Journal of Biomedical Informatics 90 (2019) 103090

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Federated electronic health records research technology to support clinical trial protocol optimization: Evidence from EHR4CR and the InSite platform

Brecht Claerhout[a], Dipak Kalra[b], Christina Mueller[a], Gurparkash Singh[c], Nadir Ammour[d], Laura Meloni[a], Juuso Blomster[e], Mark Hopley[f], George Kafatos[g], Almenia Garvey[h], Peter Kuhn[i,j], Martine Lewi[k], Bart Vannieuwenhuyse[k], Benoît Marchal[l], Ketan Patel[m], Christoph Schindler[n], Mats Sundgren[o,*]

[a] Custodix N.V., Gent, Belgium
[b] European Institute for Innovation through Health Data (i-HD), Brussels, Belgium
[c] Janssen Research & Development LLC, Fremont, United States
[d] Sanofi-Aventis R&D, Chilly-Mazarin, France
[e] Turku University Hospital, Turku, Finland
[f] Boehringer-Ingelheim, Ingelheim, Germany
[g] Amgen, Uxbridge, UK
[h] ICON PLC, Paris, France
[i] Comprehensive Cancer Center Ulm, University Hospital Ulm, Ulm, Germany
[j] Institute of Medical Systems Biology, Ulm University, Ulm, Germany
[k] Janssen Pharmaceuticals, Beerse, Belgium
[l] F. Hoffmann-La Roche, Basel, Switzerland
[m] AstraZeneca, Cambridge, UK
[n] Medizinische Hochschule Hannover, Hannover, Germany
[o] AstraZeneca, Gothenburg, Sweden

ARTICLE INFO

Keywords:
Clinical research
Electronic health record
Clinical trials
Pharmaceutical industry
Learning health systems

ABSTRACT

Objective: To determine if inclusion/exclusion (I/E) criteria of clinical trial protocols can be represented as structured queries and executed using a secure federated research platform (InSite) on hospital electronic health records (EHR) systems, to estimate the number of potentially eligible patients.

Methods: Twenty-three clinical trial protocols completed during 2011–2017 across diverse disease areas were analyzed to construct queries that were executed with InSite using EHR records from 24 European hospitals containing records of > 14 million patients. The number of patients matching I/E criteria of each protocol was estimated.

Results: All protocols could be formalized to some extent into a medical coding system (e.g. ICD-10CM, ATC, LOINC, SNOMED) and mapped to local hospital coding systems. The median number of I/E criteria of protocols tested was 29 (range: 14–47). A median of 55% (range 38–89%) of I/E criteria in each protocol could be transformed into a computable format. The median number of eligible patients identified was 26 per hospital site (range: 1–134).

Conclusion: Clinical trial I/E eligibility criteria can be structured computationally and executed as queries on EHR systems to estimate the patient recruitment pool at each site.

The results further suggest that an increase in structured coded information in EHRs would increase the number of I/E criteria that could be evaluated. Additional work is needed on broader deployment of federated platforms such as InSite.

1. Background and significance

Clinical research is a time consuming, labor intensive and costly endeavor. However, some of these challenges, in particular, those pertaining to clinical trial protocol optimization and patient selection may be mitigated by the judicious re-use of data contained in Electronic Health Records (EHR). In this study we have evaluated the ability to convert patient selection criteria from clinical trial protocols into

* Corresponding author at: Biometrics and Information Sciences, AstraZeneca, Gothenburg SE-431 83, Mölndal, Sweden.
E-mail address: mats.sundgren@astrazeneca.com (M. Sundgren).

https://doi.org/10.1016/j.jbi.2018.12.004
Received 16 May 2018; Received in revised form 30 October 2018; Accepted 21 December 2018
Available online 02 January 2019
1532-0464/

electronic queries that can be executed on EHR systems, for which we used a federated research platform (InSite). Our results demonstrate that platforms, such as InSite, provide a near to real-time method of querying hospital EHR systems, allowing for a rapid, cost-effective and iterative approach to determining the number and distribution of eligible patients in a network of hospitals and as such have the potential to optimize clinical study protocols and accelerate recruitment.

2. Introduction

Clinical research is a time consuming, labor intensive and costly endeavor [1]. Specific issues related to these bottlenecks include difficulties evaluating patient populations, identifying suitable patients for enrolment, optimizing protocols, manual and redundant data entry, reliability of data sources, and the difficulty in detecting and reporting infrequent adverse events [2]. By some estimates, it costs around $2.5 billion (USD 2013) to bring a new drug to market [3]. The high cost is due, in part, to the need to conduct large clinical trials that provide definitive evidence of efficacy and safety. It is also attributable to a requirement to fulfill regulatory requirements and a need to generate value-based evidence, which may require "real-world" studies to evaluate comparative effectiveness, safety and cost-effectiveness. In addition, there are process challenges and bottlenecks that impede the conduct of clinical trials such as sub-optimal study design, slow and lengthy patient recruitment, site selection, protocol optimization that add to time and resource requirements.

Some of these challenges, such as those pertaining to protocol optimization and patient selection may be mitigated by the judicious reuse of data contained in Electronic Health Records (EHRs). The increasing adoption of EHRs in Europe and beyond represents a vast, rich, and highly relevant health data source that has the potential to improve clinical trial implementation [4]. Potential applications for these data include testing clinical trial feasibility using computable representations of the criteria [5–8], enhancing patient selection, improving clinical trial execution and facilitating adverse event surveillance.

Analysis of high quality EHR data can help healthcare organizations become better Learning Health Systems [9–11], as envisioned by the Electronic Health Record for Clinical Research (EHR4CR) project (2011–2015). This European Innovative Medicines Initiative (IMI) project was funded to develop a trustworthy, secure and acceptable infrastructure platform to analyze EHR information across a network of European hospitals [12,13]. Its overall charter included enabling clinical trial protocol developers to more accurately understand the clinical profiles and distribution of candidate trial subjects, thereby optimizing the clinical trial protocols and enhancing recruitment. It was also intended to help hospitals identify candidate patients after a study has been initiated, but before enrolment is complete, which allows investigators to invite these individuals to participate in the trial. A cost-benefit assessment demonstrated that this approach to designing clinical trial protocols would be cost effective when scaled up across Europe [11,14]. The EHR4CR architecture and its approach to protection of patient privacy has been described elsewhere [10]. Following the completion of the EHR4CR project in 2016, a commercial version of the platform (InSite™, Custodix N.V.) has been developed and is now being deployed across Europe.

