# Genomics England Clinical Interpretation Partnership (GeCIP) Detailed Research Plan Form

| Application Summary | |
|---|---|
| **GeCIP domain name** | **Electronic Health Records** |
| **Project title** *(max 150 characters)* | **Electronic health records for genomic medicine and patient benefit: methods and tools** |

**Objectives.** *Set out the key objectives of your research. (max 200 words)*

Leveraging the rich data collected as part of usual care is of pivotal importance for Genomics England to drive transformation in the NHS through genomic medicine for patient benefit. The Farr Electronic Health Records is cross-cutting domain with one initial overarching aim: To develop the underpinning methods, tools and data linkages which will enhance the core genomics England data architecture and accelerate research at early and late translational phases across all GeCIPs.

We will deliver this aim through **6 inter-related objectives (sub-domains)**

- **Subdomain 1: Recruitment:** We will provide Genomic Medicine Centres with business intelligence tools via CogStack to improve patient recruitment and population of disease models.
- **Subdomain 2: Patient perspective:** We will implement the APPROaCH system which empowers patients to remotely record information (PROMS) about their own treatment progress thus enriching the EHR.
- **Subdomain 3: Bringing bioinformatics to health informatics:** We will develop a standardised computational description of pathophysiology that bridges networks of phenotypic measurements made by GMCs, molecular phenotypes and patient record data – an 'Ensembl for EHR'.
- **Subdomain 4: Lifelong longitudinal phenotyping** We will capture the multitude of phenotypically diverse, longitudinal data for patients that span primary and secondary care and non-health data sources using novel data linkage approaches.
- **Subdomain 5: EHR enabled, genomically informed trials** We will develop methods and tools for optimising the design, recruitment, execution and efficacy and adverse event monitoring of trials.
- **Subdomain 6: Driving better NHS outcomes through learning health systems** We will improve knowledge discovery and knowledge translation.

**Lay summary.** *Information from this summary may be displayed on a public facing website. Provide a brief lay summary of your planned research. (max 200 words)*

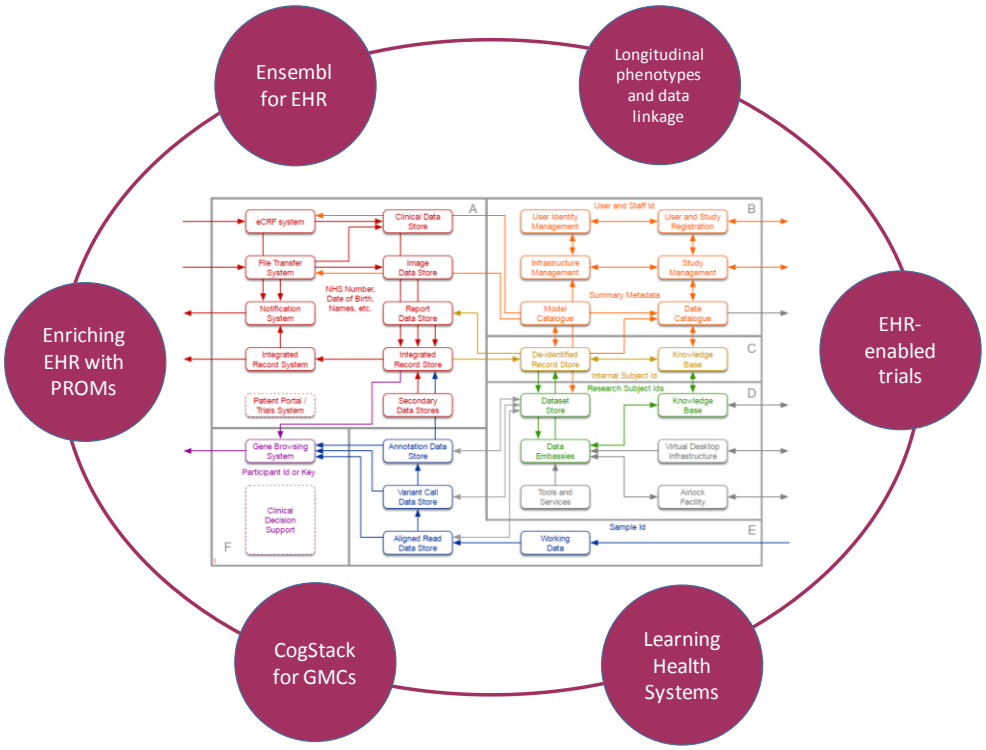
The sequence of health related events that we all experience is increasingly captured in health records, on smart phones or on devices that we wear. In order to get the most benefit from the genetic sequence in GeL for ourselves, our families and for others we need to bring together these two sequences – on the one hand our health as it unfolds over time and on the other hand the string of 3 billion letters that make up our genome.

The Farr Institute seeks to do just that. The major, exciting challenge demands developing new ways of doing research and building new methods and tools in 6 related areas. First, recruitment to help patients get into the 100k project in the first place. Second, to bring in the patients perspective (e.g. reporting how they feel on a smartphone). Third, to cross the bridge between the well established biological toolkits used in genomics and the newer field of examining health records. Fourth, to use the 'cradle to grave' strengths of the NHS to find out about health over

time. Fifth, to offer patients the opportunity to take part in more clinical trials, with less intrusion on their time. Sixth, to ensure that the NHS learns from knew knowledge in order to drive better patient outcomes from genomic medicine.

**Technical summary.** *Information from this summary may be displayed on a public facing website. Please include plans for methodology, including experimental design and expected outputs of the research. (max 500 words)*

The Farr EHR GeCIP six subdomains are seen as cogs in a mechanism to drive translational research and genomic medicine for patient benefit. They are shown related to the GeL data architecture below.




| | |
|---|---|
| **Expected start date** | **ASAP (Contingent on funding availability and decisions)** |
| **Expected end date** | **5 year programme: with near term deliverables tied to GeL priorities** |

## Lead Applicant(s)

| | |
|---|---|
| **Name** | Harry Hemingway |
| **Post** | Director, Farr Institute London |
| **Department** | Institute of Health Informatics |
| **Institution** | UCL |
| **Current commercial links** | |

## Administrative Support

| | |
|---|---|
| **Name** | Giovanna Ceroni |
| **Email** | g.ceroni@ucl.ac.uk |
| **Telephone** | |

| Subdomain leads | | |
|---|---|---|
| **Name** | **Subdomain** | **Institution*** |
| **Richard Dobson** and Jackie Cassell | 1. Recruitment | |
| **Zina Ibrahim** and Mark Duman | 2. Patient perspective and PROMS | |
| **Nick Luscombe** and Tim Hubbard | 3. Ensembl for EHR | |
| **Spiros Denaxas** and Martin Landray | 4. Longitudinal phenotypes and data linkage | |
| **JP Casas** and Folkert Asselbergs | 5. EHR enabled trials for precision medicine | |
| **Brendan Delaney** and Tjeerd van Staa | 6. Learning Health Systems | |

* With proposed UK Farr Institute, all sub-domain leads are drawn from the Farr, including the following universities: UCL, Manchester, KCL, Sussex, Oxford, Imperial, University Medical Centre, Utrecht, Francis Crick Institute. *The wider membership of the Farr GeCIP* is shown in the accompanying spreadsheet.

## Detailed research plans

See separate plans for each of the 6 subdomains:

- **Subdomain 1: Recruitment:**
- **Subdomain 2: Patient perspective:**
- **Subdomain 3: Bringing bioinformatics to health informatics:**
- **Subdomain 4: Lifelong longitudinal phenotyping**
- **Subdomain 5: EHR enabled, genomically informed trials**
- **Subdomain 6: Driving better NHS outcomes through learning health systems**

| Data requirements |
|---|
| **Data scope.** *Describe the groups of participants on whom you require data and the form in which you plan to analyse the data (e.g. phenotype data, filtered variant lists, VCF, BAM). Where participants fall outside the disorders within your GeCIP domain, please confirm whether you have agreement from the relevant GeCIP domain. (max 200 words)*<br><br>The objectives for the Farr Electronic Records domain, as a cross-cutting GeCIP, are somewhat different to that of the disease-specific GeCIPs.<br><br>The data interface for CogStack will primarily be with the GMCs and the NHS trusts.<br><br>PROM data for cancer and rare skin disorder patients will be required.<br><br>The Longitudinal Phenotypes sub-domain will primarily deal with phenotypic data from national electronic health record sources, hospitals, clinical registries and other administrative datasets that we will link GeL with. In the process of creating high-resolution longitudinal phenotypes, we will work in close collaboration with clinical GeCIPs when access to genotypic data is required. |
| **Data analysis plans.** *Describe the approaches you will use for analysis. (max 300 words)* |

We will use a mixture of supervised (i.e. support vector machines) and unsupervised (e.g. k-means clustering) and evaluate the best approaches for linking and phenotyping the different datasets. We will use international metadata standards (e.g. HL7 FHIR, DDI, SDMX, OMOP) and controlled clinical terminologies (e.g. SNOMED-CT) to harmonize data and establish a common data model compatible with clinical GeCIPS, other subdomains, regulatory requirements for trials and NHS ICT requirements.

**Key phenotype data.** *Describe the key classes of phenotype data required for your proposed analyses to allow prioritisation and optimisation of collection of these. (max 200 words)*

Clinical Practice Research Datalink and other primary care EHR data
NICOR registries
Cancer registries
Hospital Episode Statistics + mental health + hospital prescribing
Office of National Statistics mortality data
Secondary care data from hospitals
PROMS, wearables, sensors

**Alignment and calling requirements.** *Please refer to the attached file (Bioinformatics for 100,000 genomes.pptx) for the existing Genomics England analysis pipeline and indicate whether your requirements differ providing explanation. (max 300 words)*

*N/A*

**Tool requirements and import.** *Describe any specific tools you require within the data centre with particular emphasis on those which are additional to those we will provide (see attached excel file List_of_Embassy_apps.xlsx of the planned standard tools). If these are new tools you must discuss these with us. (max 200 words)*

The tools developed as part of this domain will be provided to GeL and to the GMCs as new packages.

