EHDEN
EUROPEAN HEALTH DATA & EVIDENCE NETWORK

806968 – EHDEN

European Health Data & Evidence Network

WP3 – Personalized medicine

D3.3 Assessment of regulatory requirements for patient level decision making

Lead contributor	Seamus Kent (9– NICE)
Lead contributor email	Seamus.Kent@nice.org.uk
Other contributors	Jacoline Bouvy (9 – NICE) Peter Rijnbeek (1- EMC) Ross Williams (1- EMC) Aniek Markus (1- EMC) Luis Prieto (1- EMC)
Due date	30/04/2020
Delivery date	20/05/2020
Deliverable type	R
Dissemination level	PU
DoA - Version	V1
Date	12/11/2018

TABLE OF CONTENTS

DOCUMENT HISTORY

Version	Date	Description
V1.0	29/01/2020	First draft
V2.0	02/04/2020	Revised version
V3.0	05/05/2020	Final draft for Consortium Review
V4.0	20/05/2020	Final version

DEFINITIONS

Participants of the EHDEN Consortium are referred to herein according to the following codes:

EMC	Erasmus Universitair Medisch Centrum Rotterdam- The Netherlands **(Project Coordinator)**
Synapse	Synapse Research Management Partners S.L. - Spain
UOXF	The Chancellor, Masters and Scholars of the University of Oxford - United Kingdom
UTARTU	Tartu Ulikool - Estonia
AVRU	Universidade de Aveiro – Portugal
The Hyve	The Hyve BV – the Netherlands
Odysseus	Odysseus Data Services SRO – Czech Republique
EPF	Forum Europeen des Patients (FPE) - Luxembourg
NICE	National Institute for Health and Care Excellence – United Kingdom
UMC	Stiftelsen WHO Collaborating Centre for International Drug Monitoring - Sweden
ICHOM	International Consortium for Health Outcomes measurement LTD - United Kingdom
Janssen	Janssen Pharmaceutica NV - Belgium **(Project Lead)**
Pfizer	Pfizer Limited – United Kingdom
Abbvie	AbbVie Inc - United States
IRIS	Institut De Recherches Internationales Servier - France
SARD	Sanofi Aventis Recherche & Developpement - France
Bayer	Bayer Aktiengesellschaft - Germany
Lilly	Eli Lilly and Company Limited – United Kingdom
AZ	AstraZeneca AB - Sweden
Novartis	Novartis Pharma AG - Switzerland
UCB	UCB Biopharma SPRL - Belgium
Celgene	Celgene Management SARL - Switzerland

Grant agreement	The agreement signed between the beneficiaries and the IMI JU for the undertaking of the EHDEN project (806968).
Project	The sum of all activities carried out in the framework of the Grant Agreement.
Consortium	The EHDEN Consortium, comprising the above-mentioned legal entities.
Consortium agreement	Agreement concluded amongst EHDEN participants for the implementation of the Grant Agreement. Such an agreement shall not affect the parties' obligations to the Community and/or to one another arising from the Grant Agreement.

PUBLISHABLE SUMMARY

The EHDEN network has the potential to support the development of externally validated clinical prediction models at high speed and in a transparent manner. Under the incoming Medical Device Regulations in Europe (2017/745), clinical prediction models that are used to inform further testing or treatment decisions for patients are likely to be classified as medium to high risk medical devices and be regulated accordingly. The clinical evidence required to demonstrate performance and safety over a device's lifetime and the intensity of post-market surveillance are greater than in the outgoing directives. This, in turn, is likely to lead to greater scrutiny by health technology assessment bodies and has prompted the development of evidence standards for the demonstration of clinical benefit. The development and use of such models must also adhere to other European and national regulations such as the General Data Protection Regulation.

To ensure that high-quality clinical prediction models developed through the EHDEN network can usefully inform clinical practice, it is essential that developers take into account these changing regulatory requirements. In this report we provide an overview of the incoming regulations and evidence standards for clinical algorithms, as well as best practice methods for clinical prediction model development. We discuss the challenges that this will pose for those developing clinical prediction models on the EHDEN platform and provide guidance for the future development of such models.

Models developed on the EHDEN network using OHDSI tools are well placed to provide substantial evidence on a model's performance both in-sample and in one or more external datasets and to do so in a transparent manner following the principles of open science. However, it is also important to demonstrate clinical effectiveness against usual practice, ideally through a randomised controlled trial. Developers must also establish mechanisms to collect data on the real-world use of their clinical algorithms to support post-market surveillance and vigilance activities.

Some commentators have argued that European regulations require decisions made based on clinical algorithms to be explainable (conceptualisations of which differ), which may impede the use of certain machine learning methods. However, there remains considerable debate about to what extent European regulations demand this. Others argue that demonstrable performance benefits are of utmost importance.

It is also essential that those developing prediction models engage with healthcare practitioners, managers, and patients to ensure that any model developed is acceptable to clinicians and patients and can be integrated within existing clinical software systems. Finally, ethical concerns (e.g. due to unfair discrimination) need to be placed at the forefront of algorithmic development and application.