Two key concerns with the successful scale-up of federated research platforms were expressed during the EHR4CR project. These included, (1) whether the patient selection criteria in a clinical trial protocol can be expressed as structured queries and executed on hospital EHRs; and, (2) whether the data in hospital EHR systems are of sufficient quality to enable representative inferences to be made about the number of patients that meet the patient selection criteria. The present study aimed to investigate these concerns by examining inclusion and exclusion (I/E) criteria of 23 completed trials across diverse therapeutic areas that were sponsored by seven pharmaceutical companies (Amgen, AstraZeneca, Bayer, Boehringer-Ingelheim, F-Hoffman La Roche, Janssen and Sanofi) in order to determine the proportion of I/E criteria that could be represented in a computable format and the ability to query hospital EHRs to accurately estimate the number of potentially eligible patients.

3. Methods

3.1. Research environment

During the end phase of the EHR4CR project in 2015, six of the original European Federation of Pharmaceutical Industries and Associations (EFPIA) partners who participated in the EHR4CR project (Amgen, AstraZeneca, Bayer, Janssen, Roche, Sanofi) plus Boehringer-Ingelheim and ICON Plc, agreed to collaborate with the goal of establishing a hospital network connected to the InSite™ platform. This consortium, operating as the "InSite Champion Programme", provided the necessary pre-competitive environment to permit the companies to share otherwise commercially confidential information within the research team. This study was performed on 24 hospital EHR systems connected to the InSite network and encompassed data from more than 14 million patients living in 5 European countries (Belgium, Finland, Germany, Italy, Spain) (Table 1).

The InSite platform and tools were used to execute queries on this EHR network with a computerized version of each study protocol focusing selectively on structured data elements within the hospital EHRs. Specific ethical approval was not required for this study.

3.2. Selection of study protocols and validation criteria

Every pharmaceutical company participating in the InSite Champion Programme (Amgen, AstraZeneca, Bayer, Boehringer-Ingelheim, Janssen, Roche and Sanofi) contributed two to five clinical trial protocols for this study. Protocols were selected on the basis of five criteria: (1) the protocol must be in the public domain; (2) the trial described in the protocol must have been completed; and (3) the study must have been completed within the previous five years; (4) preferably the trial had been conducted at ≥ 1 hospital in the InSite hospital network; and (5) trials were selected such that diverse therapeutic areas were represented. Application of these criteria resulted in selection of 23 protocols that had collectively recruited patients with cardiovascular diseases, asthma, diabetes mellitus, fractures, migraine headaches, cancer, schizophrenia, rheumatologic conditions, nasal polyps and chronic hepatitis C infection (Table 2).

Criteria assessed in this analysis included: (1) the percentage of patient selection criteria that could be formalized; (2) the average number of patients per site that met these criteria; and (3) reasons why some criteria could not be formalized. The execution scheme is shown in Fig. 1.

3.3. Overview of the InSite platform technology

This analysis was conducted as part of the InSite Champion Program, and used the InSite Platform that connected 24 hospitals across Europe. InSite is a "Software as a Service" (SaaS) platform that allows researchers to interact with hospital-based EHRs. Participating clinical sites connect through locally installed InSite software (InSite

Table 1
Data used in this study.

Country	Sites	Approximate number of patients
Belgium	9	4.2×10^6
Finland	1	1.3×10^6
Germany	2	3.5×10^6
Italy	1	0.7×10^6
Spain	11	5.3×10^6

Table 2
Overview of clinical trial protocols included in the study.

#	Sponsor	Study registration number and description	Phase	Disease area	Study date	Ref.
1	Amgen	**NCT00907296** Interventional study to compare AMG 785 with placebo on radiographic healing time of fresh tibial diaphysis fractures fixed with an intramedullary (IM) nail	**II**	**Fractures**	09/2009 – 09/2012	None
2	Amgen	**FOURIER (NCT01764633)** Interventional study of an additional LDL-C lowering treatment to assess tolerability and cardiovascular risk reduction in subjects with clinically evident cardiovascular disease and treated for dyslipidemia	**III**	**Cardiovascular**	02/2013 – 11/2016	Giugliano et al. [15]
3	Amgen	**LILAC (NCT01901146)** Interventional study to compare the effectiveness and safety of ABP 980 and trastuzumab in women with early breast cancer	**III**	**Oncology**	04/2013 – 01/2017	None
4	Amgen	**ARISE (NCT02483585)** Interventional study of AMG 334 on monthly migraine days in subjects with episodic migraine	**III**	**Migraine**	07/2015 – 03/2017	None
5	AZ	**ICON6 (NCT00532194)** Interventional study to assess the safety and efficacy of cediranib in combination with standard chemotherapy in patients who have relapsed with ovarian, fallopian tube or epithelial cancer, after first line treatment	**III**	**Oncology**	07/2007 – 12/2016	Ledermann et al. [16]
6	AZ	**NCT01031680** Interventional study in patients with type 2 diabetes, to assess whether dapagliflozin lowers blood glucose, body weight and blood pressure, when added to existing medications and how it compares with their usual treatment without added dapagliflozin	**III**	**Diabetes mellitus**	02/2010 – 12/2012	Cefalu et al. [17]
7	AZ	**SIROCCO (NCT01928771)** Interventional study of benralizumab on asthma exacerbations in patients who remain uncontrolled on high doses of ICS-LABA	**III**	**Asthma**	09/2013 – 04/2016	Bleecker et al. [18]
8	AZ	**SELECT-1 (NCT01933932)** Interventional study to assess the efficacy of selumetinib in combination with docetaxel vs. placebo in combination with docetaxel in patients with locally advance or metastatic NSCLC that harbor mutations of KRAS	**III**	**Oncology**	09/2013 – 12/2018	Janne et al. [19]
9	AZ	**SOCRATES (NCT01994720)** Interventional study to compare ticagrelor with aspirin in the prevention of major vascular events	**III**	**Cardiovascular**	01/2014 – 03/2016	Easton et al.[20]
10	Bayer	**ARTS-DN (NCT01874431)** Interventional study to assess the safety and efficacy of a new drug, given orally at different doses in patients with type 2 diabetes and diabetic nephropathy	**II**	**Diabetes mellitus**	06/2013 – 08/2014	Ruilope et al. [21]
11	Bayer	**ARTS-HF (NCT01807221)** Interventional study to assess a new drug, given orally at different doses, to evaluate whether it is safe and can help the well-being of patients with worsening chronic heart failure and either type II diabetes with or without chronic kidney disease or kidney disease alone	**IIb**	**Cardiovascular**	06/2013 – 12/2014	Pitt et al. [22]
12	BI	**IMPULSIS (NCT01335464)** Interventional study to investigate and confirm the efficacy and safety of BIBF 1120 at a high dose in treating patients with IPF, compared with placebo	**III**	**Idiopathic pulmonary fibrosis**	04/2011 – 10/2013	Richeldi et al. [23]
13	BI	**REDUAL-PCI (NCT02164864)** Interventional study to compare a dual antithrombotic therapy (DAT) regimen with a triple antithrombotic therapy (TAT) combination in patients with atrial fibrillation that undergo PCI with stenting	**III**	**Cardiovascular**	07/2014 – 06/2017	Cannon et al. [24]
14	BI	**RE-CIRCUIT (NCT02348723)** Interventional study to assess the safety of an uninterrupted dabigatran etexilate periprocedural anticoagulant regimen compared to an uninterrupted warfarin regimen in patients with non-valvular atrial fibrillation (NVAF)	**IV**	**Cardiovascular**	04/2015 – 11/2016	Calkins et al. [25]
15	Janssen	**NCT01559272** Interventional study to assess the pharmacokinetics, safety, and tolerability of a paliperidone palmitate 3-month formulation in patients with schizophrenia	**I**	**Schizophrenia**	02/2008 – 05/2014	Magnusson et al. [26]
16	Janssen	**CANTATA-MP (NCT01106690)** Interventional study to evaluate the efficacy and safety of 2 different doses of canagliflozin compared with placebo in patients with type 2 diabetes mellitus who are receiving treatment with metformin and pioglitazone and have inadequate glycemic (blood sugar) control	**III**	**Diabetes mellitus**	06/2010 – 07/2012	Usiskin et al. [27]
17	Janssen	**NCT01515423** Interventional study to demonstrate that a paliperidone palmitate 3-month formulation (PP3M) is as effective as the paliperidone palmitate 1-month formulation (PP1M) in the treatment of patients with schizophrenia who have been stabilized on PP1M	**III**	**Schizophrenia**	04/2012 – 05/2015	None
18	Janssen	**NCT01809327** Interventional study to assess the effectiveness of the co-administration of canagliflozin and metformin extended release (XR) compared with canagliflozin alone, and metformin XR alone in patients with type 2 diabetes mellitus with inadequate control despite treatment with diet and exercise	**III**	**Diabetes mellitus**	06/2013 – 12/2014	Rosenstock et al. [28]
19	Roche	**NCT00750880** Interventional study to investigate the safety, tolerability and efficacy of tocilizumab monotherapy, or combination therapy with non-biologic disease modifying antirheumatic drugs (DMARDs), in patients with severe active RA	**IIIb**	**Rheumatism**	09/2008 – 07/2011	None
20	Roche	**NCT01307397** Interventional study to evaluate the safety and efficacy of RO5185426 in patients with BRAF V600 mutation-positive, surgically incurable and unresectable stage IIIC or IV metastatic melanoma	**III**	**Oncology**	03/2011 – 02/2016	Larkin et al. [29]
21	Roche	**ANNAPURNA (NCT01628094)** Interventional study to evaluate the safety, efficacy and tolerability of a combination treatment in patients with chronic hepatitis C genotype 1	**II**	**Hepatitis C**	06/2012 – 11/2013	Jensen et al. [30]
22	Sanofi	**ALIGN (**NCT01061723) Interventional study to evaluate the efficacy of sarilumab in participants with ankylosing spondylitis	**II**	**Rheumatism**	02/2010 – 06/2011	Sieper et al. [31]
23	Sanofi	NCT01920893 Interventional study to evaluate the efficacy of dupilumab in the treatment of bilateral nasal polyps	**II**	**Nasal polyps**	07/2013 – 11/2014	Bachert et al. [32]

