + How have inter-annotator disagreements been resolved (i.e., what was the arbitration process)?
  + If annotations were part of the submitted dataset, how was the quality of the annotations controlled?
  + How was the annotation of each case documented?
  + Were metadata on the annotation process included in the data (e.g., is it possible to compare the benchmarking performance based on the annotator agreement)?
* Were data/label update/amendment policies and/or criteria in place?
* How was access to test data controlled (e.g., to ensure that no one could access, manipulate, and/or leak data and data labels)? Please address authentication, authorization, monitoring, logging, and auditing
* How was data loss avoided (e.g., backups, recovery, and possibility for later reproduction of the results)?
* Is there assurance that the test dataset is undisclosed and was never previously used for training or testing of any AI model?
* What mechanisms are in place to ensure that test datasets are used only once for benchmarking? (Each benchmarking session will need to run with a new and previously undisclosed test dataset to ensure fairness and no data leakage to subsequent sessions)

#### Data sharing policies

This section provides details about legalities in the context of benchmarking. Each dataset that is shared should be protected by special agreements or contracts that cover, for instance, the data sharing period, patient consent, and update procedure (see also [DEL05\_5](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7B71FE8B9D-ACB3-48CE-AA3F-136409B550A4%7D&file=DEL05_5.docx&action=default) on *data handling* and [DEL05\_6](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7B5C95327E-96A5-4175-999E-3EDB3ED147C3%7D&file=DEL05_6.docx&action=default) on *data sharing practices*).

* Which legal framework was used for data sharing?
* Was a data sharing contract signed and what was the content? Did it contain:
  + Purpose and intended use of data
  + Period of agreement
  + Description of data
  + Metadata registry
  + Data harmonization
  + Data update procedure
  + Data sharing scenarios
    - Data can be shared in public repositories
    - Data are stored in local private databases (e.g., hospitals)
  + Rules and regulation for patients’ consent
  + Data anonymization and de-identification procedure
  + Roles and responsibilities
    - Data provider
    - Data protection officer
    - Data controllers
    - Data processors
    - Data receivers
* Which legal framework was used for sharing the AI?
* Was a contract signed and what was the content?

#### Baseline acquisition

The main purpose of benchmarking is to provide stakeholders with the numbers they need to decide whether AI models provide a viable solution for a given health problem in a designated context. To achieve this, the performance of the AI models needs to be compared with available options achieving the same clinically meaningful endpoint. This, in turn, requires data on the performance of the alternatives, ideally using the same benchmarking data. As the current alternatives typically involve doctors, it might make sense to combine the test data acquisition and labelling with additional tasks that allow the performance of the different types of health workers to be assessed.

* Does this topic require comparison of the AI model with a baseline (gold standard) so that stakeholders can make decisions?
* Is the baseline known for all relevant application contexts (e.g., region, subtask, sex, age group, and ethnicity)?
* Was a baseline assessed as part of the benchmarking?
* How was the process of collecting the baseline organized? If the data acquisition process was also used to assess the baseline, please describe additions made to the process described in the previous section.
* What are the actual numbers (e.g., for the performance of the different types of health workers doing the task)?

#### Reporting methodology

*After the benchmarking, the next step is to describe how the results are compiled into reports that allow stakeholders to make decisions (e.g., which AI systems can be used to solve a pre-diagnosis task in an offline –field –clinic scenario in central America). For some topic groups, the report might be as simple as a classical AI competition leader board using the most relevant performance indicator. For other tasks, it could be an interactive user interface that allows stakeholders to compare the performance of the different AI systems in a designated context with existing non-AI options. For the latter, statistical issues must be carefully considered (e.g., the multiple comparisons problem). Sometimes, a hybrid of prepared reports on common aspects are generated in addition to interactive options. There is also the question of how and where the results are published and to what degree benchmarking participants can opt in or opt out of the publication of their performance.*

This section discusses how the results of the benchmarking runs will be shared with the participants, stakeholders, and general public.

* What is the general approach for reporting results (e.g., leader board vs. drill down)?
* How can participants analyse their results (e.g., are there tools or are detailed results shared with them)?
* How are the participants and their AI models (e.g., versions of model, code, and configuration) identified?
* What additional metadata describing the AI models have been selected for reporting?
* How is the relationship between AI results, baselines, previous benchmarking iterations, and/or other benchmarking iterations communicated?
* What is the policy for sharing participant results (e.g., opt in or opt out)? Can participants share their results privately with their clients (e.g., as in a credit check scenario)?
* What is the publication strategy for the results (e.g., website, paper, and conferences)?
* Is there an online version of the results?
* Are there feedback channels through which participants can flag technical or medical issues (especially if the benchmarking data was published afterwards)?
* Are there any known limitations to the value, expressiveness, or interpretability of the reports?