1. INTRODUCTION

The EHDEN network will facilitate the development of clinical prediction models which can provide prognostic information to patients and potentially guide clinical decisions about further testing or treatments for patients. Under the incoming European regulations for medical devices, clinical prediction models are expected to be classified as medical devices and regulated accordingly. This, in turn, is likely to have repercussions for their evaluation by health technology assessment (HTA) bodies. Clear guidance is needed on the evidentiary requirements to satisfy regulatory and HTA bodies to ensure that the EHDEN network and OHDSI tools are developed, and studies conducted, in ways that support the adoption of these models in healthcare practice.

In this report we:

- Describe the types of personalised medicine tools that can be generated through the EHDEN network
- Describe the existing and changing regulatory and HTA frameworks for medical devices, with a focus on the evidentiary requirements
- Identify the challenges in developing prognostic models in the EHDEN framework that meet regulatory requirements and support clinical decision making and provide guidance for future work

2. PREDICTIVE MODELLING IN EHDEN

The EHDEN network is a federated database network consisting of many European observational healthcare datasets all mapped to the OMOP common data model. The EHDEN network can be used to develop models and for safety surveillance. EHDEN itself is not responsible for models developed on its network.

Personalised medicine is concerned with stratifying patients to specific treatment pathways or therapies based on their specific characteristics [1]. Personalised medicine tools can provide information on disease risk, prognosis, diagnosis, or treatment response. Types of tools include complex algorithms, health apps, and omics-based biomarkers.

The EHDEN network is expected to be used predominantly to develop clinical algorithms to estimate the risks of future outcomes (positive or negative) and treatment response (i.e. prognostic models). It may also be used to predict intermediate outcomes like prescription adherence. Such tools can simply provide information to patients and healthcare practitioners or can be used to inform clinical decisions, such as whether to conduct further testing or treat a patient. They can be based on simple rule-based expert systems or generated using traditional regression modelling techniques or machine learning methods.

The open-science OHDSI collaborative has developed a suite of tools to perform patient level prediction tasks [2,3]. These currently include several machine learning methods (e.g. lasso regression, random forest, logistic regression) which predict outcomes over a large number of parameters (potentially thousands, though this is often reduced during model development) in observational databases. The fact that data are standardised across the network means the model can be tested in any external dataset in the network subject to standard data governance processes (and assuming adequate coverage and quality of included data elements). A series of dashboards are available which describe model performance in the estimation sample as well as external data sets. Performance metrics include measures of discrimination and calibration. Model development follows best practices for prognostic model development as laid out in the TRIPOD statement [4]. Finally, there is an emphasis on transparency, with all generated analytical code being available to support replication and external validation using open-source tools.

3. REGULATION OF MEDICAL DEVICES

3.1 European regulations

3.1.1 Medical device regulation 2017/745

The regulation of medical devices in the EU is currently undergoing a transition. The medical device directive 93/42 is being replaced by regulation 2017/745 (MDR) [5]. The transition period started in May 2017 and the new regulations were due to fully apply from 26 May 2020. The European Commission has delayed this deadline by one year due to the COVID-19 outbreak.

The new regulations introduce several key changes, including to:

- Regulatory responsibilities
- Definitions of devices and risk classification
- Quality of clinical evidence required to demonstrate performance, safety, and acceptable risk-benefit
- Post-market surveillance requirements

3.1.2 Who is responsible for regulating medical devices?

The regulation of medical devices is largely delegated to national competent authorities, who in turn appoint accredited notified bodies (commercial organisations) to conduct assessments. Medical devices must undergo conformity assessment to demonstrate that they meet relevant legal requirements and to ensure that they are safe and perform as intended. Assessment typically involves an audit of the manufacturer's quality control systems and, for some devices, a review of technical documentation on the performance and safety of the device [6].

The EMA is only responsible for evaluating combination products, i.e. devices which include a medicinal substance as an integral or ancillary part [7]. However, if a company submission includes a prognostic model the EMA will evaluate it, though clear criteria for their evaluation have not yet been established. The EMA prefers evidence from randomised controlled trials. The EMA encourages companies to engage with the EMA through Innovation Task Force meetings prior to submission, though these are not compulsory.

3.1.3 Medical device definition and risk classification

The MDR includes 'software' (defined as 'a set of instructions that processes input data and creates output data') intended for a 'specific medical purpose' in its definition of a medical device (Figure 1). Prognostic models may fall under this definition and therefore be subject to pre-market certification and post-market surveillance processes.

Medical Device Regulation (EU) 2017/745 Article 2(1)	'"medical device" means any instrument, apparatus, appliance, software, implant, reagent, material or other article intended by the manufacturer to be used, alone or in combination, for human beings for one or more of the following specific medical purposes: ■ diagnosis, prevention, monitoring, prediction, prognosis, treatment or alleviation of disease ■ diagnosis, monitoring, treatment, alleviation of, or compensation for, an injury or disability ■ investigation, replacement or modification of the anatomy or of a physiological or pathological process or state ■ providing information by means of in vitro examination of specimens derived from the human body, including organ, blood and tissue donations and which does not achieve its principal intended action by pharmacological, immunological or metabolic means, in or on the human body, but which may be assisted in its function by such means…'

Figure 1. Article 2(1) of the Medical Device Regulation

Medical devices are categorised into one of the following risk classes:

- Class I – low risk
- Class IIa – medium risk
- Class IIb – medium/high risk
- Class III – high risk

Risk classification determines the evidence requirements to achieve conformity assessment and post-market surveillance activities. All devices of class IIa or higher require approval from a notified body while class I devices require only self-assessment unless they are sterile, have a measuring function, or are re-usable surgical instruments.