Programming tools: Python (with Pandas and numpy and pytoolz), Perl + local CPAN repository
RDBMS: MySQL or Postgres
Workflow tools: cwltool or bpipe or nextflow, galaxy
Ensembl for EHR will require Python, R, BioConductor, and SQL

**Data import.** *Describe the data sets you would require within the analysis environment and may therefore need to be imported or accessible within the secure data environment. (max 200 words)*

**Computing resource requirements.** *Describe any analyses that would place high demand on computing resources and specific storage or processing implications. (max 200 words)*

## Omics samples

**Analysis of omics samples.** *Summarise any analyses that you are planning using omics samples taken as part of the Project. (max 300 words)*

| Data access and security | |
|---|---|
| **GeCIP domain name** | Electronics Records |
| **Project title** *(max 150 characters)* | **Electronic health records for genomic medicine and patient benefit: methods and tools** |

***Applicable Acceptable Uses.*** *Tick all those relevant to the request and ensure that the justification for selecting each acceptable use is supported in the 'Importance' section (page 3).*

X☐ *Clinical care*

X☐ *Clinical trials feasibility*

X☐ *Deeper phenotyping*

X☐ *Education and training of health and public health professionals*

X☐ *Hypothesis driven research and development in health and social care - observational*

X☐ *Hypothesis driven research and development in health and social care - interventional*

X☐ *Interpretation and validation of the Genomics England Knowledge Base*

X☐ *Non hypothesis driven R&D - health*

☐ *Non hypothesis driven R&D - non health*

X☐ *Other health use - clinical audit*

X☐ *Public health purposes*

X☐ *Tool evaluation and improvement*

***Information Governance***

X☐ The lead for each domain will be responsible for validating and assuring the identity of the researchers. The lead may be required to support assurance and audit activities by Genomics England.

Any research requiring access to the embassy will be required to complete IG Training and read and sign a declaration form. Access will only be granted once these requirements have been met.

## Detailed research plan

| Full proposal (total max 1500 words per subdomain) | |
|---|---|
| **Title**<br><br>*(max 150 characters)* | **Farr EHR GeCIP Subdomain 1: Recruitment**<br><br>**CogStack: An Open Source, Enterprise Grade Informatics Platform for NHS Business Intelligence, Audit and Research** |

**Importance.** *Explain the need for research in this area, and the rationale for the research planned. Give sufficient details of other past and current research to show that the aims are scientifically justified. Please refer to the 100,000 Genomes Project acceptable use(s) that apply to the proposal (page 6).*

To date, the GeL Genomic Medicine Centres (GMCs) have found it challenging to meet recruitment targets and provide essential covariate phenotype data for the interpretation of core genomic data. Using existing Electronic Health Records (EHRs) for this purpose has proven ad-hoc, labour intensive, and therefore expensive. It represents a poor return on investment for the GeL legacy requirement. While it is generally accepted that the EHR has a huge untapped potential to address fundamental uncertainties of medicine, the complexities of the EHR and practical considerations of embedding modern analytics into complex organisations such as the NHS has often proven to be an insurmountable obstacle.

EHRs are information-rich resources containing detailed history of the patient. These clinical records represent expensive assets when the cumulative cost of the information in the EHR is considered (data collection, infrastructure, maintenance). Due to the significant proportion of unstructured data in EHRs, the benefits of the information within these records are hard to realise fully using present EHR technologies.

This proposal specifically addresses the immediate mission-critical GeL objectives for the use of data within clinical records at each GMC site for:

1. candidate recruitment and,
2. population of GeL disease models.

These objectives will be met through the provision of an information retrieval and extraction platform: 'CogStack'. Cogstack implements best-of-breed enterprise search, natural language processing, analytics and visualisation technologies that are known to be absent from the majority of GMCs' business and clinical intelligence capabilities. We have demonstrated successes at two GMC sites (South London and The Maudsley NHS Foundation Trust (SLaM) and King's College Hospital (KCH)) where we are already seeing a step-change to GMC capabilities to expedite patient recruitment for the 100KGP and populate phenotype data models. It is already clear that this project has triggered lasting change at this GMC NHS Trust.

We aim to build a legacy within the NHS business intelligence community through the deployment of CogStack across all GMC sites. Beyond the principal goals of the 100KGP, the implementation of the proposed toolset and the associated upskilling in data mining techniques will create a transformational capability in the fields of business intelligence, research and audit. One of the key outcomes will be a dramatically reduced cost of clinical trials in routine practice.

**Research plans.** *Give details of the analyses and experimental approaches, study designs and techniques that will be used and timelines for your analysis. Describe the major challenges of the research and the steps required to mitigate these.*

**Approach:** This GeL EHR GeCIP subdomain is a cross-cutting initiative to provide tooling for GMCs and GeCIPs. The research goals here are centered around identifying, specifying and developing the most appropriate solutions.

Specifically, the new capabilities that CogStack will offer are to the GMCs are:

1. Semantically enhanced search. By integrating the latest search enhancements from the EBI backed BioSolr project, we are able to offer a world leading document search engine. This will enable a range of capabilities for different users, from simple document retrieval to complex, multifaceted ontology queries, making simultaneous use of both structured and unstructured data. This will allow NHS staff to rapidly answer questions such as, “does this patient meet the inclusion criteria for the 100K Genome Project?”, "where are the likely locations of information pertinent to disease model field X?" or, “has this patient received any high-cost treatments that have not been captured in their discharge summary?”
2. Natural Language Processing (NLP), allowing automated information extraction of medical concepts (autocoding) for both the 100KGP and the Trusts’ clinical coding function. For example, this process will convert a sentence such as, “I supplied the patient with 500mg Paracetamol, to be taken PRN,” into a structured format such as, “Drug=Paracetamol, Dose=500mg, Frequency=PRN”
3. Ontological (e.g. SNOMED, UMLS) markup of unstructured text. We will develop and/or licence third party SNOMED annotation tools, such as the Bio-YODIE pipeline developed by The University of Sheffield to provide text markup. In combination with the aforementioned NLP tools, this will enable us to infer ontological relationships between document entities (i.e. drugs, patients, diagnoses, symptoms), creating the data structure required to power semantic search.
4. Capability to perform clinical text de-identification. An option for each deployment will be to create a parallel de-identified research dataset composed of the same structured and unstructured data but with all strong identifiers masked. This would provide the means to utilise local clinical record for research purposes, with the proviso that all the appropriate governance and controls are in place.
5. Analytics and visualisation capability. Beyond the information retrieval/extraction use case, once the structured and unstructured data collected and indexed with NLP annotation in Elasticsearch, the downstream opportunities for analytics are extensive.

Immediate applications include: deep learning, clustering and modelling for patient stratification, identification of optimised care pathways, prognosis modelling and series analysis time pharmacovigilance.

Beyond the technical capabilities of this proposal, we recognise foremost that the use of clinical data in research hinges on an unwavering consideration of the information governance, ethical use and patient engagement implications. We seek to develop a significant patient engagement capacity and governance framework, based around our experiences with the CRIS governance protocol and interface with the CPRD and other national EHR research concerns. To this end, we have already formed technical and governance working groups with the specific focus on safe use of clinical data in research, and are actively progressing an agenda in collaboration with experts throughout the UK.

**Challenges:** EHR systems are all too often closed, proprietary and issues remain with the accuracy, completeness and the challenges of unstructured data. Thus most hospital data lies entirely hidden to the GMCs and wider academic community. This programme of work will deliver on the vital first step, namely unlocking the EHR. Very specifically, this application directly addresses the first directive for the GeCIP partnerships, ' to optimise clinical data and sample collection'. We have already established proof of concept on the full EHRs at the SLaM and KCH Foundation trusts and a further deployment at the UCLH GMC is in progress, through funding from a WT Strategic Award, the UCL NIHR BRC and Farr London. To this end we will be able leverage our experience and lessons learnt from these initial deployments.

**Collaborations including with other GeCIPs.** *Outline your major planned academic, healthcare, patient and industrial collaborations. This should include collaborations and data sharing with other GeCIPs. Please attach letters of support.*

We have already deployed this capability as a business intelligence function in King'S College Hospital and South London and The Maudsley NHS Foundation Trusts (Guy's and St Thomas GMC) through funding provided to the GMCs by NHS England to create legacy IT infrastructure. This is already yielding a huge benefit of streamlining the process of identifying patients fitting GeL recruitment criteria and populating the GeL phenotype models (commendation attached). Importantly, this is proving to have huge implications for the clinical audit and coding business functions. The success of this work has led to further funding (WT Strategic Award) to deploy at UCLH in collaboration with the cancer GeCIP and for neurodegeneration (Swanton, Crick; Alexander UCL).

A large component of this work package will concern the ethical and governance issues associated with the use of pseudonymised and de-identified clinical data in GMC business intelligence and research. To this end, we intend to link with the Ethics and Social Science GeCIP to define patient engagement and empowerment strategies, and localised governance frameworks for clinical

research conducted outside the immediate remit of GeL.

As part of defining the requirements for this proposal, the intention is to work directly with the Machine Learning, Quantitative Methods and Functional Genomics GeCIP, to ensure that the platform is suitable for implementing deep learning and other advanced analytical methods. In addition to complementing the existing methods for cohort identification and data model population, this has the potential to drive localised, trust-specific knowledge discovery, contributing to the GeL legacy and the transformation of the NHS into a data-driven organisation.

Collaborators and partnerships are already emerging in the wider academic and industrial sectors, which include the University of Sheffield for NLP applications such as SNOMED annotation and expertise in clinical governance at the University of Sussex. Enterprise search industrial collaborators include Flax and Opensource Connections.

Finally, this programme has the potential to deliver extensive upskilling of GMC staff, described further below. Naturally, these elements will be managed via interactions with the Education and Training GeCIP.

**Training.** *Describe the planned involvement of trainees in the research and any specific training that will form part of your plan.*

The legacy requirements of the 100KGP holds that members implement modern, sustainable informatics capabilities. In order for Trust functions (such as business intelligence, clinical coding, clinical audit and research) to benefit from these capabilities, the project should deliver a new, self-sufficient unstructured data capability. This proposal includes plans to upskill GMC staff in data mining, data management and basic Natural Language Processing techniques (NLP), in order to seed the growth of this capability. The suggested programme will produce and deliver electronic training materials for an EHR focussed GeL training course, covering the following themes:

Introduction to Big Clinical Data - Structured and Unstructured Data Problems

Clinical Natural Language Processing in Context

Clinical Ontologies and Semantic Search

Data Validation and Visualisation

Applications of Deep Learning

This training programme has been designed jointly with the KCH Deputy Director for IT and clinical coding team at KCH as part of the KCH deployment schedule.

**People and track record.** *Explain why the group is well qualified to do this research, how the investigators would work together.*

The Software Development Team at King's College London BRC for Mental Health designed and deployed the CogStack platform in two NHS trusts, a successful proof of concept demonstrating the team's capabilities for delivery in this area. This software will also form the foundation for further functionality to be built, forming the central hub from which all collaborations will extend.

With respect to governance and ethics, we have researchers with specific interest and background in medical ethics. Secondly, and perhaps more distinctively, we have a role in NHS commissioning which brings us close to the demands of regulators for high quality information system to be available in the NHS. There is very poor practical awareness of the potential of EHR among commissioners and regulators, and we feel that this is an important conversation for the research community if we are to facilitate a commissioning culture that creates and curates research-ready data across the NHS economy.

Since the governance issues necessarily inform the technological capability of the proposed systems we plan to adopt a tandem of communication between a governance working group and a technical working group.

