Secondary Use and Analysis of Big Data Collected for Patient Care

Contribution from the IMIA Working Group on Data Mining and Big Data Analytics

F. J. Martin-Sanchez1, V. Aguiar-Pulido2, G. H. Lopez-Campos3, N. Peek4, L. Sacchi5

1 Weill Cornell Medicine, Department of Healthcare Policy and Research, Division of Health Informatics, New York, USA
2 Weill Cornell Medicine, Brain and Mind Research Institute, New York, USA
3 The University of Melbourne, Health & Biomedical Informatics Centre, Melbourne, Australia
4 MRC Health e-Research Centre, Division of Informatics, Imaging and Data Science, The University of Manchester, Manchester, UK
5 Department of Electrical, Computer and Biomedical Engineering, University of Pavia, Pavia, Italy

Summary

Objectives: To identify common methodological challenges and review relevant initiatives related to the re-use of patient data collected in routine clinical care, as well as to analyze the economic benefits derived from the secondary use of this data. Through the use of several examples, this article aims to provide a glimpse into the different areas of application, namely clinical research, genomic research, study of environmental factors, and population and health services research. This paper describes some of the informatics methods and Big Data resources developed in this context, such as electronic phenotyping, clinical research networks, biorepositories, screening data banks, and wide association studies. Lastly, some of the potential limitations of these approaches are discussed, focusing on confounding factors and data quality.

Methods: A series of literature searches in main bibliographic databases have been conducted in order to assess the extent to which existing patient data has been repurposed for research. This contribution from the IMIA working group on “Data mining and Big Data analytics” focuses on the literature published during the last two years, covering the timeframe since the working group’s last survey.

Results and Conclusions: Although most of the examples of secondary use of patient data lie in the arena of clinical and health services research, we have started to witness other important applications, particularly in the area of genomic research and the study of health effects of environmental factors. Further research is needed to characterize the economic impact of secondary use across the broad spectrum of translational research.

Keywords
Electronic health record; medical informatics; data mining; clinical study; translational medical research

Published online May 8, 2017

1 Introduction

With the advent of technology, a wealth of data sources is becoming available, both for clinicians and biomedical researchers. Although this article presents major examples of secondary use of data extracted from sources in clinical settings, the main challenge for biomedical informatics is to allow for an effective and reliable integration of distributed, complex, and heterogeneous data sources.

Within this context, it is worth highlighting at least three scenarios that involve data integration. The first one is related to the need of permeating the barriers between clinical care, clinical research, and population health [1]. Current information systems have not been designed for cutting across these settings neither the security and confidentiality requirements, nor the appropriate levels of data aggregation, although this responds to different user needs. The second scenario involves a growing diversity in data sources, derived from the availability of different technologies and origins (e.g., genetic, molecular, clinical, environmental, administrative, behavioral, socio-demographic). Each of these data types, as well as the expertise required to process them, poses significant challenges from a data integration perspective [2]. Finally, the third scenario refers to the stakeholder generating the data. We have transitioned from a provider-centered model to a patient-centered one [3]. Patient-(person-) generated health data brings a new avenue for biomedical informatics research.

Only if these integration steps are adequately addressed, Big Data analytic approaches will yield actionable knowledge for medical purposes. The following sections will thus provide insight into successful experiences of the previous.

2 Methodological and Technical Issues Related to the Secondary Use of Patient Data

2a Models for Data Integration

The use of practice-based data to inform research has recently been identified as one of the main pillars for enabling precision medicine in the Learning Health System Cycle [4]. Informatics is the propellant for this process, as it provides methodologies to integrate, store, and share data from different sources, and to analyze such data for research purposes. Several efforts have recently been proposed, both in the US and in Europe,
towards this goal. All these efforts have in common the definition of a model for data integration and of a framework for querying the data, maintaining privacy and security, and complying with ethical requirements [5].

Examples of such systems are clinical research networks (CRNs), which aggregate data coming from different sources represented through a common data model (CDM), and give users the possibility to query the data and perform distributed analytics on it. Through CDMs, it is possible to manage the variety of the sources that make up CRNs, by harmonizing variables and data types found in local repositories, with the ultimate goal of potentiating the complex distributed analytic functionalities that will take place on the data, allowing them to take advantage of larger samples stored in a common format. Examples of CDMs that have been used to develop several research infrastructures are the FDA Mini-Sentinel CDM [6], the Observational Medical Outcomes Partnership (OMOP) common data model [7], and the Informatics for Integrating Biology and the Bedside (i2b2) framework [8].