Rule 11 of Annex 2 (Figure 2) states that software (which should include prediction models) will be considered as class IIa or greater if used to inform treatment decisions. Clinical algorithms providing information (e.g. risk of a cardiovascular event) but not guiding treatment decisions are considered class I medical devices.

6.3. Rule 11

Software intended to provide information which is used to take decisions with diagnosis or therapeutic purposes is classified as class IIa, except if such decisions have an impact that may cause:

— death or an irreversible deterioration of a person's state of health, in which case it is in class III; or

— a serious deterioration of a person's state of health or a surgical intervention, in which case it is classified as class IIb.

Figure 2. Risk classification of software algorithms (MDR: Annex 2, Rule 11)

3.1.4 What are the evidence requirements?

Manufacturers (or developers) must carry out a conformity assessment to demonstrate that their medical device meets the requirements in the MDR. Once complete the device is issued with a CE mark to allow marketing in the European Union.

As part of the conformity assessment procedures, manufacturers must submit technical documentation to notified bodies demonstrating that the device is suitable for its intended purpose, safe, effective, and with an acceptable risk-benefit ratio. The components of the technical documentation are specified in Table 1.

The required fields of the technical documentation are similar to those in the outgoing directives but there is general agreement that the quality of evidence required, particularly for clinical evidence, and the rigour of post-market surveillance are greater [8–11].

Table 1. Required Content of Technical Documentation as per MDR

(a) Annex II – Technical Documentation:
1. Device description and specification, including variants and accessories
1.1 Device description and specification
1.2 Reference to previous and similar generations of the device
2. Information to be supplied by the manufacturer
3. Design and manufacturing information
4. General safety and performance requirements
5. Benefit–risk analysis and risk management
6. Product verification and validation
6.1 Pre-clinical and clinical data
6.2 Additional information required in specific cases
(b) Annex III – Technical Documentation on Post Market Surveillance:
1. The post-market surveillance plan
2. The PSUR (Periodic Safety Update Report)
3. PMS Report

Clinical algorithms must be designed to ensure repeatability, reliability, and performance in line with their intended use, and development must take into account the principles of development life cycle, risk management, including information security, verification, and validation. Development teams (or manufacturers) must include at least one full-time employee with demonstrable expertise in medical device regulation. The algorithms must be developed to be compatible with the systems into which they will be integrated.

Product verification and validation requires information on the design and results of tests and evaluation of published literature applicable to the device. Developers must prepare a clinical evaluation report with sufficient amounts of detail to allow a qualified assessment of whether the device achieves the intended clinical benefits and safety when used as intended. The process to generate, collect, analyse, and assess the clinical data must be methodologically sound, systematic, and planned. Relevant data includes serious and non-serious incidents and side-effects, feedback and complaints, relevant specialist or technical literature, databases and/or registers, and publicly available information about similar medical devices. There are no clear standards with respect to clinical study design.

Developers should collect data for continuous reassessment of the benefit-risk analysis and risk management, develop a post-market clinical follow-up plan, and submit periodic safety summary and post-market surveillance reports. Developers must also undertake vigilance to track and report serious adverse events. The use of registries for post-marketing data collection is encouraged.

Where developers consider their device equivalent to an existing device, they must submit evidence to demonstrate its technical and clinical equivalence. This avoids the need for clinical evaluation but not for post-marketing surveillance.

3.1.5 General Data Protection Regulation

Another relevant piece of European legislation is the General Data Protection Regulation (GDPR; 2016/679), which defines the rights of data subjects and the responsibilities of those controlling and processing data in terms of information provision, transparency and explanation [12].

GDPR codifies a right to explanation through Article 22 as well as other Articles (e.g. 51(1)) concerning general principles of transparency [13]. Two conditions are required regarding the processing of data in Article 22: (1) that it is based solely on automated processing (i.e. there is no human involvement in the decision process), and; (2) it produces legal effects concerning or similarly affecting the data subject. In practice, healthcare professionals will use clinical prediction models as a decision tool but bring other considerations to bear on the treatment decision. As such it is not clear that these two conditions will be met in the case of clinical prediction models [13].

Similarly, there is disagreement among commentators about when explanation is required, i.e. before or after data processing, what needs to be explained, i.e. the model as a whole or individual decisions, and what kind of explanation is required [13,14].

Data controllers and processers must also ensure that they receive informed consent to use the data on the subjects for scientific research and that any use of the data is compatible with the original purpose.

3.2 US Food and Drugs Administration

Traditionally the FDA has reviewed medical devices through an appropriate pre-market pathway (e.g. pre-market clearance [510k], de novo classification, or pre-market approval). The FDA may also clear modifications to medical devices depending on the significance of the risk posed by modification, and it is argued that changes to machine learning algorithms would require pre-market review.

Recognising the limitations of this framework as they pertain to software, particularly dynamic AI models, the FDA has proposed a framework which shifts the regulatory emphasis from products to organisations and from pre-market review to post-marketing surveillance [15]. It is anticipated that this would remove the need for repetitive FDA clearances for software updates in many circumstances.