AZ = AstraZeneca; BI = Boehringer-Ingelheim NCT number refers to the registration number listed in ClinicalTrials.gov.

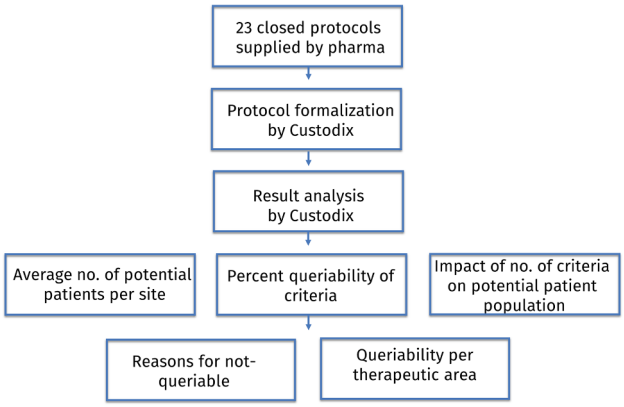

Fig. 1. Execution scheme for closed protocol validation.

Local Install). This software is integrated with a local Clinical Data Warehouse (CDW) containing patient data that has been uploaded from the EHR (Fig. 2).

The InSite Local Install contains the components necessary to support the platform operations (mainly a medical query engine) and includes a number of tools for use by the clinical site: a patient recruitment application, a data exploration tool, a basic analytics tool, etc.

A fundamental requirement for the acceptable operation of the system is that patient data never leaves the hospitals. In all usage scenarios data processing is done under the control of the hospital and central platform end-users (researchers) only get access to anonymous aggregated results. As such, InSite enabled data-protection compliant re-use of EHR data, validated by privacy experts, patient organizations and ethical review boards during the EHR4CR project (www.ehr4cr.eu).

The technology executed the following scenarios:

A. Study Design Scenario and Study Placement Scenario

For both of these scenarios, researchers used the central platform to send formalized trial criteria (medical queries) to the different sites on the InSite network. The associated queries were executed locally on the InSite local software. The platform received the number of patients from each connected site that met the formalized trial selection criteria. The number of patients was initially aggregated by site and then by country.

B. Patient Recruitment Scenario

Using the platform, researchers sent a formalized (final) protocol to the different sites on InSite. The local InSite software used this formalization to automatically scan for patients and assist the local site investigators with patient selection. Only recruitment progress was reported back to the central platform.

3.4. Formalization of criteria and execution of queries

A comprehensive list of I/E criteria were extracted from each protocol and each criterion was matched to terms contained in standard medical coding systems (diagnosis: ICD-10CM, procedures: ICD-PCS, medication: ATC, laboratory: LOINC, clinical findings: SNOMED and anatomic pathology/oncology ICD-O-3) as far as possible. Next, hospital data ware matched (mapped) during the connection process to these standard terms [13] to produce a uniform way of querying. For example, if a protocol specified that eligible patients were at least 18 years of age, had a diagnosis of type 2 diabetes mellitus and were taking medication recommended by the American Heart Association (AHA), but not insulin, these criteria were formalized as SNOMED code (age) 424144002 and the age filter was adapted to > 18 years. In addition, we used ICD-10CM code E11 (type 2 diabetes) and ATC code A10B (blood glucose lowering drugs except insulin). I/E criteria were separated within the same query by applying an AND/OR or a NOT function, eventually creating a computed copy of the selection criteria that corresponded as closely as possible to those specified in the protocol (work set). Temporal relations where added where these arose in the criteria. In the query representation formalism used, time windows can be adapted down to days, months or years and modifiers such as “at least”, “at most”, “first”, “any” or “last” refine the search. The following options are available: (1) A simple time window: an event happens at least or at most a certain time period ago; e.g. The first diagnosis of type 2 diabetes happened at least 5 years ago. (2) A specific order of events within a specified time window: a given event(s) happen before or after another (set of) event(s); e.g. The first diagnosis of being overweight with a BMI > 30 happened at most 2 years before the first diagnosis of type 2 diabetes mellitus. (3) A given event happens multiple times; e.g. At least 3 diagnoses of asthma exacerbation separated by at least 2 months (4) events happen at the same time during a visit or an observation; A diagnosis of streptococcal infection and streptococcal sepsis happened during the same visit (hospital stay), or, the diagnosis of lung cancer and stage T2 was determined during the same observation (consultation).

