BENTHAM OPEN

CrossMark

The Open Bioinformatics Journal

Content list available at: www.benthamopen.com/TOBIOIJ/

DOI: 10.2174/1875036201709010001

RESEARCH ARTICLE

Data Mining Approach to Estimate the Duration of Drug Therapy from Longitudinal Electronic Medical Records

Olga Montvida[1,2], Ognjen Arandjelović[3], Edward Reiner[4] and Sanjoy K. Paul[5,*]

[1]*Clinical Trials and Biostatistics Unit, QIMR Berghofer Medical Research Institute, Brisbane, Australia*
[2]*School of Biomedical Sciences, Institute of Health and Biomedical Innovation, Faculty of Health, Queensland University of Technology, Brisbane, Australia*
[3]*School of Computer Science, University of St. Andrews, St. Andrews, United Kingdom*
[4]*Smart Analyst Inc., New York, Unites States of America*
[5]*Melbourne EpiCentre, University of Melbourne and Melbourne Health, Melbourne, Australia*

Received: March 27, 2017 Revised: May 06, 2017 Accepted: May 12, 2017

Abstract:

Background:

Electronic Medical Records (EMRs) from primary/ ambulatory care systems present a new and promising source of information for conducting clinical and translational research.

Objectives:

To address the methodological and computational challenges in order to extract reliable medication information from raw data which is often complex, incomplete and erroneous. To assess whether the use of specific chaining fields of medication information may additionally improve the data quality.

Methods:

Guided by a range of challenges associated with missing and internally inconsistent data, we introduce two methods for the robust extraction of patient-level medication data. First method relies on chaining fields to estimate duration of treatment ("chaining"), while second disregards chaining fields and relies on the chronology of records ("continuous"). Centricity EMR database was used to estimate treatment duration with both methods for two widely prescribed drugs among type 2 diabetes patients: insulin and glucagon-like peptide-1 receptor agonists.

Results:

At individual patient level the "chaining" approach could identify the treatment alterations longitudinally and produced more robust estimates of treatment duration for individual drugs, while the "continuous" method was unable to capture that dynamics. At population level, both methods produced similar estimates of average treatment duration, however, notable differences were observed at individual-patient level.

Conclusion:

The proposed algorithms explicitly identify and handle longitudinal erroneous or missing entries and estimate treatment duration with specific drug(s) of interest, which makes them a valuable tool for future EMR based clinical and pharmaco-epidemiological studies. To improve accuracy of real-world based studies, implementing chaining fields of medication information is recommended.

* Address correspondence to this author at the Melbourne EpiCentre, University of Melbourne and Melbourne Health, Melbourne, Australia; Tel: +61 3 93428433; Fax: +61 3 93428780; E-mails: Sanjoy.Paul@mh.org.au; sambhupaul@hotmail.com

Keywords: Electronic medical records, Treatment duration, Data mining, Type 2 diabetes, Rule-based algorithm, Patient-level data aggregation.

1. INTRODUCTION

The electronic medical records (EMRs) and the administrative data from the primary/ambulatory care systems are increasingly being used in epidemiological [1 - 3], pharmaco-epidemiological [4 - 6], pharmaco-vigilance [7 - 9], clinical outcome [5, 10 - 12], health economic [13, 14] and public health related studies [15 - 18]. Analyses of large primary care based EMRs from various countries, most notably from UK, USA and Sweden, have provided significant insight into the effectiveness of changes in health care practices/polices on overall disease and health management costs [3, 15, 19, 20], in addition to population level evidences on the safety and effectiveness of various therapies and the association of disease-related risk factors on long-term outcomes [5, 6, 18, 21 - 23]. Increasing use of such large real-world patient-level data is illustrated well by the sixfold increase in EMR based published studies since 2000 [10, 24].

In structured EMRs, especially from the primary/ambulatory care systems, comprehensive patient level data are captured on different domains simultaneously and stored in the form of relational database [25, 26]. Representative examples include the UK Clinical Practice Research Database and Centricity™ EMR (CEMR) database of USA [27, 28]. The extraction, quality control and management of such voluminous longitudinal data under individual study protocols is highly methodologically and computationally involved, and challenging from data mining and analytical viewpoints [22, 29]. Data science generally considers that data preparation tasks consume about 80% of total project timeline leaving only 20% for ultimate analysis itself [30, 31]. Data completeness, systematic biases, reproducibility and quality are some of the notable limitations in such databases [18, 29, 32].