**Clinical interpretation.** *(Where relevant to your GeCIP) Describe your plans to ensure patient benefit through clinical interpretation relevant to your domain. This should specifically address variant interpretation and feedback and your interaction with the cross-cutting Validation and Feedback domain.*

This project is a capability building cross-cutting concern, therefore the patient benefit will be derived through tangible efficiency enhancements in clinical data management.

**Beneficiaries.** *How will the research benefit patients and healthcare institutions including the NHS, other researchers in the field? Are there other likely beneficiaries?*

Key benefactors of the outputs of this subdomain will principally be the GeCIPs and GMCs. Secondary benefits will include clinical informatics research groups that are able to access an anonymised version of the EHR, and thirdly, business intelligence departments of the NHS trusts that compose the GMCs.

1- Recruitment into GeL, EHR based trials and clinical decision support:

At the simplest level, the identification of patients fitting recruitment criteria for GeL or trials is a challenge. Making the data within the record immediately accessible will greatly streamline this process and have an immediate impact on the ability of GMCs to identify suitable patients, and thus, improve recruitment into GeL. However, it also enables the clinician or carer to make informed decisions about next steps in the care pathway through easier access to the relevant information in the patient history. By overlaying additional alerting capability through Watcher (https://www.elastic.co/products/watcher) and/or our software agent capability, we can provide the capacity for automating the alerting and notification process.

2- Research:

With inclusion of an optional drop-in de-identification microservice, CogStack converts the EHR into a research ready data resource (with appropriate governance in place), directly benefitting the GeCIPs and research community more generally.

3- Business Intelligence:

Part of this platform includes the provision of business intelligence tooling and NLP applications to enhance the clinical coding capacity of the Trust. Here, the objective is to determine the high value services that the Trust delivers but does not charge for (either due to insufficient capacity in the clinical coding function or inaccessibility of unstructured data) and create the necessary NLP algorithms accordingly.

Depending on the level of success achieved here, this phase may lead to a further project to expand this concept to its maximum economic viability, developing additional processes and technologies to integrate into the Trusts clinical coding SOPs. Here, the development team will create algorithms and interfaces to provide automatic coding solutions and assisted coding software to clinical coders to increase the efficiency of the manual coding process. The intended document domain is outpatient letters, which is a known source of Trust undercharging.

Clinical decision support is also a putative direction for the use of the information liberated by CogStack, here multi-agent based computer systems are a delivery route for i) tailored clinical decision support (e.g. the personalised decision to prescribe drug A), and ii) delivery of

randomised trials (e.g. intervention X in patients Y). They are autonomous pieces of software that are able to detect and act upon their environment, in this case, they act upon an events identified within the EHR, in real time.

**Commercial exploitation.** *(Where relevant to your GeCIP) Genomics England has a very explicit intellectual property policy. We and other funders need to know if the proposed research likely to generate commercially exploitable results. Do you have commercial partners in place?*

As per feedback received from KCH, it is likely that the commercial potential of this project is high. We are engaged with a variety of SMEs with expertise in various elements of the CogStack. Working with a principly open source community, we anticipate a significant contribution to various high-profile open source projects and an advancement of the open source, meritocratic business model which does not engender vendor lock-in.

Beyond the initial deployment, we intend to ensure the long term sustainability of our proposal. To this end, we include requirements that consider the handoff of day-to-day operations for a variety of concerns. We recognise the importance of stability and support that are can offered by traditional large scale NHS technology providers (Microsoft, etc), as well as the contribution of niche SMEs (Elastic.co Lucidworks, Flax, OpenSource Connections, OntoText) that specialise in the facets of our proposal.

**References.** *Provide key references related to the research you set out.*

**https://www.elastic.co/**

**https://gate.ac.uk/**

**http://www.kconnect.eu/**

## Detailed research plan

| Full proposal (total max 1500 words per subdomain) | |
|---|---|
| **Title**<br>*(max 150 characters)* | **Farr EHR GeCIP 2: Patient Perspectives**<br>**Enriching the EHR phenotype with PROM, wearable and device generated data** |

**Importance.**

We are moving into an era of patient-centred care, and patient-centred informatics. The patient is best placed to provide phenotypic data, either by directly reporting outcomes or by using wearable devices that autonomously collect data. With the patient at the hub of a rapidly increasing quantity of stored information, this has great potential for improving the qualification of disease and subsequent interpretation of genetic data.

Patient-reported Outcome measures (PROM) provide vital insight into patient experience with care and are an important clinical tool whose use has evolved from research and monitoring quality of care to supporting outcome improvement (Black'13). Not only do PROMs provide data on large numbers of patients, but are representative of typical, everyday practice, thus facilitating research on the observational effectiveness (rather than controlled efficacy) of treatments (Devlin'10). PROMs also represent high-quality standardised phenotypes enabling the quantification of patient state, comparisons of treatments and patient response to therapy. In this regard, PROMs can be used to generate a greatly an enriched phenotype with benefit to the 100k Genomics England project.

In addition to PROM data, passive remote monitoring using the sensors on mobile phones and wearable devices can greatly enhance our understanding of the clinical phenotype, and provide powerful and potentially disruptive means to support innovation in, and democratization of, healthcare delivery. Streamed data has the potential to support early diagnosis, prognosis and provide a means for stratification and intervention delivery. The data captured offers the first real opportunity to collect objective metrics at high resolution in daily life providing a more complete picture of the patient, and a new class of phenotype target to complement genetic, molecular and neuroscience research. We will aim to build on our present success in this area with a €22M (additional €1.2m contribution from tech partners Intel and SoftwareAG) IMI2 funded project: RADAR-CNS coordinated by KCL making the proposed work outlined here highly cost-effective.

To achieve all this, we need a user-centred design to build informatics tools that provide the right functionality and information for patients as well as for healthcare professionals. Our objective is to create and demonstrate the benefits of a computing infrastructure supporting an interoperable and personalized environment enabling remote patient-led phenotype provision through PROMs and wearable-generated data directly to hospital EHRs.

**Research plans**

The aim of this proposal is to extend the context of the EHR beyond the clinician-patient interface through:

**(a) PROMs:** to take our APPROaCh (Ibrahim'15) system, originally developed as a platform to address the lack of ADHD-specific PROM data and deployed at the South London and The Maudsely NHS Foundation Trust (SLaM), and apply it in at least two GeCIP exemplars disease areas (rare skin disorders (McGrath, KCL) and cancer (Swanton, Crick)) thus enriching the patient phenotype and demonstrating clinician and

patient benefit.

**(b) Mobile and Remote Monitoring Sensors:**
A rapidly growing range of mobile sensors technologies are being developed. These technologies have the potential to continuously and passively monitor patients in daily life, producing high-resolution and low-burden objective information on a patient. The research plan is to determine how to bring this information into the EHR and to the clinician in a useful actionable form. The tooling to be developed will center around:

- Patient stratification: based on objective, high-resolution and high-content clinical covariate or phenotype data from remote monitoring devices.
- Characterisation of disease etiology, trajectories and new endo-phenotypes.
- Emphasis on more effective and scalable community care: reduction hospital beds has meant that increasing numbers of patients live in the community managing chronic conditions.
- In-community monitoring: to remedy the lack of frequent monitoring of patients in-community, we will develop infrastructure to facilitate, reduce costs, and improve safety of community-based care where appropriate.
- Opportunity for real time monitoring: detection and prediction of adverse events (and precursors to adverse events), pharmaco-compliance

The APPROaCH system empowers patients to remotely record information about their own treatment progress with no reliance on clinical staff to collect data. In the case of ADHD, APPROaCh automates the collection of the Strength and Difficulties Questionnaire (SDQ) (Goodman'97), a PROM targeting the wellbeing of adolescent mental health patients. APPROaCh is deployed at SLaM, and at predefined times contacts patients/carers via emails encouraging them to remotely report their SDQ progress and provides immediate visual and textual assessments, storing SDQ results in the patient's EHR and fed back to the patient. We will extend and generalise APPROaCh to support four streams of work.

*Workstream 1: Creating an interoperable and secure platform integrated into the EHR*
A distributed and integrated infrastructure supporting PROM provision and utilisation for decision support will require the harmonisation, retrival and extraction of information from the EHR, patient-reported data and knowledge inferred using the decision support and monitoring components. It will also generate issues pertaining to ensuring the security and integrity of the data being transported over distributed systems.

Challenges Addressed:

- Supporting a unified data representation via a common ontology (Harrocks'03), describing patient profiles and PROMs, importing relevant external ontologies (e.g. medication), and embedding standardised ICD10 coding.
- Exploiting role-based security models (Xiang'07) to embed flexible access and security preserving addressing the privacy, security and integrity issues resulting from the cross-organisational exchanges of patient data as well as giving patients access to their clinical records.

We will build on the ontology and associated reasoning component developed for APPROaCh to encompass a wide range of GeI-specific disease spectra. Local expertise at SLaM with systems such as MyHealthLocker & Microsoft HealthVault will be leveraged to provide secure and data storage and visualisation personal health record (PHR) platform and define role-based and context-specific access control and data transport models.

*Workstream 2: Patient Incentivisation*
The infrastructure will achieve a set of different tasks for a given patient: issuing suitable

personalised reminders, developing habits (Lally'10), designing, displaying, and optimising feedback. Engaging patients and maintaining engagement is a non-trivial and multi-factorial problem, at the core of which is the delivery and reinforcement of personally relevant information and goals. The usability of PROMs (Benson'13) and of informatics tools like EHRs (Greenhalgh'10; Greenhalgh'09) on a day-to-day basis is critical to successful adoption.

Challenges addressed:

- Identify, understand and implement psychological aspects and user interface design methods for optimising patient experience.
- Investigate and build machine learning models for extracting and exploiting EHR and system usage information (e.g. patient disorder, stage in disorder PROM response history, response to treatment, etc.) to personalise the timing, language and visual interfaces used to request PROM data in order to maximise patient engagement and completion rates.
- Build a knowledge framework mapping PROM values to clinical interpretations of PROM results, the latter to be provided to patients as incentives upon completing PROM questionnaires or after automatically collected data (sensor) is analysed.

*Workstream 3: Clinical Decision Support*

Integrating data collection with decision support will ensure the continuous usage and create a drive to acquire more data on the clinicians' side. In our context, by decision support we mean the ability to detect regression or adversity, subsequently alerting clinicians in an appropriate, useful and timely manner.

Challenges addressed:

- Build a machine learning model to detect deleterious events from reported PROMs, system usage information (or lack thereof) as well as treatment trajectories and timeline.
- Build a decision support component based on clinical guidelines for a wide coverage of Gel-specific conditions
- Integrating genomic information into the personalisation of the decision support system.
- Developing a plan for transforming a prototype implementation to a deployed EHR-agnostic decision support system.