Criteria that drew on the, probably subjective, investigator’s opinion could not be considered. This means that it was not possible to incorporate the nuances such as “significant”, or “major” within the formalization process. Imprecise temporal expressions were modelled based on context and medical knowledge e.g. “recent” can be modelled by using time constraints. A recent antibiotic treatment could be set to “at most 6 weeks ago”, a recent operation to “at most 6 month ago”. This is, however, highly depend on the therapeutic area and the specific criterion within the protocol and thus varies with each protocol. We recognize that there is no authoritative standard for formalizing such terms. These decisions were made through discussion with clinical members of the project, but ideally their use should be discouraged within protocol criteria (see discussion).

For this study, free-text reports were not analyzed, although it is recognized that this would be desirable in the future, as discussed later.

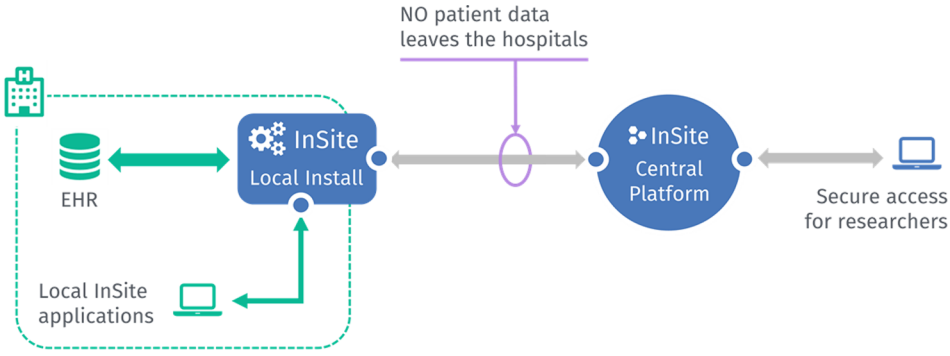

Fig. 2. Standard InSite data flow. Selected data (both structured and unstructured) from electronic health records (EHRs) are loaded into a Clinical Data Warehouse (CDW) via an ETL process (extract – transform-load) on which medical queries are executed by the InSite local installation upon requests of sponsors who formalized their eligibility criteria on the InSite central platform. Results are returned to the central platform as aggregated results.

A limitation of the method is that the criteria were formalized by one expert, which provided consistency but did not allow for a control (to assess inter-rater agreement), which is also discussed later.

Once formalized, the criteria were expressed as queries that could be executed via the InSite Platform across the network of hospitals included in this study, reported in Section 3.1 above. The InSite eligibility criteria query language is a newly developed query language with similar characteristics as ERGO[1] supporting various types of constraints (Boolean logic, temporal, encounter-based, value ranges, medical concept inference, …). It can be composed with a graphical user interface tool with drag-and-drop features and has xml and json representations for communication and persistence purposes. The InSite query engine transforms these query constructs into readily executable database queries and evaluates all specified constraints, yielding in the end a set of eligible patients or their relevant characteristics (e.g. breakdown of number of patients matching each of the specified criteria in the query). All of the formalized criteria (i.e. clinically coded, temporally modelled etc.) could be represented as InSite query expressions.

Each protocol was executed against the available data on the reference date that matched the closing date for the study as provided on clinicaltrials.gov (for example: Nov. 2016 for NCT02348723 sponsored by Boehringer Ingelheim).

4. Results

Patient selection criteria from the 23 completed clinical trial protocols covering diverse therapeutic areas (Table 2) were formalized on the InSite platform and used to query the InSite hospital EHR network.

The number of criteria that could be formalized on the InSite platform was compared with the number of selection criteria specified in each protocol (Table 3).

A median of 55% of the selection criteria in each study could be formalized (range 38–89%) (Supplemental Figure). The study with the lowest formalization rate (38%, Table 3, number 20) recruited patients with metastatic melanoma and required information on tumor staging and genetic testing that is not routinely incorporated in a structured format into hospital EHR systems for different reasons. For example, the results of some genetic tests are out-sourced and thus only found in free-text reports. (As stated earlier, we did not analyze free-text information in this study.) In contrast, the study with the highest formalization rate (89%, Table 3, number 6) recruited patients with diabetes mellitus. Patient selection criteria for patients with diabetes mellitus are well defined and routinely held in a structured format within hospital EHR systems.

Criteria that could not be formalized were analyzed and grouped according to: (1) willingness of patients to sign informed consent forms or to abstain from a particular behavior; (2) scoring systems that are not routinely contained in hospital EHR systems; (3) criteria that are applied at screening or at a particular subsequent visit, (for example, a specific laboratory test that is not routine), and would eliminate too many potential candidates when running a feasibility screen; (4) criteria that cannot be formalized (for example, conditional events such as use of medications that are exclusion criteria when used for another reason, treatment response, measurements taken at home, toxicity grades, use of terms such as, “symptomatic”, “may or may not“, “treatment naïve”, “more than one medication,” but not specifying which ones, “must have recovered from all side effects”, “uncontrolled”, “patient self-report/e-diary, etc.); (5) investigator's opinion; (6) clinical findings (or other data) not held by hospital EHRs, such as, contraindications, allergies to medications or dosage information; (7) temporary conditions for which randomization could be delayed, for example, no antibiotic treatment; (8) planned or likely future events, such as planned surgeries, occurrence of side effects or development of a disease; (9) participation in another trial, or trial-related guidance such as “not a relative or employee of the investigator/institution, or local authority approval for pharmacogenomic research, unable to provide own consent, fluent in local language, having a stable residence or a support person, etc. (Supplementary Table 1).

As shown in Fig. 3, 39% of criteria that could not be formalized depended on criteria that are not routinely recorded in hospital EHR systems including the investigator’s opinion (17%), the willingness of the patient (16%), and/or planned or likely events (6%). Another 18% of items that could not be queried relate to participation in another trial or trial related questions. Missing data elements account for 20% (unavailable scoring systems and clinical findings not provided), whereas information that could be formalized, but could only be confirmed at screening or during a subsequent visit (6%), or after resolution of a temporary condition (3%) account for 9% of criteria that could not be queried. The inability of the InSite software to execute a query was a limiting factor in the remaining 14% of non-formalizable criteria, which amounted to only 6% of the total criteria contained in a given protocol.