Most EMR databases capture large amounts of detailed information on medications provided to individuals over time, while specific form in which this information is stored varies from database to database [26]. It is usually possible to obtain the drug class, specific brand name within the corresponding class, prescription dates, dosage, and number of refills [32]. However, a significant number of entries for an individual prescription may be missing or contain errors. The problem with information completeness can also arise when the medication nomenclature is not correctly matched [29].

Clinical and pharmaco-epidemiological studies, which rely on the data from EMRs, are often interested in the effectiveness of specific therapies, therapeutical dynamics, treatments with concomitant medications, and durations thereof in specific disease areas. Such real-world analysis provides an extremely valuable means for the understanding of drug utilization patterns, treatment initiation periods following the diagnosis of a disease, the effectiveness of specific therapies on disease-related risk factors, and possible associations of therapies with long-term outcomes [1, 6]. These studies warrant appropriate extraction of longitudinal information on prescriptions or medications at individual patient level, inappropriate extraction of the data may result in misleading inferences reported [33 - 35]. Generally, pharmaco-epidemiological studies do not estimate treatment duration, but only account for the fact of one or more prescriptions for a particular drug(s) [36, 37]. Some studies calculated medication duration by extracting first prescription date from the last prescription date [38, 39], and only few studies additionally considered a drug being discontinued if the subsequent prescription was not refilled within the expected time of drug coverage [40, 41]. While some studies have discussed the challenges in the analysis of medication data from EMRs [18, 42], to the best of our knowledge no existing study has analysed the quality, consistency, and completeness of EMR prescription information, nor proposed a practical algorithm able to extract salient medication information from large and complex longitudinal data sets [43].

The aims of this explanatory and methodological study are (1) to discuss and analyse the most pressing challenges encountered by computer based methods in the process of extracting and aggregating longitudinal medication data from EMRs, (2) to describe two algorithms to extract prescription information of individual therapies and to estimate the corresponding duration of treatment, and (3) to discuss how estimates of individual medication duration are affected by the choice of the study design. The effectiveness of algorithms is compared is on a cohort of patients with a clinical diagnosis of type 2 diabetes (T2DM) using a real-world EMR database collected across the USA.

2. MATERIALS AND METHODS

2.1. Centricity Electronic Medical Records

The CEMR database contains more than 40 million patients' clinical/treatment records from 1995. CEMR represents 49 US states and a variety of ambulatory medical practices, including solo practitioners, community clinics,

academic medical centres, and large integrated delivery networks. The database has been extensively used for academic research worldwide [3, 37, 44 - 47]. The CEMR database consists of over 30,000 health care providers, of whom approximately 70% are primary care providers. For both insured and uninsured patients, this database contains comprehensive patient-level information on many aspects including demographic information, laboratory results, history of diseases, clinical diagnosis of symptoms/ diseases, vital signs, history of medications and detailed information on the ongoing medications. For this study we used longitudinal information from January 1995 to October 2014.

2.2. Medication Data in Centricity EMR database

The medications taken by an individual (medication domain) and the prescriptions for drugs provided to the individuals by the service provider registered within the EMR system (prescription domain) are extensively documented in the database by means of three tables: medication dimension (MD), medication fact (MF) and prescription fact (PF). The MF and PF belong to the medication and prescription domains respectively. The MF may include a broader list of all medications that a patient is taking including over the counter medications, herbal remedies and medications prescribed by a provider that may be out of the EMR network. MD is linked to both MF and PF. Each record in the MD contains information on individual drug, which includes the National Drug Code (NDC) and Generic Product Identifier (GPI), as well as the four ordered attributes derived from the GPI such as generic drug names. The MD also includes the medication doses corresponding to different brands' products, identified by a unique medication key value assigned to each record.

The entries in MF capture individual patient's medication prescription history and active prescriptions from all practitioners including the service provider registered within EMR system. It contains several special fields to track longitudinal patterns, such as active medication flag, which indicates if a patient was taking the drug at the database extraction moment. Active medication list is identified by records with value "Y" of active flag. The chain identification (ID) values facilitate tracking of treatment alterations (including the addition of new medications) over time, with the related chain sequence values which track medication adjustments within the same chain ID. The initiation ('start') and cessation ('stop') dates associated with different treatments are also stored in the MF. However we found that the corresponding values are missing with alarming frequencies: 67% of the cases for the former and 11% for the latter. Also, some of the start and stop date entries could be erroneous, such as stop date preceding start date. An excerpt from the MF for an individual patient is shown in Table **1**.

Table 1. Snapshot of MF table – treatment intensification.