Among CRNs, PCORnet (http://pcornet.org/) is a National Patient-Centered Clinical research network born in 2014 from an initiative of the Patient-Centered Outcome Research Institute (PCORI) for community-based observational and interventional studies [9]. One of the main features of PCORnet is the direct involvement of patients in the initiative, with the inclusion of 20 Patient-Powered Research Networks (PPRNs) focusing on single health conditions and including rare diseases. Besides PPRNs, PCORnet includes 13 Clinical Data Research Networks (CDRNs), each collecting longitudinal data of at least one million patients each. Data is organized according to the PCORnet common data model [10]. The PCORnet CDM is based on the FDA Mini-Sentinel CDM, with the idea of leveraging the large set of analytic tools and expertise developed for that system, including data characterization approaches, and tools for complex distributed analytics. The CDM, now at its fourth release, aggregates information according to several domains that depend on the source data comes from, such as data captured from healthcare delivery and patient-reported outcomes.

Within PCORnet, health centers belonging to the Scalable Collaborative Infrastructure for a Learning Health System (SCILHS) CRN are all using the i2b2 data warehouse to store and analyze patient data for research. To avoid duplicating extraction-transformation-loading efforts already performed to transfer data from the hospital databases to the i2b2 data warehouse, the i2b2 project has defined a way to translate the i2b2 data model (i2b2 ontology) directly into the PCORnet CDM [11]. The i2b2 data warehouse was successfully mapped to the PCORnet CDM database for all the sites participating to the test, demonstrating that it was feasible to enable the interchange of data between different data models.

Another interesting example of collaborative platform for data re-use is the Observational Health Data Sciences and Informatics (OHDSI) [12]. Differently from PCORnet, OHDSI is mainly focused on data analytics solutions to be applied to networks of health databases [13]. OHDSI is based on the Observational Medical Outcomes Partnership (OMOP) common data model [7], and provides a set of tools for data visualization, summary statistics, and analysis.

### 2b Omics and Big Data

Next generation sequencing technologies are evolving at a very fast pace, outpacing Moore’s Law. These advances together with the decrease in sequencing costs have resulted in a deluge of biological data being generated [14]. As a consequence, computational challenges that have to do not only with data storage but also with data processing and analysis have arisen. Furthermore, data processing is currently the most challenging of the three and has become an evident bottleneck.

For now, we can say that Big Data technologies are still in their early stages. Although powerful, there is clearly room for improvement [15]. Regarding clinical data storage, data warehouses are widely employed as a means to integrate data from different clinical sources. Clinical data warehouses (CDW), thus, provide a unified view on clinical data. Many implementations of these warehouses rely on the i2b2 infrastructure and must be appropriately adapted to handle molecular biology data, as well as to enable clinical and “omics” data integration. As part of the solutions that have emerged to deal with data management and analysis, translational research platforms seem to be on the rise. Cauvel and colleagues analyzed seven of the main non-commercial translational research platforms (BRISK [16], caTRIP [17], cBio Cancer Genomics Portal for Cancer Genomics [18], Georgetown Database of Cancer – G-DOC – [19], integrated clinical omics database – iCOD – [20], integrating data for analysis, anonymization and sharing – iDASH – [21] and tranSMART [22]) in their review [23]. Most of these platforms rely heavily on third-party widely adopted solutions for core components which deal with data storage (e.g., the i2b2 CDW model) or analysis (e.g., the R statistical framework).

Although these platforms represent a first step in the development of tools and methods that enable exploring and managing very diverse data, more advances will be required to fully take advantage of the available data. One key issue identified by the previous authors has to do with privacy. In this sense, several platforms feature a client/server architecture, for which the server is not controlled by the final user. Given the nature of the data, this could involve data privacy issues under the existing regulations. Furthermore, installing the platform in the institution that “owns” the data could be mandatory in many situations. Another major caveat pinpointed by the authors is the fact that none of the reviewed platforms directly support international exchange standards such as HL7 CDA or CDISC ODM. As a result, none of them will be able to import data from EHRs or personal health records. Furthermore, these platforms only present limited capabilities regarding the use of standard terminologies and ontologies, restricting interoperability and data sharing. Finally, the management of complex temporal data has yet to be solved.