Next, we analyzed the number of eligible patients per-hospital that InSite identified and compared this with the actual global patient enrolment and the average global per-hospital recruitment figures, assuming that the actual figures are per site rather than per health care organization (due to different contractual arrangements for each trial with each institution). The median number of potential eligible patients identified was 26 per hospital (range: 1–134).

InSite permits a user to quickly obtain an accurate estimate of the potential pool of eligible patients in the network. Formalization of a complete protocol takes a couple of hours and, after sending a query, data are received within minutes, which shows the potential for InSite to significantly reduce the time taken for feasibility assessments. An independent cost-benefit analysis has already shown that a solution like InSite can reduce the average cycle time of a phase II-III clinical trial by a “conservative” 20% [10]. The total number of patients enrolled in a completed trial and the number of study sites were taken from clinicaltrials.gov or from the primary publications of clinical trials. It should be noted that the InSite network only contains European hospitals (Table 2).

5. Discussion

In this study, we have examined recent protocols from 23 Phase 2–4 international randomized clinical trials in diverse therapeutic areas and evaluated the potential to query the structured data contained in the EHR systems of 24 hospitals in five European countries using a novel federated EHR research platform (InSite). The main objective was to evaluate the capability of the InSite platform to computably assess protocol feasibility prior to study initiation, and to provide an understanding of the potential pool of patients at sites that have this technology in place.

We first investigated whether selected clinical trial protocols could be represented in a computable format. The results demonstrate that the 23 protocols could be formalized to varying degrees. Overall a median of 55% of the patient selection criteria could be formalized using standard medical coding systems and used to communicate with the EHR systems in hospitals connected to the InSite platform. Some clinical concepts were expressed rather vaguely and could not be mapped to a terminology system. In addition, criteria such as “investigator’s opinion”, “willingness to adhere to the regimen” or “not participating in another trial” could not be formalized. For example, 89% of the patient selection criteria used to recruit patients with type 2 diabetes mellitus by Cefalu et al. [13] (NCT01031680) could be formalized because the I/E criteria corresponded to precise clinical concepts, value constraints and established disease parameters that could be matched to data elements contained in hospital EHRs. In contrast,

[1] http://rctbank.ucsf.edu/home/ergo

Table 3
Average number of enrolled patients per site per trial vs. average number of patients estimated by InSite per site.

No.	Study name/NCT number	No. of inclusion/ exclusion criteria	Percent of criteria formalized on InSite	Average no. of patients per site (Total no. of patients enrolled per trial/no. of sites)[a]	Average no. of patients/site on InSite (no. of patients/no. of sites)	Comment[b]
1	NCT00907296	3/9	**50%**	4 (402/103)	**6 (127/23)**	
2	**FOURIER** (NCT01764633)	6/25	**61%**	21 (27564/1287)	**118 (2711/23)**	The study had a broad set of inclusion criteria
3	**LILAC** (NCT01901146)	13/17	**47%**	8 (827/99)	**107 (213/2)**	Only two hospitals (Specialized in oncology) provided sufficient data on tumor staging for this protocol at this reference, yet it may represent an overestimation due to one site not recording comorbidities and thus not allow exclusion of patients with certain comorbidities
4	**ARISE** (NCT02483585)	8/26	**41%**	8 (577/76)	**14 (301/21)**	2 sites did not provide structured medication data, a key inclusion criterion in this case
5	**ICON6** (NCT00532194)	11/17	**46%**	8 (486/63)	**26 (634/24)**	
6	NCT01031680	9/9	**89%**	6 (922/141)	**60 (1251/21)**	2 hospitals have no medication data available in structured format which was a key inclusion criterion, hence only results from 10 centers
7	**SIROCCO** (NCT01928771)	14/20	**56%**	10 (2681/281)	**1 (31/2)**	1 hospital has no medication data available in structured format which was a key inclusion criterion, hence only results from 11 centers. Results furthermore rely on medication data that is, in the majority of cases, prescribed by a GP (asthma inhalers) and thus not available in the hospital records, leading to an underestimation of patients
8	**SELECT-1** (NCT01933932)	13/20	**58%**	3 (510/197)	**7 (156/21)**	3 hospitals have no medication data available in structured format, 1 center no laboratory data, at this reference date, which was a key inclusion criterion, hence only results from 9 centers
9	**SOCRATES** (NCT01994720)	5/16	**48%**	25 (13307/528)	**17 (380/22)**	1 hospital has no medication data available at this reference date
10	**ARTS-DN** (NCT01874431)	8/19	**70%**	5 (823/168)	**54 (1242/23)**	
11	**ARTS-HF** (NCT01807221)	12/18	**53%**	5 (1058/203)	**6 (69/11)**	3 HCO's (10 hospitals) have no laboratory information and 2 hospitals no medication data available for this reference date
12	**IMPULSIS** (NCT01335464)	7/18	**44%**	5 (515/98)	**22 (494/22)**	1 hospital has an unexplainable outlier in terms of number of patients and was not included in this query because the cause was not yet established
13	**REDUAL-PCI** (NCT02164864)	4/23	**59%**	6 (2725/4210)	**46 (1059/23)**	
14	**RE-CIRCUIT** (NCT02348723)	6/21	**52%**	8 (678/87)	**6 (134/23)**	
15	NCT01559272	11/21	**56%**	9 (9328/36)	**80 (1833/23)**	No scoring data available yet for schizophrenia. Result may be an overestimation
16	**CANTATA-MP** (NCT01106690)	9/31	**55%**	4 (344/83)	**134 (1738/13)**	3 HCO's (10 hospitals) did not provide laboratory data going back this far in time (ref. date Dec. 2012 and before)
17	NCT01515423	15/22	**46%**	8 (1429/177)	**98 (2254/23)**	No scoring data available yet for schizophrenia. Result may be an overestimation
18	NCT01809327	11/29	**58%**	9 (1186/137)	**100 (2106/21)**	2 hospitals have no lab data available for the reference date of the study
19	NCT00750880	8/39	**70%**	6 (1681/282)	**19 (365/19)**	5 hospitals do not have medication data available at this reference date
20	NCT01307397	11/13	**38%**	11 (3219/280)	**26 (625/24)**	
21	**ANNAPURNA** (NCT01628094)	5/13	**83%**	4 (110/31)	**107 (2454/23)**	Many hospitals do not test for HCV genotype 1, hence we had to run a general query on hepatitis patients, which means the result is an overestimation
22	**ALIGN** (NCT01061723)	3/11	**57%**	4 (301/80)	**6 (128/23)**	
23	NCT01920893	4/24	**54%**	4 (60/14)	**7 (162/23)**	

[a] It is assumed that global enrolment figures are per hospital site and not per health care organization (containing several sites) due to a diversity of contractual arrangements. This allows for a comparison with the number of hospitals in the InSite network.