#### Result

This section gives an overview of the results from runs of this benchmarking version of your topic. Even if your topic group prefers an interactive drill-down rather than a leader board, pick some context of common interest to give some examples.

* When was the benchmarking executed?
* Who participated in the benchmarking?
* What overall performance of the AI systems concerning medical accuracy, robustness, and technical performance (minimum, maximum, average etc.) has been achieved?
* What are the results of this benchmarking iteration for the participants (who opted in to share their results)?

#### Discussion of the benchmarking

This section discusses insights of this benchmarking iterations and provides details about the ‘outcome’ of the benchmarking process (e.g., giving an overview of the benchmark results and process).

* What was the general outcome of this benchmarking iteration?
* How does this compare to the goals for this benchmarking iteration (e.g., was there a focus on a new aspect to benchmark)?
* Are there real benchmarking results and interesting insights from this data?
  + How was the performance of the AI system compared to the baseline?
  + How was the performance of the AI system compared to other benchmarking initiatives (e.g., are the numbers plausible and consistent with clinical experience)?
  + How did the results change in comparison to the last benchmarking iteration?
* Are there any technical lessons?
  + Did the architecture, implementation, configuration, and hosting of the benchmarking system fulfil its objectives?
  + How was the performance and operational efficiency of the benchmarking itself (e.g., how long did it take to run the benchmarking for all AI models vs. one AI model; was the hardware sufficient)?
* Are there any lessons concerning data acquisition?
  + Was it possible to collect enough data?
  + Were the data as representative as needed and expected?
  + How good was the quality of the benchmarking data (e.g., how much work went into conflict resolution)?
  + Was it possible to find annotators?
  + Was there any relevant feedback from the annotators?
  + How long did it take to create the dataset?
* Is there any feedback from stakeholders about how the benchmarking helped them with decision-making?
  + Are metrics missing?
  + Do the stakeholders need different reports or additional metadata (e.g., do they need the “offline capability” included in the AI metadata so that they can have a report on the best offline system for a certain task)?
* Are there insights on the benchmarking process?
  + How was the interest in participation?
  + Are there reasons that someone could not join the benchmarking?
  + What was the feedback of participants on the benchmarking processes?
  + How did the participants learn about the benchmarking?

#### Retirement

*Topic driver: describe what happens to the benchmarking data and the submitted AI models after the benchmarking.*

This section addresses what happens to the AI system and data after the benchmarking activity is completed. It might be desirable to keep the database for traceability and future use. Alternatively, there may be security or privacy reasons for deleting the data. Further details can be found in the reference document of this section [DEL04](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7BC68833D1-9B31-4E8E-8A4A-3939D7DEA56F%7D&file=DEL04.docx&action=default) “*AI software lifecycle specification”* (identification of standards and best practices that are relevant for the AI for health software life cycle).

* What happens with the data after the benchmarking (e.g., will they be deleted, stored for transparency, or published)?
* What happens to the submitted AI models after the benchmarking?
* Could the results be reproduced?
* Are there legal or compliance requirements to respond to data deletion requests?

### Benchmarking version [X]

This section includes all technological and operational details of the benchmarking process for the benchmarking version [X].

*Topic driver: Provide details of previous benchmarking versions here using the same subsection structure as above.*

# Overall discussion of the benchmarking

This section discusses the overall insights gained from benchmarking work in this topic group. This should not be confused with the discussion of the results of a concrete benchmarking run (e.g., in 9.1.1.12).

* What is the overall outcome of the benchmarking thus far?
* Have there been important lessons?
* Are there any field implementation success stories?
* Are there any insights showing how the benchmarking results correspond to, for instance, clinical evaluation?
* Are there any insights showing the impact (e.g., health economic effects) of using AI systems that were selected based on the benchmarking?
* Was there any feedback from users of the AI system that provides insights on the effectiveness of benchmarking?
  + Did the AI system perform as predicted relative to the baselines?
  + Did other important factors prevent the use of the AI system despite a good benchmarking performance (e.g., usability, access, explainability, trust, and quality of service)?
* Were there instances of the benchmarking not meeting the expectations (or helping) the stakeholders? What was learned (and changed) as a result?
* What was learned from executing the benchmarking process and methodology (e.g., technical architecture, data acquisition, benchmarking process, benchmarking results, and legal/contractual framing)?

# Regulatory considerations

*Topic Driver: This section reflects the requirements of the working group on* [***Regulatory considerations on AI for health (WG-RC)***](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/wg/SitePages/WG-RC.aspx) *and their various deliverables. It is* ***NOT requested to re-produce regulatory frameworks****, but to show the regulatory frameworks that have to be applied in the context of your AIs and their benchmarking (****2 pages max****).*

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)*](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/wg/SitePages/WG-RC.aspx) compiled the requirements that consider these challenges.