Parallelized pipelines will need to be created in order to process “omics” data in a timely manner. Currently, few bioinformatics tools offer parallelization given the complexity involved in their design; some efforts in this arena are discussed by O’Driscoll and colleagues [15]. Cloud computing has also
become very popular recently and could be applied to “omics”. However, in order to take full advantage of the capabilities offered by the cloud, and given the scale of the data involved, new solutions have to be provided regarding data transfer. Uploading and downloading “omics” data from/to the cloud becomes a bottleneck, since the transmission rate is limited. Again, issues with the privacy of medical and “omics” data will have to be tackled. In clinical settings, strong regulations, such as the Health Insurance Portability and Accounting Act of 1996 (HIPAA), will apply.

Lastly, with the launch of precision medicine, new approaches will have to be devised in order to integrate phenotypic, genotypic, and environment-related data. The previously mentioned i2b2 system provides a useful framework which can be successfully applied to this type of approach [24]. Based on this framework, approaches such as BigQ have been developed to deal with genomic variants in a clinical research environment [25].

2c Main Advantages

The main advantages of re-using routinely-collected data for research is that such studies can be conducted on a larger scale, against lower costs, and within shorter time frames than traditional studies [26]. In particular, such studies can have very large cohort sizes and long follow-up times, thus enabling reliable estimation of rare outcomes as well as long-term effects of interventions. The cost of conducting these studies is much lower than for conventional studies because no staff is needed to implement study protocols and record data. So, it is mostly the ‘Volume’ aspect of Big Data that is attractive here. Moreover, the real-world setting in which the data is collected increases the external validity of such studies compared to more traditional designs such as controlled trials. For instance, there will be little or no Hawthorne effects – the phenomenon that people tend to behave differently when they know that they are being observed.

We illustrate some of these advantages with two recent studies. Matthews and colleagues showed that this type of selection is biased towards sick patients [31].

Further data quality issues relate to variation in coding practice [33], bias in recording due to financial incentives, and the use of unstructured data such as free text.

3 Review of Recent Studies on the Secondary Use of Patient Data

3a Secondary Use of Patient Data for Clinical Research

The widespread availability of data collected at different steps of the clinical management of patients (clinical routine, administrative flow, insurance claims) has opened several perspectives for reusing such data in clinical research and, subsequently, in patient care [34]. In this section, we review some...
of the most recent literature in the area, particularly referring to works published in 2015 and 2016.

Since the first experiences in data re-use [35, 36], attention has been focused on the selection and the definition of cohorts of patients to be used in clinical trials. This application well represents the potential intersections between clinical research and patient care, as several data elements useful for clinical trials execution are often already available in electronic health records (EHRs), and their re-use could lead to faster recruitment and decreasing costs, by reducing redundant data capture. Starting from this need, EHR4CR [37], an Innovative Medicine Initiative (IMI) funded project, was aimed at building a platform to support clinical trials based on routinely collected data from EHR systems in Europe [38]. Besides clinical trials execution, the reporting of serious adverse events is also addressed by the EHR4CR project [39]. The EHR4CR network, which included 11 hospitals across Europe, was designed using an architecture promoting the re-use of already available solutions and connect single-center EHRs to the platform. Semantic mapping between local terminologies and the central EHR4CR terminology led to the definition of the EHR4CR Common Information Model. The EHR4CR project ended in 2015 and the platform is currently being scaled up across Europe. Therefore we expect that analyses of data coming from the platform will start being presented. Up to now, a cost-benefit assessment analysis of using EHR data versus current practices has been proposed, showing that the EHR4CR platform would enable a 50% reduction of person-time and operational costs when conducting clinical research scenarios in an oncology clinical trial [40].

Another ongoing EU-funded initiative on data re-use for research is the European Medical Information Framework (EMIF) [41]. This project is focused on the development of a platform able to aggregate data coming not only from clinical EHRs, but also from other sources, such as already performed cohort studies, registries, administrative data, genomic data, and biobanks. The project is mainly focused on identifying predictors for metabolic complications in obesity and for Alzheimer's disease, and has already succeeded in creating large sample sets for biomarker discovery by merging patients cohorts spread across Europe using an OMOP-based common data model [42].