A pre-certification procedure would focus on the quality and transparency of the manufacturer's processes rather than on the devices themselves. Manufacturers would have to submit a "pre-determined control plan" outlining anticipated modifications and associated methodology of implementation.

Key components of the evidence requirements of this framework are presented in Figure 4.

Data Management	➢ For new training & test data: • Collection protocols • Quality assurance • Reference standard determination ➢ Auditing and sequestration of training and test sets
Re-training	➢ Re-training objectives ➢ Changes related to: • ML methods, including architecture and parameters • Data pre-processing ➢ Criteria to initiate performance evaluation
Performance Evaluation	➢ Assessment metrics ➢ Statistical analysis plans ➢ Frequency and triggers for evaluation ➢ Performance targets ➢ Methods for testing with "clinicians in the loop" when necessary
Update Procedures	➢ Software verification and validation ➢ When and how updates will be implemented ➢ Plans for global and local updates ➢ Communication and transparency to users

Figure 3. US FDA's AI framework [15]

3.3 Other perspectives

The TRIPOD statement was created in order to support good research practices in the design, conduct, analysis, and reporting of prediction models [4]. More recently, the PROBAST tool has been developed to help assess the risk of bias of studies developing, validating or updating prediction models [16,17]. Some have argued that these tools, though essential, do not fully account for the nuances of machine learning tools [18]. Vollmer et al. [18] identify twenty critical questions for the evaluation of machine learning methods in healthcare covering study design and conduct, statistical methods, reproducibility, impact evaluation, and implementation.

Similarly, Parikh et al. [19] argue that although good practice recommendations are useful for improving study quality, they are not sufficient to demonstrate that an algorithm delivers real clinical benefits, and that predictive analytics should meet regulatory standards of clinical evidence. They identify five pillars for the evaluation and regulation of predictive algorithms:

1. Use meaningful endpoints in clinical evaluation, i.e. established standards of clinical benefit (this is consistent with International Medical Device Regulators Forum guidelines [20])
2. Compare performance to an appropriate benchmark (e.g. clinician's predictions or guideline-based prediction scores), ideally within a randomised controlled trial design
3. Algorithms should be interoperable and generalisable
4. Downstream interventions should be explicitly stated
5. Algorithms should be subject to rigorous audit mechanisms

Van Calster et al. [21] argue that algorithms, where used to make treatment decisions, should be publicly available. This would support independent external validation, assessment of performance heterogeneity across settings and over time, and algorithm refinement or updating. Parikh et al. [19] recognise a need to balance transparency with the proprietary interests and intellectual property of the algorithms developers.

Despite the strong interest in developing clinical prediction models, there is too little focus on embedding these models within clinical practice, thereby limiting their impact [22,23]. A survey of healthcare practitioners in England found the main barriers to use were that they were often time-consuming and difficult to use and did not always add value [24]. Some clinical areas such as cardiovascular disease primary prevention have a surfeit of clinical prediction models, with over 800 constructed [17] and at least 10 in widespread use [25] (most clinical prediction models are never used [18]).

There is too little focus on assessing the real-world impact of the use of these different clinical prediction models and clear guidance on preferred models [18,22,26]. In evaluating a model's performance it is important to provide information on clinical utility in addition to statistical measures of a model's performance (e.g. calibration, discrimination), and it should be compared to current clinical practice [18]. Performance should be evaluated in one or more external data sets and there should be a continuous (or regular) process to counteract the potential for model deterioration over time (due to the future diverging from the past).

There is a growing recognition that ethical concerns need to be placed at the forefront of algorithmic development and application [27]. Algorithms can encode bias and discrimination, for instance, by applying only to specific social groups or by selection of certain outcomes, or they can be used in a way that creates unfair discrimination.

4. Health Technology Assessment of Medical Devices

4.1 Consideration of medical devices by HTA bodies

HTA bodies and payers determine coverage, reimbursement and/or pricing of medical technologies. They assess comparative effectiveness against alternative technologies. Some bodies additionally consider the net benefits taking into account additional costs [28]. Many countries have bodies evaluating medical devices, as well as pharmaceuticals [29]. However, we have not yet identified examples of HTA bodies evaluating clinical prediction algorithms, although there are a number of examples of health economic evaluations of prognostic models [30]. This may relate to the limited regulatory oversight applied to such models to date and may change in response to the incoming MDR.

Even though HTA bodies may not evaluate clinical prediction models alone, they are accepted when accompanying health technology assessment submissions for medicinal products or other health technologies and in the development of clinical guidelines [31]. Betts et al. [31] survey the use of various primary prevention cardiovascular disease prediction models by HTA bodies and developers of clinical guidelines. Where criticisms of these models are raised these commonly relate to a lack of applicability to the jurisdiction (e.g. the use of Framingham risk equations in the UK), being outdated, poor calibration to specific populations (by personal characteristics, risk, or treatment status), and inappropriate covariates.

4.2 The Evidence Standards Framework

NICE, in collaboration with NHS England, Public Health England, and MedCity, recently developed the Evidence Standards Framework for Digital Health Technologies (DHTs) [32]. The framework describes standards for the evidence that should be available, or developed, for DHTs to demonstrate their value in the UK health and care system. This includes evidence of effectiveness relevant to the intended use(s) of the technology and evidence of economic impact relative to the financial risk.