*Workstream 4: Information for the Patient*

Decision support systems and PHRs also need to deliver the goal of providing the right information at the right time to patients and gene carriers (Modell'00). We will draw on the Accessible Publishing of Genetic Information (APoGI) project developing metadata for patient information in haemaglobinopathies; and on work on tailored risk predictions (Martin'08), where genomic information is expected to contribute to more behaviour change (Silarova'15).

Challenges address:

- Build a general metadata model of genomic information for patients.
- Integrate genomic information into tailored risk prediction systems for patients.
- Extend HER/genomic connectivity into the PHR space.

**Collaborations including with other GeCIPs.**

We have a rich portfolio of collaborations covering academic, tech industry, pharma, and the NHS through our EHR, wearables & devices and genomics research programmes. Technology partners Intel, FitBit and best of breed real-time data streaming company SoftwareAG, EFPIA partners Janssen, Biogen, Merck, Lundbeck, UCB, as well as The

Michael J Fox and TranSMART foundations. Key hospital partners include SLaM, KCH and UCLH.

At SLaM we have created a hospital software dev & staging environment through the Centre for Translational Informatics. We will mirror this model at other GMC sites and work with the trials subdomain to formally trial the effectiveness of each PROM

**Training**

We will build local capacity and develop training in the critical areas of health informatics, computing and data science research by working with the Farr institute to establish suitable PhD studentships and utilising potential capacity from the Master's programmes in Health Informatics and in Data Science at the Farr and partner institutions.

**People and track record**

Our sub-domain will fully utilise the expertise of its members whose expertise span: software engineering, health and medical informatics, clinical Bioinformatics, Artificial Intelligence, data science, machine learning, omics, data analysis, clinical decision support and medical knowledge representation. We will draw on the existing partnership between the UCL Centre for Beahviour Change and UCL Institute of Digital Health to recruit interdisciplinary expertise. We will seek to actively collaborate with the best national and international research groups including Professor Michael Luck (KCL), Dr Simon Miles (KCL) and Professor Brendan Delaney (Imperial).

**Clinical interpretation.**

We will deliver this working in collaboration with the other clinical GECiPs, and will firstly identify exemplar case studies through the cancer GeCIP (Swanton), rare skin disease GeCIP (McGrath, KCL), and Gastroenterology and Hepatology (Hirschfield, Birmingham) beyond our current exemplars in mental health at SLaM. The aim would then be to further deploy in other targeted GeCIPs.

**Beneficiaries**

Automating and enriching the process of PROM collection, interpretation, presentation and utilisation will benefit:

1. **Clinical workers**: by improving the availability of standardised outcome data, increasing efficiency and maximising resource allocation.
2. **Organisations**: improve the transparency for care by producing readily available data for audits and evaluation purposes.
3. **Patients**: empowering patients by creating an environment where they take charge of their progress, leading to increased awareness and engagement in treatment.
4. **Research:** increasing the amount of quality phenotypic data generated during routine clinical care.

**Commercial exploitation.**

We will write applications that are agnostic to the data source, once harvested we will leverage platforms such as Open mHealth and the Human API providing standardised structured health data, which will assist healthcare, commercial and individual stakeholders exchange data and reuse software code.

**References.**

Benson T, Potts HWW, Whatling JM, Patterson D (2013). Comparison of howRU and EQ-5D measures of health-related quality of life in an outpatient clinic. Informatics in Primary Care, 21(1), 12-7

Black, N; 2013; Patient reported outcome measures could help transform healthcare; British Medical Journal; 346:f167

Greenhalgh T, Stramer K, Bratan T, Byrne E, Russell J, Potts HWW (2010). Adoption and

non-adoption of a shared electronic summary record in England: A mixed-method case study. BMJ, 340, c3111.

Greenhalgh T, Potts HWW, Wong G, Bark P, Swinglehurst D (2009). Tensions and paradoxes in electronic patient record research: A systematic literature review using the meta-narrative method. Milbank Quarterly, 87(4), 729-88.

Devlin, N and Appleby, J; 2010; Getting the most out of PROMs; The King's Fund.

Horrocks, I; PatelSchneider, P;Van Harmelen;F; 2003; From SHIQ and RDF to OWL: The Making of a Web Ontology Language; Journal of Web Semantics; 1(1);726.

Ibrahim, Z; Fernandez de la Cruz, L; Stringaris, A; Goodman, R; Luck, M; Dobson, R; 2015; A Multi Agent Platform for Automating the Collection of PatientProvided Clinical Feedback; Proceedings of The 14th International Conference on Autonomous Agents and Multiagent Systems; 157165.

Lally P, van Jaarsveld CHM, Potts H, Wardle J (2010). How are habits formed: Modelling habit formation in the real world. European Journal of Social Psychology, 40(6), 998-1009.

Martin CJ, Taylor P, Potts HWW (2008). Construction of a comprehensive odds model of cardiovascular and coronary heart disease using published information: The Cardiovascular Health Improvement Model (CHIME). BMC Medical Informatics & Decision Making, 8, 49.

Modell B; Darlison M; New Developments in Genetics for the New Millennium: The Concept of Clinical Bioinformatics. Community Genet 2000;3:184–189.

Silarova B; Lucas J; Butterworth AS; et al.; 2015. Information and Risk Modification Trial (INFORM): design of a randomised controlled trial of communicating different types of information about coronary heart disease risk, alongside lifestyle advice, to achieve change in health-related behavior. BMC Public Health; 15:868.

L. Xiang, L. et al; 2007; An adaptive security model for multi-agent systems and application to a clinical trials environment. In *IEEE COMPSAC* 2:261–268.

## Detailed research plan

### Full proposal (total max 1500 words per subdomain)

| | |
|---|---|
| **Title**<br>*(max 150 characters)* | **Farr EHR GeCIP 3:**<br>**Framework for disease and syndrome phenotyping – An Ensembl database for Electronic Health Records** |

**Importance.** *Explain the need for research in this area, and the rationale for the research planned. Give sufficient details of other past and current research to show that the aims are scientifically justified. Please refer to the 100,000 Genomes Project acceptable use(s) that apply to the proposal (page 6).*

There is ongoing debate regarding disease classification and developing novel phenotypic cohorts of patients that may encompass one or more disease. A "new taxonomy of disease" based on underlying biology, rather than traditional descriptives, will allow dramatically different approaches to diagnosis, treatment and prognosis for risk evaluation. Many diseases currently studied by GeL have not been formally defined in computable terms using existing clinical terminologies. Additionally, for diseases where diagnosis code do exist, the resolution provided is coarse and as a result substantial overlap and ambiguity between data elements exists Raw genomic and EHR data are not research nor clinic-ready and a substantial amount of work is required in order to transform them into a resource that can be analysed and interpreted. The EHR GeCIP provides an ideal opportunity to curate omic, biomarkers, imaging, EHR and other data, and present them as clinically meaningful disease models.

The **aims** of this sub-domain are:

(i) to develop novel computational software infrastructure for automatically curating genomic and EHR data;
(ii) to use these curations to help develop a "new taxonomy" of disease, based on computable disease models;
(iii) to feed back these models to GMCs as a way to continually improve patient care and data collection;
(iv) to provide programmatic access to curated disease models for researchers.

We will develop and evolve the **EHR research community** by fundamentally shifting the cultural landscape, providing a 'go-to' resource of information, tools and knowledge exchange and promoting sharing and transparency.

The subdomain is closely aligned with the principle aims of the Farr Institute and the major MRC Medical Bioinformatics Awards, and it builds on ongoing initiatives such as the UK Biobank.

**Research plans.** *Give details of the analyses and experimental approaches, study designs and techniques that will be used and timelines for your analysis. Describe the major challenges of the research and the steps required to mitigate these.*

<u>What is Ensembl?</u>

The Ensembl genome infrastructure provides a successful combination of (i) an automated annotation pipeline to curate genomic data, (ii) a graphical user interface, as well as (iii) APIs/webservices to navigate, query and analyse the curated information. By using the genomic coordinate as a reference, Ensembl integrates disparate types of annotation: some information is intrinsic to the genome (eg, exons, LD blocks), whereas others are extrinsic (eg, expression levels, signalling pathways linking gene products). The underlying curated data and coordinate system allows for (i) chromosome annotation across different scales (*e.g.* from single SNPs to whole LD blocks), (ii) graphical navigation, as well as (iii) APIs/webservices for the programmatic access for

large-scale querying and analytics. By combining structural, evolutionary and functional features of the genome provide the basis for extracting information and biological understanding of the organism in question. This curated form of information also allows users to compare genomes across species, bridge genomic data to functional databases, as well as disease records. The open-source infrastructure is generally applicable to different genomes; the main database hosted at the Sanger Institute and European Bioinformatics Institute receives ~10M unique accesses annually and it has become an essential resource for genomic researchers.

An Ensembl for EHR

The lessons learnt from managing knowledge of chromosomal structure and function are applicable to the development of an infrastructure to manage knowledge about drug and disease mechanisms implicitly embedded in data collected by GeL. Using a patient's life history as the underlying coordinate system, it will be possible to map events recorded in Electronic Health Records, phenotypic measurements and genomic information. By curating such information across patients, we shall curate disease models that can be interpreted for clinical decisions.

In practice, we shall develop a standardised computational description of pathophysiology that bridges relevant (i) networks of phenotypic measurements collected by the GMCs with (ii) networks of measurements describing molecular phenotypes and (iii) patient record data. For example, prototyping this approach discussed above in the area of rare diseases of the kidney, the deliverables of the subdomain would include:

1) a multiscale knowledge model of renal anatomy, in terms of standard reference ontologies for gross, cellular and subcellular anatomy [http://obofoundry.org/].
2) a review, in collaboration with key opinion leaders in the field, of pathophysiology mechanisms documented in rare kidney disease, in particular:
    a. Anti Glomerular Basement Membrane (GBM) disease;
    b. Cystinuria;
    c. Hyperoxaluria;
    d. EAST syndrome;
    e. Nephrotic Syndrome.
3) for each condition reviewed in #3 above, the authoring of a network linking clinical and molecular phenotypic measurements, in terms of the reference kidney model in #1 and stable identifiers for:
    a. clinical terminologies in EHRs (*e.g.* SNOMED-CT), and
    b. molecular resources at the European Bioinformatics Institute (*e.g.* Ensembl, Intact, ArrayExpress), as well as IMI platforms for pharmaceutical knowledge (*e.g.* OpenPHACTS).
4) Tools for the cross-querying, browsing and visualisation of pathophysiology networks of renal disease over the spectrum of mechanisms in #3a-3e.

We will build on frameworks developed from OMOP, ORPHA.NET (www.orpha.net Phenotype terminologies in use for genotype-phenotype databases: A common core for standardisation and interoperability HVP5), with an initial focus on rare disease (http://www.slideshare.net/variomeproj/phenotype-terminologies-in-use-for-genotypephenotype-databases-a-common-core-for-standardisation-and-interoperability) and *European Nutritional Phenotype Assessment and Data Sharing Initiative* (ENPADASI). http://www.healthydietforhealthylife.eu/index.php/enpadasi, BBMRI-LPC (Biobanking and Biomolecular Resources Research Infrastructure – Large Prospective Cohorts) http://www.bbmri-lpc.org/ and the international serious adverse events consortium, iSAEC Phenotype standardisation project (http://www.saeconsortium.org/?q=node/31). We will include validation for phenotype based on minimal dataset requirements.