GPI category 4	Medication key (M)	Patient key (P)	Create date (C)	Start date (B)	Stop date (S)	Active flag (F)	Chain ID (H)	Chain seq (G)
METFORMIN HCL	41467	288859	6-May-09	6-May-09		N	307667619	0
METFORMIN HCL	41467	288859	11-Jun-10	11-Jun-10		N	307667619	1
METFORMIN HCL	41467	288859	25-Apr-11	11-Jun-10	25-Apr-11	N	307667619	2
LIRAGLUTIDE	3347202	288859	25-Apr-11	25-Apr-11		N	812855070	0
LIRAGLUTIDE	3347202	288859	10-May-11	10-May-11		N	812855070	1
LIRAGLUTIDE	3347202	288859	10-May-11	10-May-11		N	820957274	0
LIRAGLUTIDE	3347202	288859	14-Dec-11	10-May-11	14-Dec-11	N	820957274	1
LIRAGLUTIDE	3347202	288859	14-Dec-11	10-May-11	14-Dec-11	N	812855070	2
INSULIN GLARGINE	682327	288859	27-Feb-12			N	1092145628	0
INSULIN ISOPHANE HUMAN	682834	288859	27-Feb-12			N	1092145627	0
INSULIN GLARGINE	682327	288859	26-Sep-12	26-Sep-12		N	1092145628	1
INSULIN ISOPHANE HUMAN	682834	288859	14-Nov-12			N	1092145627	1
INSULIN GLARGINE	682327	288859	14-Nov-12			Y	1092145628	2
INSULIN LISPRO (HUMAN)	682825	288859	26-Feb-14	26-Feb-14		Y	1092145627	2

The entries in the PF capture the prescription date and the associated number of refills only for medications that have been prescribed by the responsible provider within the EMR network. The MF dataset contains a broader set of entry sources, moreover the form of recording potentially comprises more details than corresponding data in the PF. Nevertheless it was determined that PF may contain unique entries that are not stored in MF. Therefore, the MF was considered as the primary source of medication information and the PF as a complimentary one.

3. METHODS

In this section, we introduce a novel algorithm for mining large-scale longitudinal EMRs with the ultimate goal of estimating the duration of treatment of a particular individual with a drug(s) of interest. The first method we introduce ("chaining") relies on chain ID and chain sequence values recorded in the MF. This feature of the approach allows to account for treatments which include alternative drug use. To assess the importance and power of longitudinal chain information, we also describe a modification of the "chaining" method ("continuous") which disregards chain ID and chain sequence values, and instead relies only on the chronology of patient's records of particular drug(s). In the current literature, the latter approach is used more frequently.

3.1. Data Pre-processing: Auxiliary Fields

Although erroneous entries generally cannot be identified, various types of global consistency rules may be applied to reduce the error. Chronology of the events may be corrected by incorporating two additional fields: patient's last available follow-up date and patient's date of birth (DOB).

CEMR database stores last available follow-up date for each patient. As initial data pre-processing step, erroneous follow-up date entries were identified and corrected by the latest record creation dates of all activities within the database for corresponding patients.

Similar to many anonymized EMRs, the exact DOB was not available within CEMR. Simple procedure was applied to approximate DOB:

1. Obtain multiple DOB estimates per patient by subtracting reported 'valid' age from the record creation date for all activities within the database. CEMR groups patients older than 80 years under a single age key. The non-missing age data and the non 80+ age keys were considered as 'valid' age entries.
2. Approximate DOB as minimum of all estimates from Step 1.
3. For patients without reported activities estimate DOB from the dataset containing demographic information by subtracting reported 'valid' age from the database extraction date.

The parameters for the mathematical formulations are identified in the Table **2** below.

Table 2. Mathematical Formulation

Scalars	
n	number of records in MF table
k	number of records in PF table
sd	standard prescription duration for individual drug
mx	maximal number of prescription refills for individual drug
u	number of unique patient keys in the cohort of interest
Sets	
$PS = \{ps_1, ps_2, \ldots\ldots ps_u\}$	set of unique patient keys in the cohort of interest
V	set of missing values
MS	set of medication keys of selected drug(s)
$F_Y = \{f_i \mid f_i = \text{"Y"}, i = \overline{1,n}\}$	set of active drugs
MF={M,P,C,B,S,F,H,G} dataset	
$M = (m_1, m_2, \ldots, m_n)^T$	medication keys for drugs
$P = (p_1, p_2, \ldots, p_n)^T$	patient keys
$C = (c_1, c_2, \ldots, c_n)^T$	record creation dates
$B = (b_1, b_2, \ldots, b_n)^T$	start dates of individual records
$S = (s_1, s_2, \ldots, s_n)^T$	stop dates of individual records
$F = (f_1, f_2, \ldots, f_n)^T$	active medication flag values
$H = (h_1, h_2, \ldots, h_n)^T$	chain identification values
$G = (g_1, g_2, \ldots, g_n)^T$	chain sequence values
PF={M,P,C,B,R} dataset	
$M = (m_1, m_2, \ldots, m_k)^T$	medication keys for individual prescriptions
$P = (p_1, p_2, \ldots, p_k)^T$	patient keys