The framework distinguishes between fixed algorithms and adaptive algorithms and applies only to the former. Separate standards, the code of conduct for data-driven health and care technology, developed by

the UK Department of Health and Social Care, apply to such technologies [33]. The code of conduct has additional considerations pertaining specifically to adaptive AI and machine learning technologies.

The evidence standards framework categorises DHTs into risk categories and specifies evidence requirements for each risk category (called tiers). For each tier there are two sets of evidence standards: a minimum set and a best practice set. Evidence requirements are cumulative such that technologies in any given tier must meet the requirements pertaining to that tier and to all preceding tiers.

Clinical prediction algorithms are likely to be in the highest risk tier, tier 3b, defined as: "Digital health technologies with measurable user benefits, including tools used for treatment and diagnosis, as well as those influencing clinical management through active monitoring or calculation." They must therefore meet all evidence standards in the framework (see Annex 1).

In brief, developers must show that the DHT generates accurate, reproducible, and relevant data in the target population, and demonstrate effectiveness by undertaking, at a minimum, a high-quality intervention study showing improvements in patient relevant outcomes compared to the current care pathway. Ideally, randomised controlled trials would be conducted. Ongoing data collection is required to show usage in real clinical practice and evaluate user outcomes.

5. IMPLICATIONS FOR EHDEN

5.1 Regulatory processes and risk classification

Clinical prediction models will be regulated as medical devices. Evaluations will be undertaken by notified bodies on behalf of competent authorities (i.e. national regulators). Where prognostic models are used to guide further treatment decisions, they are likely to be classified as medium or high-risk medical devices (i.e. class IIa or higher). Such models may also be evaluated by the EMA when used in relation to a specific medicinal product.

5.2 Producing clinical evidence

If prognostic models are classified as medium or high-risk medical devices, then developers will have to provide clinical evidence to support the use of their models in clinical practice.

The EHDEN network and OHDSI tools are well placed to provide comprehensive information on model construction and performance both in the data set(s) in which the model was constructed and in other data sets in the network. The open-source tools and dashboards support the transparent reporting of this information in line with best practices laid out in the TRIPOD statement [2,4].

However, to demonstrate clinical effectiveness it may be necessary to conduct investigative studies. This is certainly a requirement of the evidence standards framework and may also be required for regulatory approval. These investigative studies would likely have to be conducted extraneously to the EHDEN network. The MDR is not prescriptive as to the form that these evaluations should take. Following Parikh et al.'s [19] guidelines for the evaluation of clinical algorithms, NICE's Evidence Standards Framework [32], and Vollmer et al. [18] they should be assessed in high-quality investigative studies (ideally RCTs) using established measures of clinical benefit and appropriate comparators. Further work is required to elucidate the processes model developers should follow in demonstrating clinical effectiveness for regulatory approval.

5.3 Demonstrating performance

A key component of the MDR and Evidence Standards Framework is demonstrating that the algorithms can be deployed in clinical practice. This requires consideration of how algorithms developed in EHDEN can be implemented in existing clinical software systems.

Some clinical algorithms developed under the principle of parsimony require only a few key parameters which are routinely collected (e.g. QRISK [34]). These are easier to implement in existing software systems or online calculators can be used.

Though simple models can, and have been, developed using OHDSI tools, the standard approach to algorithmic development in the network creates challenges. The models typically contain a much larger number of parameters and some of these parameters will not appear in all electronic healthcare systems, or the quality of data or extent of missing data may differ across these systems. This prevents the use of simple online calculators and poses challenges for implementation of these models in existing software systems. These constraints should be taken into account in the development of algorithms within the EHDEN network that are intended to inform clinical practice.

The development of clinical prediction models should involve strong and early engagement with healthcare practitioners to ensure that the model is designed to answer important clinical questions and is implementable such that it can contribute to decision making and with patients to ensure that the model is acceptable to them [18].

There are also questions about the use of a model in datasets with different amounts of missing data and data quality which may impact the performance of the model [17]. Developers should consider imposing constraints on the application of an algorithm, e.g. setting minimum standards for data quality or contents, or demanding recalibration. Where recalibration is performed, the performance of the algorithm needs to be understood [35].

5.4 Post-marketing surveillance activities

The new medical device regulations require developers to proactively and systematically collect data on the use of devices in clinical practice and observed outcomes.

The EHDEN network could potentially be used as a source of post-marketing surveillance data. Although this would depend on how well the data sets in the network capture the use of software in clinical practice in addition to concerns about data comprehensiveness, representativeness, and quality.

5.5 Development processes and responsibilities

Under the new medical device regulations, model developers must meet regulatory standards including post-marketing surveillance activities and employment of at least one full-time person with expertise in medical device regulation. This may require substantial resources to be devoted to the development and maintenance of clinical algorithms. The EHDEN network will facilitate the development of models but it will not be responsible for further regulatory activities.

A related question is whether algorithms developed on the EHDEN network will be freely available and non-proprietary as argued for by some commentators [21]. However, this may have implications for the capacity of developers to meet the regulatory requirements.