We will develop and evolve the **EHR research community** by fundamentally shifting the cultural landscape, providing a 'go-to' resource of information, tools and knowledge exchange and promoting sharing and transparency. We will establish awareness and engagement through user groups and scientific meetings across the health data science communities. The resource will be implicitly promoted by all research output that gets generated based on it across collaborators, nationally and internationally.

Risks and challenges

The main risk associated with the work in this proposal is the ever changing landscape of legal and information governance requirements with regards to patient sensitive identifiers which are required for linkage. We will systematically explore and validate the best approach for performing linkages between sources including the use of trusted third party services, HSCIC accredited safe havens, GMC safe havens and pseudonymization approaches in order to mitigate the risk of not being able to access the required identifiers. Patient confidentiality is of critical importance and we will seek to develop privacy-aware approaches to ensure its upheaval.

**Collaborations including with other GeCIPs.** *Outline your major planned academic, healthcare, patient and industrial collaborations. This should include collaborations and data sharing with other GeCIPs.*

We will leverage our extensive network of on-going cross-disciplinary collaborations in order to undertake the research outlined in this proposal.

Academic: Farr national network of institutions and partners across England, Scotland and Wales, Francis Crick Institute, Alan Turing Institute, Oxford Big Data Institute, UCL Institute of Big Data, QMUL, LSHTM, UCL CTU, eMedLab and other MRC-funded Medical Bioinformatics centres, Francis Crick Institute, King's, Administrative Research Data Network across England, Scotland, Wales and Norther Ireland, UK Biobank, CTSU

Initiatives: EMIF, EuroRec, EHR4CR, CDISC, TRANSFoRM, Global Alliance for Genomics and Health

Industry: ABPI, Cerner, EMIS, Intel, IBM

Healthcare: AHSC's distributed across academic partners mentioned

Patient: Genomics England Patient and Public Participation Network

Public sector: HSCIC, ONS, PHE

We will work across all GeCIPs and more widely within the Farr network, EU projects and with the USA and Japan to develop exemplars, test, iterate and evaluate systems based on these basic information concepts.

**Training.** *Describe the planned involvement of trainees in the research and any specific training that will form part of your plan.*

We will create and deliver short courses across all levels (e.g. undergraduate, postgraduate and continuous professional development) on electronic health record and administrative data linkages. We will seek to develop modules in existing or emerging postgraduate degrees across all partners and create online delivery modules for popular online platforms such as Coursera.

Our training programme will have a strong emphasis on providing the crucial bridging between scientific disciplines and domains, particularly between clinical practitioners, health informaticians/computer scientists and bioinformaticians. Our training and educational programme will be outward looking and inclusive and we will seek to collaborate with relevant partners, organisations, industry and doctoral training programmes and the NHS in order to deliver our education objectives – examples include but are not limited to: ELIXIR UK, Health Education England MSc in Genomic Medicine, Modernizing Scientific Careers Clinical Bioinformatics Programme.

The developed material will provide attendees with a solid understanding of the fundamental principles of data linkage and relevant methodologies (e.g. probabilistic vs. deterministic) and with hands-on experience through a set of computer practical sessions using contemporary and realistic training datasets.

**People and track record.** *Explain why the group is well qualified to do this research, how the investigators would work together.*

Investigators belong to the Farr Network and/or an MRC Medical Bioinformatics Consortium; therefore the work is already supported through an existing organisational structure. The expertise in the group extends across clinical practise and research, public health, epidemiology, bioinformatics, health informatics and computer science. Denaxas is a Senior Lecturer in Biomedical Informatics at the Farr Institute in London; Luscombe is a bioinformatician and academic lead of the eMedLab MRC Medical Bioinformatics Award; de Bono is a Principle Research Fellow in Health Informatics; Hubbard is Professor of Bioinformatics and Head of Bioinformatics at Genomics England; Molohkia is a clinician and Senior Lecturer in Clinical Epidemiology; Dobson is Senior Lecturer and Head of Bioinformatics at the NIHR Biomedical Research Centre for Mental Health and the South London and Maudsley NHS Trust.

**Clinical interpretation.** *(Where relevant to your GeCIP) Describe your plans to ensure patient benefit through clinical interpretation relevant to your domain. This should specifically address variant interpretation and feedback and your interaction with the cross-cutting Validation and Feedback domain.*

As part of a cross-cutting GeCIP, we will deliver this in collaboration with other clinical domains.

**Beneficiaries.** *How will the research benefit patients and healthcare institutions including the NHS, other researchers in the field? Are there other likely beneficiaries?*

The work will deliver organised and programmatic access to curated patient phenotype information. Beneficiaries are the NHS and patients, academic community, GeL researchers, clinical GeCIPs, industry and commercial partners.

**Commercial exploitation.** *(Where relevant to your GeCIP) Genomics England has a very explicit intellectual property policy. We and other funders need to know if the proposed research likely to generate commercially exploitable results. Do you have commercial partners in place?*

The research undertaken as part of this proposal will facilitate the broader commercial exploitation of GeL by adding value and expanding the scope of existing and future data. We will leverage existing and forge new relationships with industry partners from the pharmaceutical industry, EHR vendors and regulatory organizations in order to ensure the development of industry-accredited methods and software that can be subsequently commercial exploited.

**References.** *Provide key references related to the research you set out.*

## Detailed research plan

| Full proposal (total max 1500 words per subdomain) | |
|---|---|
| **Title**<br>*(max 150 characters)* | **Farr EHR GeCIP 4: Lifelong phenotyping**<br>**Developing high-resolution longitudinal phenotypes through data linkages** |

In order to drive NHS transformation, GeL has to deliver patient benefit in terms of improved clinical outcomes. The current data models for the majority of GeL phenotypes are defined based on a *a priori*-selected set of data elements which are being provided by the GMC's through a mixture of approaches including case report forms. These data elements are of limited scope as they capture cross-sectional data that is generated during secondary care interactions within the healthcare system. There is a multitude of phenotypically diverse and longitudinal data for patients that span primary and secondary care and non-health data sources that are of high interest to GeL and are currently not systematically captured or extracted.

These rich data sources can be used to construct longitudinal disease phenotypes by establishing the transition between disease states from onset to progression capturing all episodes of care (from initial presentation in primary care to diagnosis and treatment in secondary care). High-resolution longitudinal phenotypes can be utilized for hypothesis driven and hypothesis free observational research across all clinical GECiPs. Furthermore, the linkage with national coded primary and secondary care electronic health records will enable scalable and cost-effective long-term outcomes for clinical trials. Similarly, the linkage with administrative and social datasets will facilitate high-impact research across clinical domains at the intersect between health and social care by enabling the creation of non-health related phenotypes.

These EHR data however are stored across diverse sources and in different formats and require a substantial amount of processing before they can be integrated with existing GeL data models and statistically analyzed. The **aims** of this sub-domain are

(i) To develop and evaluate novel computational methods for linking, harmonizing and integrating deeper health and non-health phenotypic data for GeL participants across hospitals, national EHR data sources and administrative datasets in a standards-driven manner.

(ii) To deliver standardized approaches to the measurement of short and long term patient outcomes through health and social care information systems.

a) Data linkages

We will seek to maximize the overlap between individual patient records in different care settings spanning primary, secondary, tertiary and social care through a set of novel data linkages. We will build, test and evaluate technical solutions for enabling the linkage between multiple sources including the handling of sensitive identifiers, anonymization and the secure transport of data between data sources and the GeL environment. In Oxford we have built the infrastructure for accessing, importing, cleaning, processing and presenting HES, cancer and death data for 100,000 genomes project (based on our UKB experience,). The first set of HES data arrived at the GEL data centre in February 2016. We will create and evaluate methods for ascertaining, validating and phenotyping clinical outcomes for observational and interventional studies using electronic health records. We will develop a standardized approach for performing and evaluating data linkages between GeL participants and **national data sources** including Hospital Episode Statistics (held by

the HSCIC), Cancer registration data (held by PHE) and cause-specific mortality data (held by the ONS). We will work with the Clinical Practice Research Datalink (CPRD, a joint DH/MHRA venture), primary care information system providers (e.g. EMIS, TPP, Vision) and regional practice networks around GMCs to link and extract the primary care EHR for GeL participants. Working in close collaboration with the LHS subdomain (Delaney), we will seek to replicate best practices and approaches developed in other major initiatives such as TRANSFoRm and increase the interoperability of our approach by following relevant standards (e.g. CDISC) where possible.

We will re-use and expand the successful infrastructure created by the NIHR Health Informatics Collaborative and EHR4CR in order to increase the depth and breadth of the hospital data that gets recorded. We will work with relevant hospital information systems providers (e.g. Cerner, EPIC) to build and deploy sustainable middleware technical infrastructure to facilitate the on-going extraction of data from hospitals. We will create and evaluate computational methods for extracting, harmonizing and integrating **deeper phenotypic data from hospitals** across multiple modalities (e.g. coded pathology data, unstructured text, laboratory values, imaging). We will work with the HSCIC to extract hospital prescribing and mental health data. Working with colleagues in computer science, we will explore and systematically evaluate computational methods for performing natural language processing tasks on clinical text for extracting clinically meaningful markers (and the complex semantic relationships between them) from raw data and integrating them with existing and future GeL disease data models.

We will seek to develop a common approach for linking with **national and procedure disease registries** (e.g. cancer, cardiovascular disease registries in NICOR, National Radiotherapy Dataset, Systemic Anti-Cancer Therapy Dataset) by providing a set of standardized methods and an information governance roadmap.

Working with the Administrative Data Research Centre in England (ADRC-E) and the Administrative Data Research Centre Network, we will seek to expand our data linkage approaches to other **administrative and social care data** including environmental (weather, pollution, sensor), financial, geospatial, and build environment data.

<u>b) Data curation</u>

We will develop and evolve the **EHR research community** by fundamentally shifting the cultural landscape, providing a ‘go-to’ resource of information, tools and knowledge exchange and promoting sharing and transparency. We will establish awareness and engagement through user groups and scientific meetings across the health data science communities. The resource will be implicitly promoted by all research output that gets generated based on it across collaborators, nationally and internationally.