[b] The comments explain why the number of patients is overestimated. If it was not an oncology protocol, the oncology center was not included in the queries, thus by default there were 23 available hospitals.

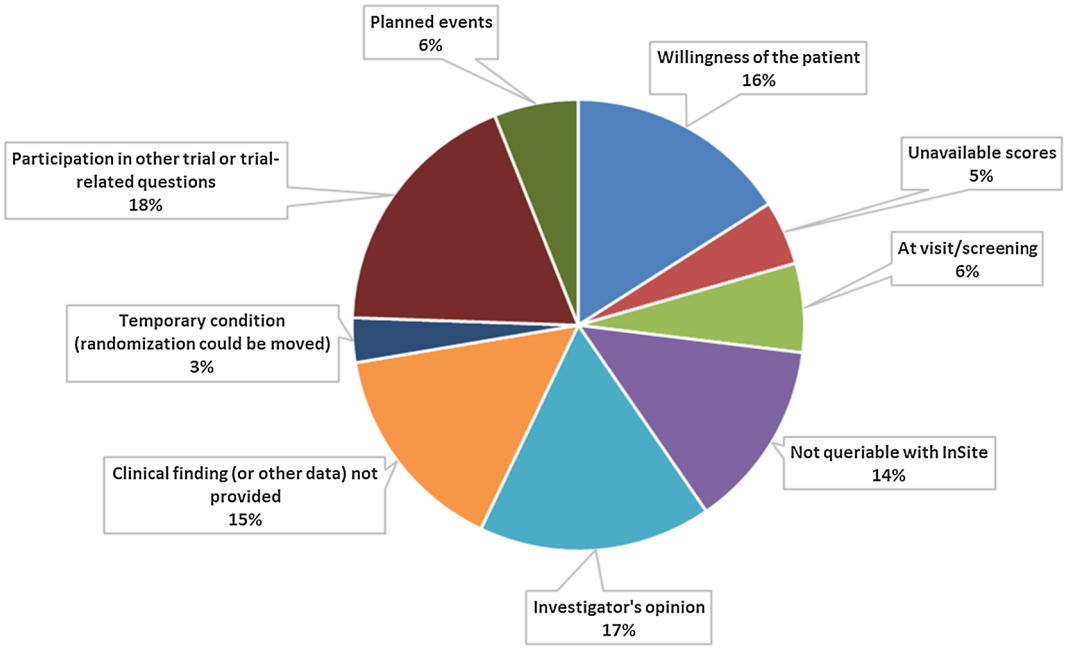

Fig. 3. Reasons why criteria could not be queried.

only 38% of patient selection criteria in the protocol for a study recruiting patients with schizophrenia (NCT01307397) could be formalized, largely because many criteria included disease-specific scoring systems and outcome measures that are currently only available in free-text fields within the EHRs.

As a second step, we investigated whether sufficient data were contained in hospital EHR systems at participating hospitals to identify candidate subjects for enrolment in the 23 protocols that we evaluated. The results showed that a substantial amount of structured data is contained in these EHR systems that correspond to the patient selection criteria in the protocols. On average, approximately 26 potential patients (median) per site could be identified per study by InSite. As might be expected, there was variation in the number of patients identified per hospital for each protocol. There are several factors that contribute to this variation. For example, not all the computable criteria corresponded to the data structure in the EHR system in every hospital, because the EHR systems vary between hospitals. Additionally, the incidence and prevalence of a given disease varies geographically, by hospital and by disease. Thus, it is easier to identify patients with a disease with a high prevalence such as diabetes mellitus than with a low prevalence such as a particular form and stage of cancer.

Nevertheless, it is very clear that a solution like InSite can rapidly identify potential patients suitable for recruitment. For example, using the I/E criteria from the study protocols used in this study we were able to identify a median of 26 suitable individuals from an average of 596,000 patients at each site (14.3 million distributed across 24 sites), which amounts to 0.004% of the overall hospital population. These numbers serve as a guide to the potential for this system to identify potential patients.

Our findings suggest that on average more than half of the patient selection criteria in a clinical trial protocol can be mapped to a structured data scheme with value constraints to allow for viable queries to hospital EHR systems. Given that some of the criteria that could be mapped are of a subjective nature, and rely more on investigator opinion than on objective patient characteristics, the proportion of relevant criteria could be higher. In addition, several criteria will most likely never be queriable with such a system, such as “Involvement in the planning and/or conduct of the study.” After removing these types of criteria from the overall count, the percentage of queriable items increases by an average of 10–15%. This solution permits a user to quickly obtain an accurate estimate of the potential pool of eligible patients in the network.

5.1. Strengths and limitations

We examined a diversity of protocols provided by different companies and pertaining to a wide range of therapeutic areas; nonetheless our sample is not comprehensive. As demonstrated by the contrast between clinical trials in the domains of diabetes and schizophrenia, patient selection criteria in certain therapeutic areas lend themselves well to translation into computable criteria. The ability to accurately estimate patient numbers also depends upon the richness of the data within hospital EHRs. We would advocate that the scope of our work should be expanded to include a broader range of therapeutic areas to provide better insight into which therapeutic areas are well-suited to computable feasibility assessments. Our work suggests that a future study should also include information extracted from free text, which we will examine in future research. As a corollary to this, we believe that there is a need to educate protocol authors about the ways in which patient selection criteria can, and cannot, be expressed computationally, and frequently occurring data items in hospital EHR systems, so that future protocols might be better constructed in the first place for computable feasibility assessments. Terms requiring subjective investigator interpretation should be discouraged.

The work of structuring each protocol and mapping patient selection criteria to suitable value constraints was largely undertaken by one expert, who had previously done this task. The involvement of more than one independent terminologist may have resulted in more robust validity. We recognize this as a limitation; however, we believe the work was done sufficiently well for a proof of concept study.