5.6 Meeting GDPR requirements

Sensitive pseudonymised personal data are used by data holders within the OHDSI and EHDEN data networks in the construction and testing of clinical prediction models. Data controllers must adhere to GDPR regarding data subject consent and use of data.

The EHDEN network is a federated data network, meaning no individual person-level data is shared across the network, but only analytical code and summary results.

The extent to which GDPR requires models to be explainable and what this means in practice are discussed in Section 5.7.

5.7 Human interpretability

There are no specific requirements in the MDR regulations itself that models be human interpretable or subject to white box testing. Instead, the focus is on safety and effectiveness which can be evaluated even when the model lacks human interpretability. Furthermore, the contributions of particular parameters to a prediction do not possess causal meaning. However, it has been argued that human interpretability may support the requirements for validation and verification of the model.

The right to explanation in the GDPR may be more directly pertinent, although there is substantial disagreement about this [11,12]. Potential risks of a lack of human interpretability include bias and discrimination. Algorithms developed in EHDEN should consider the ethical implications of any algorithms developed and mitigate risks of unfair discrimination or bias.

Not all machine learning models lack human interpretability, and for those which do, methods are available to improve interpretability. Explainable AI is a rapidly evolving field that aims to improve human interpretability of models and decisions to create appropriate trust [31]. Human interpretability can be tackled by developing an intrinsically interpretable model, possibly at the cost of predictive accuracy, or by accompanying a model with a post-hoc explanation. Potential explanations include the importance of features, influential instances or a surrogate model. However, what a good explanation is or how human interpretability can be formally evaluated, are still open research problems [36].

Vollmer et al. [18] argue that more limited interpretability compared to human decision making or the use of simplified rules must be justified by improvements in predictive accuracy and clinical outcomes.

5.8 Fixed versus dynamic models

Clinical algorithms may be completely fixed, i.e. developed at one point in time using a given data cut, or they may be dynamic, i.e. regularly (or, continuously) updated to include more data or to change the model (e.g. different specification or covariates).

For machine learning models, new data requires retraining of the model such that it may have different performance characteristics. Models can be updated at regular but infrequent intervals (e.g. QRISK is updated annually), or on a continuous or very regular basis whenever new data becomes available or alternative methodologies are applied.

MDR 2017/745 does recognise the possibility of changes to devices over time albeit in a very structured way. It is less well suited to dynamic models which are updated on a continuous or very regular basis. The feature is explicitly recognised in the FDA AI framework [15].

The frequency with which prediction models developed on the EHDEN platform would be updated is likely to depend on the specific application and data set(s) used. Where models undergo substantial amendments, either to data or methods, they may need to be evaluated as a new device. Where these changes are more modest, developers may claim equivalence, which would then be demonstrated in post-market surveillance activities. The frequency and extent of updates has important implications for the workload imposed on model developers.

5.9 Other

The focus of this report has been on the evidence requirements for prognostic models to meet regulatory and HTA requirements. We have not discussed the implications for regulators and HTA bodies themselves, but these are substantial. The market for clinical algorithms is very different to pharmaceuticals, with low costs of entry and substantially more competing products [1]. For instance, there are around 800 prognostic models in cardiovascular disease alone [17]. In addition, the value proposition of these technologies depends on the downstream tests and treatments, and changes to these may necessitate re-evaluation of the algorithm [1]. This is likely to have major implications of how prognostic models are regulated and evaluated in practice.

6. CONCLUSIONS

Clinical algorithms are classified as medical devices in the incoming European regulations and will be regulated accordingly. For their algorithms to be used in clinical practice, developers will need to demonstrate performance and safety over their life cycle. The increased evidence and surveillance requirements should be taken into account by those developing prognostic models on the EHDEN network. Finally, it is essential that developers design models for integration into existing clinical systems if they are to inform clinical decisions. EHDEN will continue to follow developments in this field and engage further with regulatory bodies.

7. REFERENCES

1 Love-Koh J, Peel A, Rejon-Parrilla JC, *et al.* The Future of Precision Medicine: Potential Impacts for Health Technology Assessment. *PharmacoEconomics* 2018;**36**:1439–51. doi:10.1007/s40273-018-0686-6

2 Observational Health Data Sciences and Informatics. *The Book of OHDSI*. 2019. https://ohdsi.github.io/TheBookOfOhdsi/

3 Reps JM, Schuemie MJ, Suchard MA, *et al.* Design and implementation of a standardized framework to generate and evaluate patient-level prediction models using observational healthcare data. *Journal of the American Medical Informatics Association* 2018;**25**:969–75. doi:10.1093/jamia/ocy032

4 Collins GS, Reitsma JB, Altman DG, *et al.* Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD). *Annals of Internal Medicine* 2015;**162**:735. doi:10.7326/l15-5093-2

5 European Parliament. Medical Device Regulation 2017/745. 2017. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32017R0745

6 How medical devices are regulated in the UK. Drug Ther. Bull. 2010;**48**:82–4. doi:10.1136/dtb.2020.03.0023

7 EMA. Medical devices. 2019.https://www.ema.europa.eu/en/human-regulatory/overview/medical-devices

8 BSI Group. Medical Devices Regulation What you need to know. 2017. https://www.bsigroup.com/globalassets/localfiles/es-mx/dispositivos-medicos/bsi-md-mdr-resumen-webinar-06-05-2017.pdf