In the US, some important projects related to data re-use are also present. One example is the Shared Health Research Information Network (SHRINE) [43], developed at Harvard Medical School as one of the first networks that allowed researchers to perform federated queries on populations of patients aggregated across many hospitals. In analogy with i2b2, which is the default platform SHRINE is designed to work on, SHRINE common data model is based on an ontology (called the Core Ontology) that defines the set of medical concepts that can be used to query the system. On the commercial side, a successful example is represented by the Trinetx network (http://trinetx.com/), a cloud-based health research platform that allows healthcare organizations, pharmaceutical companies, and research organizations to collaborate in the process of trial design and patient recruitment.

The automatic selection of patients eligible for a clinical trial is one of the applications of what is generally known as electronic phenotyping [44], a concept that is strictly linked to the reuse of patient data for research and that refers to a series of methodologies to derive relevant features from EHR data and use such features to answer relevant research questions [45]. Beyond supporting the recruitment of patients for clinical trials, electronic phenotyping can be used for the stratification of patient cohorts and for the implementation of decision support tools to be embedded within EHR systems [45]. Several advanced and robust methodologies have been proposed to extract phenotypes from routinely collected clinical data and reuse the results for research [46]. As part of the i2b2 project, one general approach was defined that allows developing a robust phenotype algorithm to be applied to several clinical domains, like diabetes, depression, multiple sclerosis, and rheumatoid arthritis [47]. Such algorithm, besides considering the structured data included in EHRs, also uses natural language processing (NLP) to extract information from clinical text reports and notes. Once relevant domain-specific features are defined for the domain of interest and mapped to standard terminologies, they are extracted from structured and unstructured data and ranked using a feature selection technique. A logistic regression classification algorithm is finally used to assign to each patient a probability of having the desired phenotype.

To overcome the limitations of manual curation of the initial set of features that defines a phenotype, automatic algorithms have been proposed. In [48], features are selected using publicly available knowledge sources and NLP. The recently proposed “deep patient” framework uses unsupervised feature learning to extract a set of general, domain-independent, features from EHR databases [49]. Deep patient exploits a deep learning approach based on multi-layer neural networks to extract a compact, high-level, representation of the patient. The features extracted using this methodology are used to predict patients’ future diseases.

Time has also been identified as a relevant feature that is naturally embedded by the data collection process in clinical practice. Taking into account the dynamics of the collected data allows specifying a novel way to stratify the population, called dynamical phenotyping [50]. Besides clustering patients based on their dynamics [51], novel techniques have been proposed to mitigate some of the biases that characterize data collected in clinical routine, namely irregular sampling time and non-stationarity [52], and to preserve temporal relationships in a privacy-safe context [53].

Besides integrating data and making query services available through users, systems like i2b2 have been recently enhanced with some advanced analytics functionalities that go beyond the original goal of cohort selection. An example is given by the functionalities for statistical analysis implemented in the Analytics cell for i2b2 [54]. The applications currently included in this framework allow performing summary statistics on a patient set, testing hypotheses with suitable data visualization, and the possibility to draw survival curves.

While data reuse for research has been broadly addressed in the past two years, its effective adoption in patient care is less common. Although the majority of
the systems based on clinical data reuse are expected to have an impact on patient care in the future, the number of systems that are already adopted in clinical practice is still very limited. This is due to the fact that closing the learning healthcare system cycle by transferring the results of research to the clinical practice is a complex process. Nevertheless, some examples of systems that already use available clinical data to improve decision support tools already exist. One example is given by a web-based population risk surveillance dashboard developed to predict and visualize the risk of future admissions to the emergency departments relying on data coming from EHRs and Health Information Exchange (HIE) [55]. The system is developed as a web application that integrates a predictive model validated on a large cohort of patients that provides real-time visualization of the population risk profile. The system is running on the Maine State Health Information Exchange system. Another example of a system based on the re-use of clinical data to support patient care is described in [56]. This system is aimed at supporting clinicians and healthcare managers in the management of type 2 diabetes and prevention of complications. This system is designed as a dashboard including several analytics algorithms working both at the population and the single-patient level. The developed models are derived thanks to the re-use of routinely collected clinical and administrative information, which are integrated exploiting the i2b2 technology.