The project was conducted exclusively within a network of European hospitals because they were members of the InSite Champion Programme network. The hospitals in this network, many of which are academic centers, are active in clinical research, have good quality EHR systems and have made efforts to establish and maintain comprehensive patient databases. It is possible that hospitals with a less sophisticated or less well-used EHR system might not yield the same precision in estimating patient numbers. An additional point to note is that clinical documentation within EHRs is still evolving towards the greater capture of structured and coded data, as opposed to free text, and that

CRITERION	SWEDEN	FINLAND	GERMANY	SPAIN
All patients	200	571	[192, 212]	200
More or equals to 18 Years old (inclusion)	200	571	[192, 212]	200
last administration of **BLOOD GLUCOSE LOWERING DRUGS, EXCL. INSULINS** first diagnosis of **Type 2 diabetes mellitus** (inclusion)	181	[542, 550]	[165, 197]	181
last diagnosis of **Heart failure** happened at least 2 times separated by at most 6 Months in between (inclusion)	0	153	0	0
last administration of **metformin** last administration of **INSULINS AND ANALOGUES** (exclusion)	0	17	0	0
last recording of **Hemoglobin A1c/Hemoglobin.total in Blood** is Less than 7 (included) % (exclusion)	0	17	0	0

Fig. 4. Example of an attrition table result provided by the InSite platform on a small demonstration data.

efforts are being made to improve EHR data quality. It is inevitable that poor data quality will result in inaccurate estimates of patient numbers, which we find acts as an incentive for hospitals to assess and improve that quality. The availability of data corresponding to eligibility criteria should therefore increase in the coming years.

Our study was limited to the use of only one technical platform and not any other emerging ecosystems in the field [33]. However, this ensured that the queries were executed in a consistent manner within a single large network of hospital EHRs on which it was permitted to execute these confidential protocols.

A major strength was the ability to represent the majority of patient selection criteria into relevant clinical concepts using standard medical terminology. We found that all of the computable data items derived from the protocols could be mapped to the data within hospital EHRs. It is therefore our belief that the choice of technical platform was appropriate for this study and did not impose a significant limitation on the results.

Our investigation was restricted to studies that had been completed and is thus retrospective. Ideally, we would have conducted our research as an electronic query-based recruitment tool at a series of hospitals in parallel to traditional recruitment processes at other hospital sites. We are in the process of planning a prospective study that will provide a contemporaneous comparison of patient recruitment.

5.2. Further research and opportunities

This study has several important implications. First, more value could be derived from technology such as InSite if patient selection criteria were aligned with the terminology and data structure that exists in EHR systems. Our study shows that, at present, selection criteria do not consider what is contained in existing EHR systems, but, rather, must be translated to take advantage of this rich data source. The results demonstrate that approximately 45% of patient selection criteria in recent protocols cannot be translated into a format that matches the data structure of EHR systems. Criteria that contain potentially ambiguous phrases and those that rely on subjective elements such as “investigator opinion” are problematic in this regard. Greater understanding of hospital coding practices, coupled with a desire to incorporate this knowledge into clinical trial protocols in development would likely enhance the ability to translate patient selection criteria into successful data queries by InSite. This in turn might increase the total number of patients enrolled per hospital or globally by trial. This learning is important for hospitals that aspire to become better Learning Health Systems, and may incentivize them to enhance the quality of terminology-based data elements in their own EHR systems. This study has identified data domains (e.g. medications…) of interest and very commonly used in protocol eligibility criteria, that could be prioritized by hospitals for data enrichments initiatives. This should also enhance their capacity to use this platform for internal quality and safety queries, in addition to making a larger contribution to academic and industry-sponsored clinical trials.

Our results provide evidence of the potential value of EHR data to facilitate and probably expedite the recruitment of patients into clinical trials. More evidence and further data are needed to establish the value of this technology across all therapeutic areas and across diverse clinical settings. This technology could also prove useful in evaluating the feasibility of a given protocol. At present, it takes an average of 3 months to assemble the disparate data required to assess the feasibility of a study protocol [14], whereas distributed queries such as those performed by InSite could be used in almost real time to evaluate the impact of specific patient selection criteria on patient recruitment. For example, the technology provides an opportunity to identify criteria that have the greatest impact on restricting the pool of eligible patients, and that can then be represented as an attrition table (Fig. 4).

6. Conclusion

Our results suggest that, for protocols in several therapeutic areas, a sufficient number of clinical trial protocol eligibility criteria can be structured computationally and executed as distributed queries across multiple hospital EHR systems to provide a reasonable estimate of the patient recruitment pool at each site. Federated EHR research platforms, exemplified by the InSite platform, provide a near to real-time method of querying hospital EHR systems, thereby allowing a rapid and iterative approach to determining the impact of changing different selection criteria on the number and distribution of eligible patients. They offer promise as a cost-effective tool for validating and optimizing clinical study protocols, and for accelerating recruitment by hospitals. Our results suggest that, with an increasing proportion of structured and coded information in EHR systems, used against protocol eligibility criteria written as originally computable elements, the potential for cost and time savings in the conduct of clinical trials should increase. This can be further supported with a broader deployment of hospital network solutions such as InSite in Europe, and beyond.

Acknowledgements

The authors thank Rainer Thiel and Veli Stroetmann, Empirica, Bonn, Germany for help with strategy planning; Blair Jarvis ELS,

Ottawa, Canada for editing the manuscript; Bayer AG, Leverkusen, Germany for providing closed protocols for this study; and Lieven Vaneeckhaute of UZ Gent, Gent, Belgium for editing.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jbi.2018.12.004.