9 BSI Group. Technical Documentation and Medical Device Regulation. 2019. https://www.bsigroup.com/en-GB/medical-devices/resources/whitepapers/downloads/

10 Ordish, Johan, Murfet, Hannah, Hall A. Algorithms as medical devices. 2019. https://www.phgfoundation.org/documents/algorithms-as-medical-devices.pdf

11 BSI Group. The European Medical Devices Regulations: What are the requirements for vigilance reporting and post-market surveillance? 2019. https://www.bsigroup.com/en-GB/medical-devices/resources/whitepapers/downloads/

12 European Parliament. General Data Protection Regulation 2016/679. 2016. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679

13 PHG Foundation. A right to explanation? 2019. https://www.phgfoundation.org/documents/discussion-paper-right-to-explanation.pdf

14 PHG Foundation. Regulating algorithms in healthcare: The GDPR and IVDR in practice. 2019.

15 FDA. Artificial Intelligence and Machine Learning in Software as a Medical Device. 2019. https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-software-medical-device

16 Wolff RF, Moons KGM, Riley RD, *et al.* PROBAST: A tool to assess the risk of bias and applicability of prediction model studies. *Annals of Internal Medicine* 2019;**170**:51–8. doi:10.7326/M18-1376

17 Moons KGM, Wolff RF, Riley RD, *et al.* PROBAST: A tool to assess risk of bias and applicability of prediction model studies: Explanation and elaboration. Ann. Intern. Med. 2019;**170**:W1–33.

doi:10.7326/M18-1377

18 Vollmer S, Mateen BA, Bohner G, *et al.* Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics, and effectiveness. *BMJ* 2020;**368**:l6927.

19 Parikh, Ravi B, Obermeyer, Ziad, Navathe A. Regulation of predictive analytics in medicine: Algorithms must meet regulatory standards of clinical benefit. *Science* 2019;**363**:810–2.https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6557272/

20 International Medical Device Regulators Forum. Software as a Medical Device (SaMD): Clinical Evaluation. 2017. http://www.imdrf.org/docs/imdrf/final/technical/imdrf-tech-170921-samd-n41-clinical-evaluation_1.pdf

21 Van Calster B, Wynants L, Timmerman D, *et al.* Predictive analytics in health care: how can we know it works? *Journal of the American Medical Informatics Association : JAMIA* 2019;**26**:1651–4. doi:10.1093/jamia/ocz130

22 Salisbury AC, Spertus JA. Realizing the Potential of Clinical Risk Prediction Models: Where Are We Now and What Needs to Change to Better Personalize Delivery of Care? Circ. Cardiovasc. Qual. Outcomes. 2015;**8**:332–4. doi:10.1161/CIRCOUTCOMES.115.002038

23 Amarasingham R, Patzer RE, Huesch M, *et al.* Implementing electronic health care predictive analytics: Considerations and challenges. *Health Affairs* 2014;**33**:1148–54. doi:10.1377/hlthaff.2014.0352

24 Brown B, Cheraghi-Sohi S, Jaki T, *et al.* Understanding clinical prediction models as 'innovations': A mixed methods study in UK family practice. *BMC Medical Informatics and Decision Making* 2016;**16**. doi:10.1186/s12911-016-0343-y

25 Plüddemann A, Wallace E, Bankhead C, *et al.* Clinical prediction rules in practice: Review of clinical guidelines and survey of GPs. Br. J. Gen. Pract. 2014;**64**. doi:10.3399/bjgp14X677860

26 E. W, M.J.M. U, B. C, *et al.* Impact analysis studies of clinical prediction rules relevant to primary care: A systematic review. *BMJ open* 2016;**6**:e009957.http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=prem&NEWS=N&AN=27008685%5Cnhttp://bmjopen.bmj.com/content/6/3/e009957.full.pdf+html%5Cnhttp://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=emed13&NEWS=N&AN=20160240924

27 Morley, J, Floridi L. An ethically mindful approach to AI for health care. *The Lancet* 2020;**395**:25–31.https://doi.org/10.1016/S0140-6736(19)32975-7

28 EUnetHTA. Analysis of HTA and reimbursement procedures in EUnetHTA partner countries. 2017.https://www.eunethta.eu/national-implementation/analysis-hta-reimbursement-procedures-eunethta-partner-countries/ (accessed 22 Jul 2019).