3b Secondary Use of Patient Data for Genomic Research and to Study the Effect of the Environment on Health

Besides secondary use of patient data for clinical research and patient care, there are other applications, most of which are at the intersection of EHRs and genetics. In this context, biorepositories play a crucial role in generating data for potential secondary use. However, in general, the data available for this type of secondary use is very heterogeneous. Although several efforts are being made regarding data integration and standardization, there exist many private biorepositories scattered around the globe, which in many cases may have been created as a result of studies initiated by different groups of investigators. Most of this data has not been carefully organized for future reuse. As mentioned earlier, integrative approaches have to be developed to overcome the current landscape. In this section, we will review the most recent (2015-2016) and relevant works that deal with secondary use of patient data in the field of genomics and regarding the effect of the environment on human health.

With the decrease in cost of technology, it is now possible to perform multi-omics high-throughput sequencing at a reasonable price and time. However, in the case of whole genome sequencing, for example, the interpretation of the extremely high number of variants obtained both in coding and non-coding regions represents a computational challenge that has yet to be fully solved. Dried blood spots (DBSs) have been collected on a regular basis from newborns in many countries for screening purposes. Although not free of controversy (especially in the US), these collections are becoming a wealth of biological data for research. According to a study carried out in the Netherlands, Dutch mothers lean towards supporting secondary use of this data. However, more efforts are required in order to involve parents in developing policies that deal with DBSs [57]. Several researchers have used DBS data to gain more insight into different conditions and disorders. DBSs have been even used to study how the effect of the environment can affect several generations as demonstrated by Sen and colleagues in their work on inheritance of epigenetic changes associated with exposure to lead [58].

Other examples of secondary use of routine genetic data involve the monitoring of HIV transmission hotspots [59], and the prediction of long-term risk of severe infection using data from three different population-based studies, namely the Dietary, Lifestyle, and Genetic Determinants of Obesity and Metabolic syndrome (DIL-GOM) Study, the FINRISK Study, and the Cardiovascular Risk in Young Finns Study (YFS) [60]. In the former, researchers built a monitoring system around a cached database query, which contains de-identified data for all individuals enrolled in the British Columbia drug treatment program. Data is clustered based on phylogenetic distance and clusters are then annotated with clinical and demographic information. Finally, the resulting analysis is presented to the user through a graphical interface as a network. In the latter, network analysis is used to obtain modules of co-expressed genes which are then annotated with Gene Ontology (GO) and KEGG to look for enrichment. Regression models are also used to analyze the rest of the data.

The Electronic Medical Records and Genomics network (eMERGE) is a remarkable initiative organized and funded by the National Human Genome Research Institute (NHGRI). eMERGE was created to evaluate the possibility of using data from routine clinical care retrieved from electronic medical records (EMRs) to characterize disease phenotypes in genome wide association studies (GWAS) [61]. Within this initiative, a variety of methods and best practices were developed for utilizing EHRs as a tool for genomic research [62, 63]. Before widely exploiting this data for research, several authors worked to prove the benefits of EHR-derived phenotyping. Thus, Verma and colleagues used EMR data from this network to validate the utility of EMRs to detect non-sporious and relevant associations in glaucoma [64]. For this purpose, authors carried out a GWAS using logistic regression models. After that, gene-gene interaction and pathway analyses were performed, replicating the study on the NEIGHBOR consortium dataset. EHR phenotyping has been further validated in other contexts, such as the study of liver function [65].

An electronic phenotyping algorithm was developed within the eMERGE network. This algorithm allowed Almoguera and colleagues to obtain more comprehensive information on asthma [66]. Furthermore, electronic phenotyping enabled these researchers to obtain a more accurate and objective definition of the condition, basing its diagnosis on ICD9 codes and the prescription of specific medication. As a result, the authors were able to unravel pathways that could help in improving the treatment of underlying causes of asthma. In addition to the algorithms devised as part of the eMERGE initiative, other methods have arisen to complement traditional
genotype-phenotype studies also deriving
phenotypes from EMRs, such as those de-
veloped and applied in the Epidemiologic
Architecture for Genes Linked to Environ-
ment (EAGLE) study [67, 68].