The deliverables with relevance for regulatory considerations are [DEL02](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7BF2F46A99-7457-4BC8-81A3-0E1E63D6072A%7D&file=DEL02.docx&action=default) *“AI4H regulatory considerations”* (which provides an educational overview of some key regulatory considerations), [DEL02\_1](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7B6AF7C004-8BCE-4151-9F44-45F041A1EB1D%7D&file=DEL02_1.docx&action=default) *“Mapping of IMDRF essential principles to AI for health software”,* and[DEL02\_2](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7B1ED0D4D1-876C-4A0F-AEF7-06D3F445F5E6%7D&file=DEL02_2.docx&action=default) *“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). [DEL04](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7BC68833D1-9B31-4E8E-8A4A-3939D7DEA56F%7D&file=DEL04.docx&action=default) 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.

## 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.

* What existing regulatory frameworks cover the type of AI in this TDD (e.g., MDR, FDA, GDPR, and ISO; maybe the systems in this topic group always require at least “MDR class 2b” or maybe they are not considered a medical device)?
* Are there any aspects to this AI system that require additional specific regulatory considerations?

## 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).

* Which certifications and regulatory framework components of the previous section should be part of the metadata (e.g., as a table with structured selection of the points described in the previous section)?

## 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.

* Which regulatory frameworks apply to the benchmarking system itself?
* Are viable solutions with the necessary certifications already available?
* Could the TG implement such a solution?

## Regulatory approach for the topic group

*Topic Driver: Please select the points relevant for your type of AI and the corresponding benchmarking systems. If your AIs and your benchmarking are not a medical device, this might be quite short.*

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 [DEL02](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7BF2F46A99-7457-4BC8-81A3-0E1E63D6072A%7D&file=DEL02.docx&action=default) *“AI4H regulatory considerations.”*

* Documentation & Transparency
  + How will the development process of the benchmarking be documented in an effective, transparent, and traceable way?
* Risk management & Lifecycle approach
  + How will the risk management be implemented?
  + How is a life cycle approach throughout development and deployment of the benchmarking system structured?
* Data quality
  + How is the test data quality ensured (e.g., the process of harmonizing data of different sources, standards, and formats into a single dataset may cause bias, missing values, outliers, and errors)?
  + How are the corresponding processes document?
* Intended Use & Analytical and Clinical Validation
  + How are technical and clinical validation steps (as part of the lifecycle) ensured (e.g., as proposed in the IMDRF clinical evaluation framework)?
* Data Protection & Information Privacy
  + How is data privacy in the context of data protection regulations ensured, considering regional differences (e.g., securing large data sets against unauthorized access, collection, storage, management, transport, analysis, and destruction)? This is especially relevant if real patient data is used for the benchmarking.
* Engagement & Collaboration
  + How is stakeholder (regulators, developers, healthcare policymakers) feedback on the benchmarking collected, documented, and implemented?

# References

*Topic driver: Add the bibliography here.*

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 |
| 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 |  |
| FGAI4H | Focus Group on AI for Health |  |
| AI | Artificial intelligence |  |
| ITU | International Telecommunication Union |  |
| WHO | World Health Organization |  |
| DEL | Deliverable |  |
| CfTGP | Call for topic group participation |  |
| AI4H | Artificial intelligence for health |  |
| IMDRF | International Medical Device Regulators Forum |  |
| MDR | Medical Device Regulation |  |
| ISO | International Standardization Organization |  |
| GDPR | General Data Protection Regulation |  |
| FDA | Food and Drug administration |  |
| SaMD | Software as a medical device |  |
| AI-MD | AI based medical device |  |
| LMIC | Low-and middle-income countries |  |
| GDP | Gross domestic product |  |
| API | Application programming interface |  |
| IP | Intellectual property |  |
| PII | Personal identifiable information |  |
| […] |  |  |

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

In accordance with the ITU transparency rules, this section lists the conflict-of-interest declarations for everyone who contributed to this document. Please see the guidelines in FGAI4H-F-105 “ToRs for the WG-Experts and call for experts” and the respective forms (Application form & Conflict of interest form).

Note that this part of the document, as all other parts of it, is a work in progress and it does not contain all information at this point. Currently, it contains member names and some general information about them and/or their organisation.

*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, [More info may be provided in the next version]

[To be added.]

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.

Azadur Rahman Sarker, [More info may be provided in the next version]

[To be added.]

Ashwini Sathnur, [More info may be provided in the next version]

[To be added.]

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.

Company/Institution/Individual XYZ

A short explanation of the company’s area of activity and how the work on this document might benefit the company and/or harm competitors. A list of all people who contributed to this document on behalf of this company and any personal interest in this company (e.g., shares).

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

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. [↑](#footnote-ref-1)
2. To be created. [↑](#footnote-ref-2)