29 Crispi F, Naci H, Barkauskaite E, *et al.* Assessment of Devices, Diagnostics and Digital Technologies: A Review of NICE Medical Technologies Guidance. *Applied Health Economics and Health Policy* 2019;**17**:189–211. doi:10.1007/s40258-018-0438-y

30 van Giessen A, Peters J, Wilcher B, *et al.* Systematic Review of Health Economic Impact Evaluations of Risk Prediction Models: Stop Developing, Start Evaluating. Value Heal. 2017;**20**:718–26. doi:10.1016/j.jval.2017.01.001

31 Betts MB, Milev S, Hoog M, *et al.* Comparison of Recommendations and Use of Cardiovascular Risk Equations by Health Technology Assessment Agencies and Clinical Guidelines. *Value in Health*

2019;**22**:210–9. doi:10.1016/j.jval.2018.08.003

32 National Institute for Health and Clinical Excellence [NICE]. Evidence standards framework for digital health technologies. 2019. doi:10.1007/978-1-349-95810-8_867

33 Department of Health and Social Care. Code of conduct for data-driven health and care technology. 2019. https://www.gov.uk/government/publications/code-of-conduct-for-data-driven-health-and-care-technology

34 Hippisley-Cox J, Coupland C, Vinogradova Y, *et al.* Derivation and validation of QRISK, a new cardiovascular disease risk score for the United Kingdom: Prospective open cohort study. *British Medical Journal* 2007;**335**:136–41. doi:10.1136/bmj.39261.471806.55

35 Van Calster B, McLernon DJ, Van Smeden M, *et al.* Calibration: The Achilles heel of predictive analytics. *BMC Medicine* 2019;**17**. doi:10.1186/s12916-019-1466-7

36 Guidotti R, Monreale A, Ruggieri S, *et al.* A survey of methods for explaining black box models. *ACM computing surveys (CSUR)* 2018;**51**:93.

ANNEXES

Annex 1 – NICE Evidence Standards Framework for tier 3b digital health technologies

Category	Minimum standard	Best practice
Demonstrating effectiveness	High quality intervention study (experimental or quasi-experimental design) showing improvements in relevant outcomes, such as diagnostic accuracy, patient-reported outcomes (preferably using validated tools) including symptom severity or quality of life, other clinical measures of disease severity or disability, healthy behaviours, physiological measures, user satisfaction and engagement. Generic outcome measures may also be useful when reported alongside condition-specific outcomes. The comparator should be a care option that is reflective of the current care pathway, such as a commonly used active intervention.	High quality randomised controlled study or studies done in a setting relevant to the UK health and social care system, comparing the DHT with a relevant comparator and demonstrating consistent benefit including in clinical outcomes in the target population, using validated condition-specific outcome measures. Alternatively, a well-conducted meta-analysis of randomised controlled studies if there are enough available studies on the DHT.
Reliable information content	Be able to show that any health information provided by the DHT is valid, accurate, up to date, reviewed and updated by relevant experts at defined intervals (e.g. annually), sufficiently comprehensive.	Evidence of endorsement, accreditation or recommendation by NICE, NHS England, a relevant professional body or recognised UK patient organisation. Alternatively, evidence that the information content has been validated though an independent accreditation such as The Information Standard or HONcode certification.
Ongoing data collection to show usage of the DHT.	Commitment to ongoing data collection to show usage of the DHT in the target population, and commitment to share, when available, with relevant decision-makers such as commissioners in a clear and useful format.	Evidence that data on usage is being collected in line with the minimum standards and can be made available to relevant decision-makers.
Ongoing data collection to show value of the DHT.	Commitment to ongoing data collection to show user outcomes (if relevant) or user satisfaction (using non-patient identifiable information) to show ongoing value, and commitment to share, when available, with relevant decision-makers such as commissioners in a clear and useful format.	Evidence that data on outcomes or user satisfaction is being collected in line with the minimum standard and can be made available to relevant decision-makers.
Quality and safeguarding.	Show that appropriate safeguarding measures are in place around peer-support and other communication functions within the platform. Describe who has access to the platform and their roles within the platform. Describe why these people or groups are suitable and qualified to have access. Describe any measures in place to ensure safety in peer-to-peer communication, for example through user agreements or moderation.	As for the minimum evidence standard.

Category	Minimum standard	Best practice
Credibility with UK health and social care professionals	Be able to show that the DHT has a plausible mode of action that is viewed as useful and relevant by professional experts or expert groups in the relevant field. Either: a) show that relevant clinical or social care professionals working within the UK health and social care system have been involved in the design, development or testing of the DHT, or b) show that relevant clinical or social care professionals working within the UK health and social care system have been involved in signing-off the DHT, indicating their informed approval of the DHT	Published or publicly available evidence documenting the role of relevant UK health or social care experts in the design, development, testing or sign-off of the DHT.
Relevance to current care pathways in the UK health and social care system	Evidence to show that the DHT has been successfully piloted in the UK health and social care system, showing that it is relevant to current care pathways and service provision in the UK. Also evidence that the DHT is able to perform its intended function to the scale needed (for example, having servers that can scale to manage the expected number of users).	Evidence to show successful implementation of the DHT in the UK health and social care system.
Acceptability with users	Be able to show that representatives from intended user groups were involved in the design, development or testing of the DHT. Provide data to show user satisfaction with the DHT.	Published or publically available evidence to show that representatives from intended user groups were involved in the design, development or testing of the DHT and to show that users are satisfied with the DHT.
Accurate and reliable measurements	Data or analysis which shows that the data generated or recorded by the DHT is accurate, reproducible, relevant to the range of values expected in the target population. Also data showing that the DHT is able to detect clinically relevant changes or responses	As for the minimum evidence standard.
Accurate and reliable transmission of data	Technical data showing that numerical, text, audio, image-based, graphic-based or video information is not changed during the transmission process and not biased by the data 'value' expected from the target patient population.	As for the minimum evidence standard, but with quantitative data.