The possibility of conducting several
GWAS at the same time with no additional
cost (especially economic) is very benefi-
cial. The previous studies prove that EHRs
provide a new avenue for reusing already
existing genotypes by studying their rela-
tionships with different phenotypes. In
addition to disease and trait genetics, EHRs
are also a precious resource for investigat-
ing pharmacogenetic traits [69] and develop-
ing reverse genetics approaches such as phe-
nome-wide association studies (PheWAS)
[70]. PheWAS allow generating hypotheses
by testing a selected group of single nu-
cleotide polymorphisms (SNPs) and many
phenotypes. Therefore, in the past few years,
with the proliferation of data available for
secondary use, these studies have been on
the rise. A comprehensive review of the use
of EHRs in this type of study has been done
by Bush and colleagues [71]. Finally, Denny
and colleagues underscored the relevance of
PheWAS in the context of precision medi-
cine, identifying this type of approach as key
to new discoveries [70].

Although relatively little attention has
been given to the extraction of data regard-
ing determinants of health from EMRs,
this is an area with many opportunities that
will flourish in the near future. This type of
data will help improve not only healthcare
of individual patients, but will also enable
population-level applications [72]. For this
purpose, social and medical data will need
to be integrated and aggregated, leading to
major changes in healthcare, improving val-
ue-based payment, and eventually reflect-
ing on the advancement of population health.

3c Secondary Use of Patient Data for
Evaluation of Health Interventions

An increasing number of studies re-use
routinely collected health data to evaluate
the effects of health interventions. Typically,
these studies retrospectively extract baseline
and follow-up data from individuals that
were exposed to the intervention, and from
a similar group of individuals not exposed
serving as controls. Often, data from differ-
ent sources are linked to ensure that health
outcomes are reliably captured. Regression
methods are used to estimate the causal
effects of the interventions on health.

A key prerequisite for this type of re-
search is that there is sufficient variation in
real-world practice to evaluate interventions.
If, in routine practice, everyone with a cer-
tain condition receives intervention X, then
there is no opportunity to evaluate the effects
of X with routine data, simply because no
comparison can be made. But once there
is variation in practice, i.e. some patients
receive the intervention while similar others
do not, this offers the opportunity to view the
real world as a natural experiment and draw
inferences from that. So, for the evaluation
of interventions, it is the ‘Variation’ aspect
of Big Data that is essential.

As described earlier, there are many
advantages to re-using routinely collected
health data for research. This type of research
has lower costs, larger sample sizes, longer
follow-up times, and is more representative
of real-world clinical practice. However,
there are also significant risks of bias, threat-
ening the validity of these studies. Research
that re-use previously recorded health data
has, by definition, a retrospective, obser-
vational (non-randomized) study design.
This design is relatively weak for assessing
causal effects, especially when compared
to prospective, randomized, study designs.
The main thread to validity of these research
works is confounding, i.e., a lack of com-
parability between exposed and unexposed
groups. The presence of confounding means
that the exposed group is essentially differ-
ent from the unexposed group at baseline.
Had the exposed subjects actually been un-
exposed, their outcomes would be different
from those in the actually unexposed group.
For instance, it is conceivable that some of
the people who underwent CT scanning in
the study by Matthews and colleagues [27],
described earlier in this paper, were tested
because they had unexplained symptoms
that later turned out to be caused by a can-
cer diagnosis. The authors have minimized
the risk that this happened by excluding all
patients that received a cancer diagnosis
within one year after the CT scan – but we
cannot exclude the possibility that there
were other patients with a longer lead
time. Similarly, neonatal care has probably
improved during the years 2007-2011. Been
and colleagues’s interrupted time series
design cannot distinguish the causal effects
of such improvements from the effects of
the smoking ban, which thus may be over-
estimated in the study [28].

A common method to address confound-
ing is propensity scoring [73]. A propensity
score is the estimated probability that an
individual will be exposed to the inter-
vention, conditional on observed baseline
characteristics. A propensity score allows
one to analyze observational data so that it
mimics the characteristics of a randomized
controlled trial. Propensity scores are typi-
cally constructed by training a probabilistic
classifier (such as a logistic regression
model) using the baseline characteristics
as predictive features and exposure status
as a class label. For instance, De Vries
and colleagues [74] used generalized boosted
regression to estimate the probability of
receiving outpatient cardiac rehabilitation
after hospitalization for a cardiac event or
intervention, in a large insurance claims
database (3.7 million people) from the
Netherlands. Subsequently they applied Cox
proportional hazards regression analysis
with an inverse propensity score weighing
to assess the effect of cardiac rehabilitation
on survival. They found that rehabilitation
reduced mortality risks by 35% in the first
four years after the intervention.