We will enrich the add value to the linked datasets by developing and testing standards-driven data cleaning methods and develop an **EHR phenotyping toolkit** of innovative computational methods for collaboratively generating, curating, sharing and validating EHR-derived machine-readable phenotypes. The toolkit will contain validated approaches and script implementations of common problems and operations encountered in these sources i.e. deduplication of simultaneous admissions or dimensionality/temporal reduction of multiple diagnostic codes in HES, assessing and quantifying data quality on a practice level in primary care EHR, approaches for dealing with missing data and linking records at a family level. Aligned with work in the LHS subdomain, we will evaluate different models (e.g. OHDSI OMOP) and approaches in order to standardize and harmonize data sources and associated phenotypes using a common data model. We will critically appraise metadata standards (e.g. Data Documentation Initiative, SDMX) order to identify the optimal manner in which metadata on datasets, medical ontologies and

controlled clinical terminologies can be managed, displayed and queried by the wider scientific community. We will create a resource that will both support the harvesting and integration of metadata from external sources and direct curation by researchers.

The main risk associated with the work in this proposal is the ever changing landscape of legal and information governance requirements with regards to patient sensitive identifiers which are required for linkage. We will systematically explore and validate the best approach for performing linkages between sources including the use of trusted third party services, HSCIC accredited safe havens, GMC safe havens and pseudonymization approaches in order to mitigate the risk of not being able to access the required identifiers. Patient confidentiality is of critical importance and we will seek to develop privacy-aware approaches to ensure its upheaval.

**Collaborations including with other GeCIPs.**

We will leverage our extensive network of on-going cross-disciplinary collaborations in order to undertake the research outlined in this proposal. We will work across all GeCIPs and more widely within the Farr network, EU projects and with the USA and Japan to develop exemplars, test, iterate and evaluate systems based on these basic information concepts.

Academic: Farr national network of institutions and partners across England, Scotland and Wales, Francis Crick Institute, Alan Turing Institute, Oxford Big Data Institute, UCL Institute of Big Data, QMUL, LSHTM, UCL CTU, eMedLab and other MRC-funded Medical Bioinformatics centres, Francis Crick Institute, King's, Administrative Research Data Network across England, Scotland, Wales and Norther Ireland, UK Biobank, CTSU

Research initiatives: EMIF, EuroRec, EHR4CR, CDISC, TRANSFoRM, Global Alliance for Genomics and Health, eMERGE

Industry: ABPI, Cerner, EMIS, TPP

Healthcare: AHSC's distributed across academic partners mentioned

Patient: Genomics England Patient and Public Participation Network

Public sector: HSCIC, ONS, PHE

## Training

We will create and deliver short courses across all levels (e.g. undergraduate, postgraduate and continuous professional development) on electronic health record and administrative data linkages. We will seek to develop modules in existing or emerging postgraduate degrees across all partners and create online delivery modules for popular online platforms such as Coursera.

Our training programme will have a strong emphasis on providing the crucial bridging between scientific disciplines and domains, particularly between clinical practitioners, health informaticians/computer scientists and bioinformaticians. Our training and educational programme will be outward looking and inclusive and we will seek to collaborate with relevant partners, organisations, industry and doctoral training programmes and the NHS in order to deliver our education objectives – examples include but are not limited to: ELIXIR UK, Health Education England MSc in Genomic Medicine,

Modernizing Scientific Careers Clinical Bioinformatics Programme.

The developed material will provide attendees with a solid understanding of the fundamental principles of data linkage and relevant methodologies (e.g. probabilistic vs. deterministic) and with hands-on experience through a set of computer practical sessions using contemporary and realistic training datasets.

**People and track record**

Our sub-domain will build on the outstanding strengths and expertise across its members. As demonstrated by our attached CV's, our subdomain includes several experiences UK academics whose expertise spans: health informatics, interoperability and health data standards, biomedical knowledge representation, machine learning, genomics, data linkage, data quality assessment, EHR linkages, medical software engineering, clinical decision support and privacy-aware software platforms. Furthermore, subdomain members are active in several relevant high impact research initiatives and institutions: Farr Institute, eMedLab, UK Biobank, ADRN, Oxford Big Data Institute, CALIBER, IMAGINE ID, EHR4CR, TRANSFoRm. We will seek to actively collaborate with the best national and international research groups.

**Clinical interpretation.**

We will deliver this working in collaboration with the other clinical GECiPs.

**Beneficiaries**

The beneficiaries of this subdomain include but are not limited to the NHS, patients, national and international academic community, GeL researchers, the clinical community, other clinical and cross-cutting methodological GeCIPs.

**Commercial exploitation.**

The research undertaken as part of this proposal will facilitate the broader commercial exploitation of GeL by adding value and expanding the scope of existing and future data. We will leverage existing and forge new relationships with industry partners from the pharmaceutical industry, EHR vendors and regulatory organizations in order to ensure the development of industry-accredited methods and software that can be subsequently commercial exploited.

**References.**

Kohane, I. Using electronic health records to drive discovery in disease genomics, Nat Rev Genet 2011; 12:417-428.

Jensen P, Jensen L, Brunak S. Mining electronic health records: towards better research applications and clinical care. Nat Rev Genet 2012;13(6):395-405.

Pathak J, Wang J, Kashyap S, Basford M, Li R, Masys D, Chute C. Mapping clinical phenotype data elements to standardized metadata repositories and controlled terminologies: the eMERGE Network. JAMIA 2011;18(4):376-86.

Gottesman O, Kuivaniemi H, Tromp G, Faucett A, Li R, Manolio T, Sanderson S, Kannry J, Zinberg R, Basford M, Brilliant M, Carey D, Chisholm R, Chute C, Connolly J, Crosslin D, Denny J, Gallego C, Haines J, Hakonarson H, Harley J, Jarvik G, Kohane I, Kullo I, Larson E, McCarty C, Ritchie M, Roden D, Smith M, Böttinger E, Williams M. The Electronic Medical Records and Genomics (eMERGE) Network: past, present, and future. Genetics in medicine 2013;15(10):761-71.

Rasmussen L, Thompson W, Pacheco J, Kho A, Carrell D, Pathak J, Peissig P, Tromp G, Denny J, Starren J. Design patterns for the development of electronic health record-driven phenotype extraction algorithms. Journal of biomedical informatics 2014;51:280-86.

Springate D, Kontopantelis E, Ashcroft D, Olier I, Parisi R, Chamapiwa E, Reeves D. ClinicalCodes: An Online Clinical Codes Repository to Improve the Validity and Reproducibility of Research Using Electronic Medical Records. PLoS ONE 2014;9(6):e99825.

Rapsomaniki E, Timmis A, George J, Pujades-Rodriguez M, Shah A, Denaxas S, White I, Caulfield M, Deanfield J, Smeeth L, Williams B, Hingorani A, Hemingway H. Blood pressure and incidence of twelve cardiovascular diseases: lifetime risks, healthy life-years lost, and age-specific associations in 1·25 million people. Lancet 2014;383(9932):1899-911.

Shah AD, Langenberg C, Rapsomaniki E, Denaxas S, Pujades-Rodriguez M, Gale CP, Deanfield J, Smeeth L, Timmis A, Hemingway H. Type 2 diabetes and incidence of cardiovascular diseases: a cohort study in 1.9 million people. The lancet Diabetes & endocrinology 2015;3(2):105-13.

## Data requirements

**Data scope.** *Describe the groups of participants on whom you require data and the form in which you plan to analyse the data (e.g. phenotype data, filtered variant lists, VCF, BAM). Where participants fall outside the disorders within your GeCIP domain, please confirm whether you have agreement from the relevant GeCIP domain. (max 200 words)*

Our sub-domain will primarily deal with phenotypic data from national electronic health record sources, hospitals, clinical registries and other administrative datasets that we will link GeL with. In the process of creating high-resolution longitudinal phenotypes, we will work in close collaboration with clinical GeCIPs when access to genotypic data is required.

**Data analysis plans.** *Describe the approaches you will use for analysis. (max 300 words)*

We will use a mixture of supervised (i.e. support vector machines) and unsupervised (e.g. k-means clustering) and evaluate the best approaches for linking and phenotyping the different datasets.

We will use international metadata standards (e.g. HL7 FHIR, DDI, SDMX, OMOP) and controlled clinical terminologies (e.g. SNOMED-CT) to harmonize data and establish a common data model compatible with clinical GECIPS, other subdomains, regulatory requirements for trials and NHS ICT requirements.

**Key phenotype data.** *Describe the key classes of phenotype data required for your proposed analyses to allow prioritisation and optimisation of collection of these. (max 200 words)*

Clinical Practice Research Datalink and other primary care EHR data
NICOR registries
Cancer registries

| |
|---|
| Hospital Episode Statistics + mental health + hospital prescribing<br>Office of National Statistics mortality data<br>Secondary care data from hospitals<br>PROMS, wearables, sensors |
| **Alignment and calling requirements.** *Please refer to the attached file (Bioinformatics for 100,000 genomes.pptx) for the existing Genomics England analysis pipeline and indicate whether your requirements differ providing explanation. (max 300 words)*<br><br>Not applicable to this subdomain. |
| **Tool requirements and import.** *Describe any specific tools you require within the data centre with particular emphasis on those which are additional to those we will provide (see attached excel file List_of_Embassy_apps.xlsx of the planned standard tools). If these are new tools you must discuss these with us. (max 200 words)*<br><br>Programming tools: Python (with Pandas and numpy and pytoolz), Perl + local CPAN repository<br>RDBMS: MySQL or Postgres<br>Workflow tools: cwltool or bpipe or nextflow, galaxy |
| **Data import.** *Describe the data sets you would require within the analysis environment and may therefore need to be imported or accessible within the secure data environment. (max 200 words)*<br><br>See above. |
| **Computing resource requirements.** *Describe any analyses that would place high demand on computing resources and specific storage or processing implications. (max 200 words)*<br><br>Storage and processing requirements depend on dataset availability and granularity. |

| Full proposal (total max 1500 words per subdomain) | |
|---|---|
| **Title**<br>*(max 150 characters)* | **Farr EHR GeCIP 5**<br>**Tools for EHR enabled trials in precision medicine** |

**Importance**

Delivering precision medicine for the NHS requires a suite of new, genomically informed, randomised trials. Generating this evidence demands innovation in the efficient use of NHS data in order to optimise each stage of trial design, conduct and implementation) as well as specific challenges and opportunities afforded by the availability of genomic sequence data. This is crucial to address questions such as testing dose, timing and combinations of existing drugs, repurposing hypotheses, licensed drug adjuncts to existing treatments (which may arise from genetic discoveries) as well as new innovative treatments. It will also be crucial to streamline and better power new trials for patients with rare disease and ethnic groups under-represented in trials.

The overarching **aim** of this Farr GeCIP sub-domain is to intersect OMICs technologies and EHRs to enable precision medicine trials within the NHS.

The **specific objectives** of this sub-domain are

(i) To drive EHR phenotypic augmentation of OMICs Data for evaluating trial feasibility, point of care randomisation and follow up for trial safety and efficacy outcomes applicable to any trial design

(ii) To generate an intelligent computational platform that builds on OMICs data to optimise trial design and execution across a wide range of trial designs.