References

[1] B. Harper, N. Christison, Fixing the protocol feasibility process, J. Clin. Res. Best Pract. 8 (2012) 1–6.
[2] J. Mestre-Ferrandiz, J. Sussex, A. Towse, The R&D Cost of a New Medicine, Office of Health Economics, London, UK, 2012, pp. 1–86.
[3] J.A. DiMasi, H.G. Grabowski, R.W. Hansen, Innovation in the pharmaceutical industry: new estimates of R&D costs, J. Health Econ. 47 (2016) 20–33.
[4] P.B. Jensen, L.J. Jensen, S. Brunak, Mining electronic health records: towards better research applications and clinical care, Nat. Rev. Genet. 13 (2012) 395–405.
[5] C. Weng, X. Wu, Z. Luo, M.R. Boland, D. Theodoratos, S.B. Johnson, EliXR: an approach to eligibility criteria extraction and representation, JAMIA 18 (1) (2011) i116–i124, https://doi.org/10.1136/amiajnl-2011-000321.
[6] F. Köpcke, D. Lubgan, R. Fietkau, A. Scholler, C. Nau, M. Stürzl, R. Croner, H.U. Prokosch, D. Toddenroth, Evaluating predictive modeling algorithms to assess patient eligibility for clinical trials from routine data, BMC Med. Inform. Decis Mak. 13 (2013) 134, https://doi.org/10.1186/1472-6947-13-134 PMID: 24321610.
[7] Y. Ni, S. Kennebeck, J.W. Dexheimer, C.M. McAneney, H. Tang, T. Lingren, Q. Li, H. Zhai, I. Solti, Automated clinical trial eligibility prescreening: increasing the efficiency of patient identification for clinical trials in the emergency department, J. Am. Med. Inform. Assoc. 22 (1) (2015) 166–178, https://doi.org/10.1136/amiajnl-2014-00288 Epub 2014 Jul 16. PMID: 25030032.
[8] J. Doods, K. Holzapfel, M. Dugas, F. Fritz, Development of best practice principles for simplifying eligibility criteria, Stud. Health Technol. Inform. 192 (2013) 1153.
[9] C.P. Friedman, N.J. Allee, B.C. Delaney, et al., The science of learning health systems: foundations for a new journal, Learn Health Sys. 1 (2017) e10020.
[10] P. Coorevits, M. Sundgren, G.O. Klein, et al., Electronic health records: new opportunities for clinical research, J. Intern. Med. 274 (2013) 547–560.
[11] D. Dupont, A. Beresniak, A. Schmidt, et al., Assessing the financial impact of reusing electronic health records data for clinical research: results from the EHR4CR European project, J. Health Med. Informat. 7 (2016) 1–7.
[12] G. De Moor, M. Sundgren, D. Kalra, et al., Using electronic health records for clinical research: the case of the EHR4CR project, J. Biomed. Inform. 53 (2015) 162–173.
[13] R. Bache, A. Taweel, S. Miles, B.C. Delaney, An eligibility criteria query language for heterogeneous data warehouses, MethodsInf. Med. 54 (2015) 41–44.
[14] A. Beresniak, A. Schmidt, J. Proeve, et al., Cost-benefit assessment of using electronic health records data for clinical research versus current practices: contribution of the electronic health records for clinical research (EHR4CR) European project, Contemp Clin. Trials 46 (2016) 85–91.
[15] R.P. Giugliano, T.R. Pedersen, J.G. Park, et al., Clinical efficacy and safety of achieving very low LDL-cholesterol concentrations with the PCSK9 inhibitor evolocumab: a prespecified secondary analysis of the FOURIER trial, Lancet 390 (2017) 1962–1971.
[16] J.A. Ledermann, A.C. Embleton, F. Raja, et al., Cediranib in patients with relapsed platinum-sensitive ovarian cancer (ICON6): a randomised, double-blind, placebo-controlled phase 3 trial, Lancet 387 (2016) 1066–1074.
[17] W.T. Cefalu, L.A. Leiter, T.W. de Bruin, I. Gause-Nilsson, J. Sugg, S.J. Parikh, Dapagliflozin's effects on glycemia and cardiovascular risk factors in high-risk patients with type 2 diabetes: a 24-week, multicenter, randomized, double-blind, placebo-controlled study with a 28-week extension, Diabetes Care 38 (2015) 1218–1227.
[18] E.R. Bleecker, J.M. FitzGerald, P. Chanez, et al., Efficacy and safety of benralizumab for patients with severe asthma uncontrolled with high-dosage inhaled corticosteroids and long-acting beta2-agonists (SIROCCO): a randomised, multicentre, placebo-controlled phase 3 trial, Lancet 388 (2016) 2115–2127.
[19] P.A. Janne, H. Mann, D. Ghiorghiu, Study design and rationale for a randomized, placebo-controlled, double-blind study to assess the efficacy and safety of selumetinib in combination with docetaxel as second-line treatment in patients with KRAS-mutant advanced non-small cell lung cancer (SELECT-1), Clin. Lung Cancer 17 (2016) e1–e4.
[20] J.D. Easton, M. Aunes, G.W. Albers, et al., Risk for major bleeding in patients receiving ticagrelor compared with aspirin after transient ischemic attack or acute ischemic stroke in the SOCRATES study (acute stroke or transient ischemic attack treated with aspirin or ticagrelor and patient outcomes), Circulation 136 (2017) 907–916.
[21] L.M. Ruilope, R. Agarwal, J.C. Chan, et al., Rationale, design, and baseline characteristics of ARTS-DN: a randomized study to assess the safety and efficacy of finerenone in patients with type 2 diabetes mellitus and a clinical diagnosis of diabetic nephropathy, Am. J. Nephrol. 40 (2014) 572–581.
[22] B. Pitt, S.D. Anker, M. Bohm, et al., Rationale and design of MinerAlocorticoid Receptor antag nist Tolerability Study-Heart Failure (ARTS-HF): a randomized study of finerenone vs. eplerenone in patients who have worsening chronic heart failure with diabetes and/or chronic kidney disease, Eur. J. Heart Fail 17 (2015) 224–232.
[23] L. Richeldi, R.M. du Bois, G. Raghu, et al., Efficacy and safety of nintedanib in idiopathic pulmonary fibrosis, N Engl. J. Med. 370 (2014) 2071–2082.
[24] C.P. Cannon, D.L. Bhatt, J. Oldgren, et al., Dual antithrombotic therapy with dabigatran after PCI in atrial fibrillation, N Engl. J. Med. 377 (2017) 1513–1524.
[25] H. Calkins, S. Willems, E.P. Gerstenfeld, et al., Uninterrupted dabigatran versus warfarin for ablation in atrial fibrillation, N Engl. J. Med. 376 (2017) 1627–1636.
[26] M.O. Magnusson, M.N. Samtani, E.L. Plan, et al., Population pharmacokinetics of a novel once-every 3 months intramuscular formulation of paliperidone palmitate in patients with schizophrenia, Clin. Pharmacokinet. 56 (2017) 421–433.
[27] K. Usiskin, I. Kline, A. Fung, C. Mayer, G. Meininger, Safety and tolerability of canagliflozin in patients with type 2 diabetes mellitus: pooled analysis of phase 3 study results, Postgrad. Med. 126 (2014) 16–34.
[28] J. Rosenstock, L. Chuck, M. Gonzalez-Ortiz, et al., Initial combination therapy with canagliflozin plus metformin versus each component as monotherapy for drug-naive type 2 diabetes, Diabetes Care 39 (2016) 353–362.
[29] J. Larkin, M. Del Vecchio, P.A. Ascierto, et al., Vemurafenib in patients with BRAF (V600) mutated metastatic melanoma: an open-label, multicentre, safety study, Lancet Oncol. 15 (2014) 436–444.
[30] D.M. Jensen, M. Brunda, R. Elston, et al., Interferon-free regimens containing setrobuvir for patients with genotype 1 chronic hepatitis C: a randomized, multicenter study, Liver Int. 36 (2016) 505–514.
[31] J. Sieper, J. Braun, J. Kay, et al., Sarilumab for the treatment of ankylosing spondylitis: results of a Phase II, randomised, double-blind, placebo-controlled study (ALIGN), Ann. Rheum. Dis. 74 (2015) 1051–1057.
[32] C. Bachert, L. Mannent, R.M. Naclerio, et al., Effect of subcutaneous dupilumab on nasal polyp burden in patients with chronic sinusitis and nasal polyposis: a randomized clinical trial, JAMA 315 (2016) 469–479.
[33] U. Topaloglu, M.B. Palchuk, Using a federated network of real-world data to optimize clinical trials operations, JCO Clin. Cancer Inform. (2018) doi: 101200/CCI1700067.