Another way to address confound-
ing is to select controls that are very similar to the
exposed individuals — so-called ‘matched
controls’. Once a matched sample has
been formed, the exposure effect can be
estimated by directly comparing outcomes
between exposed and unexposed subjects
in the matched sample. Direct matching
uses known confounders (e.g., age, sex,
and diagnosis) to identify controls with
exactly the same values for those variables.
The number of matching variables and their
categories must be kept small to prevent that
no control would be found. An alternative
is to match by propensity score, which typi-
cally allows inclusion of a larger number of
confounders. However, if the propensity
score is mis-specified, the covariates in the

UNIA Yearbook of Medical Informatics 2017
matched sample will be imbalanced, which can lead to a conditional bias. Therefore, a third, more recently developed approach, uses an evolutionary search algorithm to directly maximize covariate balance between exposed individuals and matched controls [75]. It does not depend on knowing or estimating a propensity score. Steventon and co-workers [76] used this ‘genetic matching’ technique in their evaluation of the effects of a telephonic alert system in patients with chronic obstructive pulmonary disease (COPD). The alert system used routine meteorological and communicable disease reports to identify times of increased health risks and then alerted the participants of the study. They found that this intervention increased hospital admission rates but reduced mortality.

When there is missing data (which will very often be the case when health data is re-used for research), this can pose significant challenges to the adjustment for confounding. For instance, in the administrative dataset used by De Vries and colleagues [74], there was no data available on cardiac function, a well-known confounder of survival for cardiac patients. It was probably assessed for most included patients but not included in the insurance claims that were re-used for the study. To address this problem, the authors used Lasso regression [77], a statistical feature selection method, to construct a set of proxies for cardiac function and other potential confounders from more than 900 variables that were available in the dataset (hospital diagnosis-treatment combinations, outpatient prescriptions, medical devices, occurrence of lab tests, GP visits, ICU days, and other services). The resulting set of 99 proxies included variables such as ‘use of digoxis glycosides’, a potent cardiovascular drug that is typically prescribed for patients with poor cardiac function.

This example illustrates that the breadth of routinely collected datasets can compensate for their modest fit with research purposes if we apply advanced machine learning and artificial intelligence techniques to discover the relevant features in data. Recently, researchers have started to use unsupervised deep learning methods to derive compact, feature-based, patient representations from EHR data that are strongly predictive of health outcomes [49]. Such methods could further strengthen the existing array of propensity scoring and matching techniques.

4 Analysis of Economic Benefit

The development of EHRs and the use of their information for research purposes has been a recurrent theme in biomedical informatics, and it is a key and pivotal element for the development of what is currently known as precision medicine [78]. It must be noticed that in most of the cases when we refer to the use of EHR data, we mostly refer to data that is derived from an “ad-hoc” patient data repository or data warehouse, rather than the operational EHR itself [79].

Since the beginning of EHR implementation, there has been a clear interest in assessing and analyzing which were the economic impacts of introducing these technologies [80-84]. Most of the research carried out so far has been focused on the analyses in different areas such as:

- Clinical care
- Education
- Population and Global health
- Quality management and patient safety

However, and despite the increasing attention and the number of publications describing the secondary use of clinical data for research purposes, little research has been focused on the analysis of the economic aspects derived from this secondary (research) use of EHRs.

We conducted a series of literature searches in PubMed to assess to what extent this kind of analyses has been performed so far and what fields of biomedical research were the most commonly assessed from an economic perspective. These literature searches have been based on the use of different MeSH terms (“Biomedical Research”, “Electronic Health Records” and “Economics”). At the time of writing this article (December 5, 2016), the Boolean search of these three terms together using the “AND” operator yielded a total of 90 documents, spanning from 2008 to 2016. In addition to that search, we also performed a second search using “Electronic Health Records/economics” [85] and “Biomedical Research/economics”[MeSH Terms] resulting in 7 articles.