**Research plans**

**Workstream 1: Delivering EHR phenotypic Augmentation of OMICs Data for evaluating trial feasibility, point of care randomisation and follow up for safety and efficacy outcomes**

Central to more efficient trial feasibility, design and conduct is rapid access to good quality phenotypic data through the EHR. The specific focus here is on compliance with trial relevant standards (GCP, CDISC) this is a complementary but distinct focus from subdomain 4 Longitudinal Phenotyping (where the emphasis is on endotype discovery) and 6 Learning Health Systems (where the emphasis is on system relevant outcomes). Answering clinically relevant omics informed trial questions requires fine-grained information that was not available at the time of capturing the patient genetic data, such as clinical and radiological assessments, laboratory results, pathology diagnoses and staging investigations, in- and out-patient medications and other medical treatments. In addition, the effect of an ever changing environment in the patient and the effect of constantly evolving clinical practices, medicines and treatments, might invalidate past phenotypical characterisations, and therefore are important factors for understanding genetic associations. Understanding clinical and imaging manifestations of pharmacogenetics over time, whether therapeutic or adverse events, requires an integrated approach to record all confounders, especially those are rarely captured during studies but are in EHRs, semantically linked them to OMICs data, and interrogate them near real-time.

The research in this workstream will address these challenges by supporting the integration of numerous and diverse data sources, including EHR, imaging, pathology, laboratories and medications, ancilliary clinical systems and of course omics, following a principled, standard-based approach, resulting in a scalable and reusable federated dataset across all GMCs and associated trustworthy interrogation services, that will support all stages in the conduct of clinical trials. The research will build on the results of several international initiatives (EHR4CR, TRANSFoRm, i2b2, eMERGE, RPGEH, BioVU) to formalise models of use and meaning of the patient information, making use of bed-side standards (HL7v3, EN13606, DICOM) , clinical study information models (CDISC CDASH, BRIDG), and relevant terminologies (SNOMED-CT, LOINC, MeDRA, ICD) to ensure the semantically correct phenotypic augmentation of the OMIC data. The research will also address the study of high-performance and scalable computational methods to allow this process to perform in a cost-effective fashion, avoiding ad-hoc and error-prone manual processes. The research will also explore models of data-provenance to ensure that all phenotypic augmentation of the OMIC data is traceable and verifiable.

The research in this workstream will build on the EHR-phenotype augmented and federated OMICs dataset to develop methods and tools for optimising the design, recruitment, execution and efficacy and adverse event monitoring of trials. We will build on our experience in project such as EHR4CR and TRANSFoRm, that have already developed the blueprint of the key computational services for these stages, and further advance their results by developing bioinformatics and computational methods, to automatically identify scenarios that will boost the performance and cost-effectiveness of each stage in the trial life cycle. Capturing clinically cared populations in EHR systems, will allow to optimise trials for rare diseases that require specialised treatment in secondary and tertiary care centres, and also for under-represented ethnic populations in current cancer studies, who have normal to high representation across healthcare centres.

One approach to delivering this trials platform is to evaluate an ‘EHR mirror’ of a conventionally designed and conducted trial, suchas the DARWIN I trial in lung cancer (part of TRACERx).

**Workstream 2: Optimising EHR-driven Omics Trial Life Cycle**

This workstream will research the impact on trial design, recruitment and execution of several optimisation strategies, specifically:

a) **Exonic variants**: Whole Genome Sequencing (WGS) provides a unique opportunity to discover exonic variants that associated with amino-acid changes and 3D structural composition of the encoded protein. This knowledge is more likely to affect the binding of large-molecules (i.e. Monoclonal Antibodies) as opposed to small molecules, thus such genomic knowledge would useful to generate trials of genomic-guided therapy for MAbs. To realise this strategy, the bioinformatics methods will semantically augment exonic-variants with structural consequences (CATH database) together with bioinformatics knowledge on MAbs (CHembL). The interventions to be derived from such knowledge can be dose-finding studies of the same Mab, and comparisons with alternative treatment if available.

b) **Genetic determinants of biomarkers of therapeutic efficacy**: the genomic knowledge can emerge from genotypic-risk scores of common variants identified as GWAs-hits affecting concentrations of a Biomarker of Therapeutic Efficacy (BTE) such as VEGF levels for Anti-VEGF therapies, or from rare mutations discovered through GE with extreme effects on BTE. Borrowing from novel analytical techniques that used knowledge of the entire WGS (as opposed to "hits") to predict levels of BTE could be trial as the tool to guide therapy. The intervention to be derived from such knowledge would be either dose-finding studies to increase the efficacy/safety ratio or the use of alternative treatments when available. The same strategy and scenarios for clinical trials can be applied to known targets for small-molecules, but focusing on the targets, for instance VEGF genes that encodes of VEGF instead of all other genes that may affect VEGF levels.

c) **Genetic determinants on ADME (Absorption, Distribution, Metabolism and Excretion):** this type of trial can be based on known common variants, as well as new rare variants (to be discovered by GE) involved in the ADME genes already established. This knowledge together with the PK/PD for each drug-compound can serve to identify ideal scenarios for this type of trial are the presence of high-penetrance variants, plus drug-compounds with narrow-margins of efficacy/safety and the presence of serious adverse events due to drug-compound and absences of therapeutic alternatives.

d) **Genetic determinants of serious adverse events:** this is perhaps one of the most advanced in use of genomics in clinical care and the research in this stream should lead to to generation of new variants and also to identify genetic determinants on the known side-effects.

The novel research in workstream will study the implementation of trial design and feasibility services that allow the computational modelling and execution of these strategies, in a federated environment, distributed across all GMC. The trial design platform will allow the exploration of designs such as parallel arms, cross-over and N-1, and different randomisation levels (individual or cluster) in terms of the type of intervention, absence or existence of carry over effects and the nature of the primary outcome.

Building on the results of the phenotypic agumentation of the OMICs, the research in this workstream will enhance existing trial execution and adverse monitoring services by studying secure and trustworthy computational methods for distributed and federated investigation of surrogate end-points such as imaging and biomarkers, that are routinely generated as well as research ones embedded in clinical pathway, at pre-specified points in time and when given conditions are met.

**Collaborations including with other GeCIPs.**

There will be a concerted international approach with respect to the use of standards in collaboration with the European Institute of Electronic Health Records, EHR vendors, CDISC, EN13606 , HL7 and DICOM. We will collaborate across all GeCIP and more widely with similar international initiatives ( EHR4CR, TRANSFoRm, i2b2, eMERGE, RPGEH, BioVU).

**Training.**

We will build on the success of the Health Informatics MSc and also Data Science MSc at UCL and partner's institutions, expanding these programmes through research-based teaching that will ensure the dissemination of our results and training of future generations of leaders in this area.

**People and track record.**

Our sub-domain will build on the outstanding strengths and expertise across its members. As demonstrated by our attached CV's, our subdomain includes several experiences UK academics whose expertise spans: health informatics, interoperability and health data standards, biomedical knowledge representation, machine learning, genomics, data linkage, data quality assessment, EHR linkages, medical software engineering, clinical decision support and privacy-aware software platforms. Furthermore, subdomain members are active in several relevant high impact research initiatives and institutions: Farr Institute, eMedLab, UK Biobank, ADRN, Oxford Big Data Institute, EHR4CR, TRANSFoRm.

**Clinical interpretation.**

We will deliver this working in collaboration with the other clinical GECiPs.

**Beneficiaries.**

The project will deliver methods and a software platform for the support of trials that will benefit Patients, NHS, Pharmaceutical companies, EHR Vendors and Academic centres.

**Commercial exploitation.**

The project aim to develop a blueprint of the computational platform for the benefit of the wider community that would seek proper IP management and exploitation. The software platform services implemented will require the development of a business model.

**References.**

1. GoDARTS and UKPDS Diabetes Pharmacogenetics Study Group *et al.* Common variants near ATM are associated with glycemic response to metformin in type 2 diabetes. *Nat. Genet.* **43,** 117–120 (2011)

2. Holmes, M. V., Perel, P., Shah, T., Hingorani, A. D. & Casas, J. P. CYP2C19 genotype, clopidogrel metabolism, platelet function, and cardiovascular events: a systematic review and meta-analysis. *JAMA* **306,** 2704–2714 (2011).

3. Bergmeijer, T. O. *et al.* CYP2C19 genotype-guided antiplatelet therapy in ST-segment elevation myocardial infarction patients-Rationale and design of the Patient Outcome after primary PCI (POPular) Genetics study. *Am. Heart J.* **168,** 16–22.e1 (2014).

4. Harakalova, M. *et al.* Dominant missense mutations in ABCC9 cause Cantú syndrome.

*Nat. Genet.* **44,** 793–796 (2012).

5. Holmes, M. V. *et al.* Fulfilling the promise of personalized medicine? Systematic review and field synopsis of pharmacogenetic studies. *PLoS ONE* **4,** e7960 (2009).

6. Baranova, E. V., Verhoef, T. I., Asselbergs, F. W., de Boer, A. & Maitland-van der Zee, A.-H. Genotype-guided coumarin dosing: where are we now and where do we need to go next? *Expert Opin Drug Metab Toxicol* **11,** 509–522 (2015).

7. Mahmoudpour, S. H. *et al.* Pharmacogenetics of ACE inhibitor-induced angioedema and cough: a systematic review and meta-analysis. *Pharmacogenomics* **14,** 249–260 (2013).

8. Holmes, M. V., Casas, J. P. & Hingorani, A. D. Putting genomics into practice. *BMJ* **343,** d4953 (2011).

9. Donnelly, L. A. *et al.* Robust association of the LPA locus with low-density lipoprotein cholesterol lowering response to statin treatment in a meta-analysis of 30 467 individuals from both randomized control trials and observational studies and association with coronary artery disease outcome during statin treatment. *Pharmacogenet. Genomics* **23,** 518–525 (2013).

10. Baranova, E. V., Asselbergs, F. W., de Boer, A. & Maitland-van der Zee, A. H. The COAG and EU-PACT trials: what is the clinical benefit of pharmacogenetic-guided coumarin dosing during therapy initiation? *Curr. Mol. Med.* **14,** 841–848 (2014).

11. Schuemie, M. J. *et al.* Using electronic health care records for drug safety signal detection: a comparative evaluation of statistical methods. *Med Care* **50,** 890–897 (2012).

## Detailed research plan

| Full proposal (total max 1500 words per subdomain) | |
|---|---|
| **Title** | **Farr EHR GeCIP Subdomain 6: Essential informatics concepts for the Learning Health System** |

**Importance.**

For the potential benefits of GeL's sequence data to be translated into patient benefits requires a matching transformation in information systems. The Learning Health System (LHS), as defined by the US Institute of Medicine (2008, 2012) is a term that describes the formal linkage of research in routine healthcare settings and application of the knowledge created. The widespread adoption of electronic health records (EHRs) and their use in real time during the consultation enables the LHS to develop as an integrated technological system at potentially much greater scope and scale than traditional human-social systems. The LHS is a wider concept that personalized medicine alone as it has two important implications:

1. Research, 'Big Data/'omics/personalized medicine' is insufficient in isolation to create impact on patients, it requires a 'Big Knowledge' strategy to accompany it.
2. To progress at scale and without duplication, creating silos and fragmentation we require a significant step up in informatics approaches to facilitate the LHS.

We propose to methods research to develop core informatics standards, models, approaches for system integration across the EHR GeCIP subdomains. We will build on and collaborate with the very best UK, European and International groups in this area.

**Research plans.**

*Workstream 1: Improved knowledge discovery from routinely collected data*

The Achilles heel of all routine data analysis is the unknown and unmeasurable bias in data collected in totally uncontrolled ways as part of routine healthcare. This data does not reflect a complete record of care, and missingness is most certainly not at random. In addition the problems of interpretation, imprecision and overlapping clinical concepts, both within and across terminologies make interpretation of such data that does exist difficult. One approach to this is to enable collection of 'better' data and data enrichment within EHR systems. Approaches to using CDISC standards to collect data from the EHR in real time are being developed (IHE, TRANSFoRm).

Research challenges addressed:

- Informatics methods for supporting data mining and machine learning methods for extracting knowledge from routine data
- Methods for managing clinical concepts and maintenance of meaning along the data-knowledge cycle. This means moving beyond existing 'terminology services' for mapping, such as the National Library of Medicines UMLS system to a deeper use of ontologies to track and control key clinical concepts.
- Evaluation of integration of detailed clinical element capture within the EHR, fitting into clinical workflow at the point of care and paying due diligence to human-computer interaction.

*Workstream 2: Supporting clinical trials and alternative methods of establishing knowledge:*

The CPRD TrialVIZ tool, and the Farsite tool, developed by NW eHealth enable practice

data to be queried and suitable subjects invited for recruitment to clinical trials. The NIHR CRN have also developed a set of standard Read codes for the coding of research activities in health records. These are based on a standard model of research in practices. This allows CPMS study numbers to bound against specific read codes for things like 'consented into study', enabling systems to track study accrual and extract follow up data by individual study. Previously individual study teams were requesting a plethora of 'study specific' read codes from HSCIC. Prevalent case recruitment and live flagging of eligible subjects in a consultation for clinical trials has been piloted using the 'Live Eligible Patient Identification System (LEPIS), which continues to be used for a study of Influenza surveillance (FLUCATS). IMI-EHR4CR has developed a system for identifying clinical trial subjects in specialist databases (oncology etc). TRANSFoRm has developed a platform based on CDISC standards to integrate clinical trial workflow into the Primary Care EHR, along with data collection forms and patient related outcome measures.

Research Challenges addressed:

- Comparison of methods for clinical trial functions semi-automated within the EHR?
- Comparison of advanced non-randomised methods (propensity scores, virtual cohorts, rapid machine learning) with randomised controlled trials?
- How can the design and conduct of randomised trials be simplified using better informatics systems (such as Trials Within Cohorts designs)
- Evaluation the potential for learning from rapid iteration of changing practice provide an alternative means of reaching 'evidence'?

*Workstream 3: Supporting knowledge translation.*

TRANSFoRm has been designed as a model-based 'knowledge aware' infrastructure that uses core concepts and services to support integration of data and knowledge into clinical activity. An evaluation of an integrated decision support system (DSS) for diagnosis in primary care, in conjunction with the EHR system provider InPS (Vision3) has just been completed. Results of the functional prototype with simulated patients are similar to the earlier computer-based simulation showing an 8% improvement in diagnostic accuracy. As regards knowledge translation for clinical interventions and prognosis, Prof Charles Friedman, University of Michigan is developing standards for 'Digital Knowledge Objects', small computable evidence statements. This work could be used to develop further NICE tools for guidelines deployment in EHR systems to provide a rapid translational route for clinical evidence.

Research Challenges addressed:

- Extending diagnostic models and data to wider coverage of common conditions.
- Automating NICE treatment guidelines into the EHRs providing decision support while also highlighting the uncertainties in evidence base for particular patient groups
- Development and evaluation of international standards in the representation of diagnostic data, including genomic prediction rules.
- Integration of genomic and other personalized models into diagnostic DSS
- Evaluation of DSS for therapeutic interventions following clinical and cognitive workflows, and investigating means of optimising the content, format and timing of delivery.
- Building and evaluating a framework for maintaining the knowledge base.

**Collaborations including with other GeCIPs***.*

There will be a common approach to international standards in four areas:

- Clinical (phenotype) model of meaning: Biomedical ontologies, Detailed clinical

models (ISO13606, HL7 DCM.
- Research model of meaning: BRIDG and PCROM
- Knowledge representation for action: Diagnostic ontology, ATHENA and OpenCDS
- Clinical data content carrier: HL7 Fast Healthcare Interoperability Resources
- Research data content carrier: CDISC Operational Data Model

We will work across all GeCIPs and more widely within the Farr network, EU projects and with the USA and Japan to develop exemplars, test, iterate and evaluate systems based on these basic information concepts.

**Training**.
We will work with the Farr institute in establishing appropriate PhD studentships in the are, and in using the Masters in Health Informatics courses at partner's institutions.

**People and track record.**
All investigators are part of the Farr Network and have extensive experience of collaborating together on clinical informatics projects. The group includes Clinicians, Public Health, Computer Science, Trialists and Epidemiologists. Delaney and Curcin led the FP7 TRANSFoRm project, and have also collaborated on the IMI EHR4CR project (Denaxas, D'acosta). Buchan, Van Staa, and Peek have allied programmes in Farr Manchester (HeRC). Hemingway, Denaxas and Ray similarly at Farr London, Conley, Farr Scotland and Lyons Farr Wales. Ainsworth leads the Connected Health Cities Hub. Davies is CTO Genomic England.

**Clinical interpretation**
We will deliver this working in conjunction with other clinical domains.

**Beneficiaries.**
Beneficiaries include patients who will benefit from safer more effective healthcare, society via more efficient translation of new knowledge and industry by simpler means of adopting standards. Development of a LHS is not simply a process of using ICT tools to extract data, but requires a coordinated and standards based approach to defining, creating, curating and re-using a variety of information artefacts from individual data elements to data syntheses and statements of knowledge and clinical recommendations. The LHS represents a system for manipulating these artefacts in order to promote safe, efficient and effective healthcare at individual, organisational and population levels. A LHS requires a large degree of automaticity, in order for a cyber-human-social system to operate at scale it is necessary to transfer significant human activities to computational systems. This process is well advanced in many industries, but barely begun in healthcare.

**Commercial exploitation.**
This proposal aims to build on world class research in systems integration, data standards, knowledge representation, clinical modelling and data provenance to create a high-functioning LHS based on UK research and the UK National Health System. This will place the UK at the forefront of the LHS with significant advantages for UK science, healthcare and industry. Innovation in healthcare takes place in a context of public-private co-operation. Development of additional functionality in EHR systems requires a sustainable business model for the data analyses and the maintenance of the knowledge created.

Challenges:
- Simple interface designs to allow rapid adoption of new technologies
- Development and maintenance of international standards

**References.**

McGinnis, J. M. Evidence-based medicine - engineering the Learning Healthcare System. Stud Heal. Technol Inf. 153, 145–157 (2010).

(IOM) Medicine, I. of. Digital Infrastructure for the Learning Health System: The Foundation for Continuous Improvement in Health and Health Care. Found. Contin. Improv. Heal. Heal. Care Work. Ser. Summ. (2011).

Brendan C. Delaney, Vasa Curcin, Anna Andreasson, Theodoros N. Arvanitis, Hilde Bastiaens, Derek Corrigan, Jean-Francois Ethier, Olga Kostopoulou, Wolfgang Kuchinke, Mark McGilchrist, Paul van Royen, and Peter Wagner. Translational Medicine and Patient Safety in Europe: TRANSFoRm—Architecture for the Learning Health System in Europe. Biomed Research International 2015

Ethier, Jean-Francois; Dameron, Olivier; Curcin, Vasa; McGilchrist, Mark M; Verheij, Robert A; Arvanitis, Theodoros N; Taweel, Adel; Delaney, Brendan C; Burgun, Anita. A unified structural/terminological interoperability framework based on LexEVS: application to TRANSFoRm. Journal of the American Medical Informatics Association. 2013;20;207-216

AR Tate, N Beloff, B Al-Radwan, J Wickson, S Puri, T Williams, T Van Staa, A Bleach. Exploiting the potential of large databases of electronic health records for research using rapid search algorithms and an intuitive query interface. J Am Med Inform Assoc. 2014 Mar-Apr;21(2):292-8. doi: 10.1136

Sarah Thew, Gary Leeming, John Ainsworth, Martin Gibson and Iain Buchan. FARSITE: evaluation of an automated trial feasibility assessment and recruitment tool. From Clinical Trials Methodology Conference 2011. Trials Journal, Published: 13 December 2011

Speedie SM, Taweel A, Sim I, Arvanitis TA, Delaney BC, Peterson KA. The Primary Care Research Object Model (PCROM): A Computable Information Model for Practice-Based Primary Care Research. Journal of The American Medical Informatics Association. 2008;15:661-70; doi:10.1197/jamia.M2745

Van Staa T-P, Goldacre B, Gulliford M, Cassell J, Pirmohamed M, Taweel A, Delaney BC, Smeeth L. Pragmatic randomised trials using routine electronic health records: putting them to the test. BMJ 2012;344:e55 doi: 10.1136

Van Staa T-P, Dyson L, McCann G, Padmanabhan S, Belatri R, Goldacre B, Cassell J, Pirmohamed M, Torgerson D, Ronaldson S, Adamson J, Taweel A, Delaney B, Mahmoud M, Baracaia S, Round T, Fox R, Hunter T, Gulliford M, Smeeth L. Pragmatic point-of-care randomised trials using routinely collected electronic records: qualitative and quantitative research of opportunities and challenges in the implementation of two exemplar trials. Health technology Assessment 2014:18;43.

J.-F. Ethier, V. Curcin, A. Barton, M. M. McGilchrist, H. Bastiaens, A. Andreasson, J. Rossiter, L. Zhao, T. N. Arvanitis, A. Taweel, B. C. Delaney, A. Burgun. Clinical Data Integration Model: Core Interoperability Ontology for Research Using Primary Care Data. Methods Inf Med. 2014 Jun 18;53(4).

Olga Kostopoulou, Andrea Rosen, Round Thomas, Ellen Wright, Brendan Delaney, and Abdel Douiri. Early diagnostic suggestions improve accuracy of GPs: an experimental study. BJGP 2014: 65 (630), e49-e54