Most of this research about the economic impacts of the secondary use of clinical information derived from EHRs for research purposes has been focused on a couple of areas, namely clinical trials and biobanks. The outcomes of these analyses are reflected in a still incipient literature which has been focused around two major dimensions:

- Quantitative vs. qualitative analyses. The major difference between these two approaches lies in the explicit analysis of the economic variable (quantitative analyses) or the use of other proxies and surrogate measurements that could be potentially linked with economic parameters such as time (qualitative analyses).
- Research phase. This dimension reflects the different stages (i.e. recruitment or execution) where the benefits of using data derived from EHRs may have an economic impact.

The European project EHR4CR is an excellent example of these research approaches. As a consequence of this work, A. Beresniak and colleagues published in January 2016 a research article that clearly shows the quantitative economic impact derived from the use of EHR-extracted data in three different clinical scenarios (protocol feasibility, patient selection, and study execution) [40]. On the other hand, Geisinger’s MyCode project [86] is a recent example of some of the aforementioned application fields and dimensions (qualitative analyses on a biobank). Some other examples identified from the literature can be checked in Table 1.

Although there is an increasing body of literature focusing on some of the “applied research” areas such as clinical trials, an important aspect that remains obscure in the literature is the impact of EHRs on fundamental research. To our knowledge, the actual economic impact of these systems applied to the so-called “fundamental/basic” research has not been yet assessed.

A recurrent topic for the development of what initially was known as personalized medicine and more recently has been called precision medicine was the use of data and
Table 1 Examples of assessment of the economic impact derived from using EHRs for research.

<table>
<thead>
<tr>
<th>Subject Recruitment</th>
<th>Quantitative</th>
<th>Qualitative</th>
</tr>
</thead>
<tbody>
<tr>
<td>Massachusetts General Hospital and Partners Healthcare Inc. Boston estimated savings at $74M per year [87].</td>
<td>Northwestern University was the #1 enrolling U.S. sites (among 155 total sites) in a clinical study of heart failure [89].</td>
<td></td>
</tr>
</tbody>
</table>

| Study Execution | | |
|-----------------| At Vanderbilt University, the cost per pharmacogenetic study was approx. $77,000 when using VESPA to link biobank with EHRs and enable cost effective research as compared to $438,000 per study when not using VESPA [88]. | Northwestern University Feinberg School of Medicine needed to collect follow-up data on heart failure patients admitted through the Northwestern Memorial Hospital. In four days, they electronically queried and refined the records for 40,953 ED admissions, 4,333 of which met the criteria for the study [90]. |

5 Conclusions

This article has provided an overview of secondary use of patient data from a Big Data perspective. During our review, several common themes emerged. These include issues about data models, Big Data approaches to individual omics data, main advantages of these approaches, and concerns about data quality. We present them here in a special section before our review of recent and relevant examples of secondary use of patient data for purposes of clinical, genomic and health services research. We have also reviewed the economic impact of secondary use of clinical data. As we will witness in the next years the rise of very large collaborative research initiatives, such as the Precision Medicine Initiative (PMI, All of Us Research Program) funded by the US government, we hope that this review may contribute to the identification of barriers, opportunities, and potential solutions. The PMI is a scientific endeavor that explicitly acknowledges the need to integrate environmental, behavioral, and genetic data with EHRs and patient-(person-) generated data. The PMI Cohort Program aims at collecting data from one million or more participants, which will be made available to researchers, clinicians, and participants as a public repository. Precision medicine could both benefit and contribute immensely to the progress of the Learning Healthcare System paradigm. This model pursues an information ecosystem in which every clinical encounter becomes a source of data for clinical research, closing the loop between clinical care and biomedical research. This approach is a key step towards facilitating a truly personalized healthcare system.

References

22. Szalma S, Koka V, Khasanova T, Perakslis ED. Effective knowledge management in translational


Correspondence to:
Fernando J. Martin-Sanchez, Ph.D. FACHI, FACMI
Weill Cornell Medicine
Department of Healthcare Policy and Research
Division of Health Informatics
425 East 61st Street
New York, NY 10065, USA
Tel: +(1) 646 962 9438
E-mail: fem2008 @ med.cornell.edu