(Table 2) contd.....

$C = (c_1, c_2,\ldots, c_k)^T$	record creation date
$B = (b_1, b_2,\ldots, b_k)^T$	prescription dates
$R = (r_1, r_2,\ldots, r_k)^T$	number of refills for individual prescription

The scalars *sd* and *mx* may be defined on the basis of the standard prescription protocol for individual drugs. The default values of *sd =1* and *mx =24* were considered in our analyses.

MS may be identified by text-mining the MD dataset. For example, glucagon-like peptide-1 receptor agonist (GLP-1RA) may be identified by searching for "GLP-1 RECEPTOR AGONIST" in the second order GPI attributed field.

3.2. "Chaining" Method

The algorithm for the first approach to extract and aggregate data for the estimation of duration of treatment is elaborated below.

1. Merge the following to the MF dataset by patient key:

1.1) date of birth $DOB = (db_1, db_2,\ldots,db_n)^T$.

1.2) last available follow-up date $L = (l_1, l_2,\ldots,l_n)^T$. The extended MF dataset would be of the form.

$$MF^1 = \{M, P, C, B, S, F, H, G, DOB, L\}$$

2. Replace erroneous values of start dates ($b_i \notin V \wedge (b_i<db_i \vee b_i>s_i \vee b_i>l_i)$, $i = \overline{1,n}$) with missing values

3. Sort by patient key ascending, chain ID ascending within the same patient, chain sequence descending within the same chain ID.

MF^1: $a)\ p_i \le p_{i+1}, i = \overline{1,n}$

$b)\ h_i \le h_{i+1}, \forall i: p_i = p_{i+1}$ - post *a)* sorting

$c)\ g_i \ge g_{i+1}, \forall i: h_i = h_{i+1} \wedge p_i = p_{i+1}$ - post *b)* sorting

4. Set initial values $p_0 = 0$, and approximate individual medication end dates $E = (e_1, e_1,\ldots,e_n)^T$ on the basis of the following rules:

4.1) if stop date is not missing, then end date equals to stop date.

4.2) else, if active flag is "Y", then end date equals to last follow-up date.

4.3) else, if first unique value of patient key or first unique value of chain ID, and start date is not missing, then end date equals to start date plus standard prescription duration.

4.4) else, if first unique value of patient key or first unique value of chain ID, and start date is missing, then end date equals to record creation date plus standard prescription duration.

4.5) else, end date equals to the create date of a previous record.

$$e_i = \mathbb{I}_{\{b_i \notin V\}} \cdot \left(\mathbb{I}_{\{p_i \neq p_{i-1}\}} + \mathbb{I}_{\{h_i \neq h_{i-1}\}} - \mathbb{I}_{\{p_i \neq p_{i-1}\}} \cdot \mathbb{I}_{\{h_i \neq h_{i-1}\}}\right) \cdot \left(s_i \cdot \mathbb{I}_{s_i \notin V} + l_i \cdot \mathbb{I}_{f_i \in F_y} \cdot \mathbb{I}_{s_i \in V} + (b_i + sd) \cdot \mathbb{I}_{f_i \notin F_y} \cdot \mathbb{I}_{s_i \in V}\right) +$$

$$\mathbb{I}_{\{b_i \in V\}} \cdot \left(\mathbb{I}_{\{p_i \neq p_{i-1}\}} + \mathbb{I}_{\{h_i \neq h_{i-1}\}} - \mathbb{I}_{\{p_i \neq p_{i-1}\}} \cdot \mathbb{I}_{\{h_i \neq h_{i-1}\}}\right) \cdot \left(s_i \cdot \mathbb{I}_{s_i \notin V} + l_i \cdot \mathbb{I}_{f_i \in F_y} \cdot \mathbb{I}_{s_i \in V} + (c_i + sd) \cdot \mathbb{I}_{f_i \notin F_y} \cdot \mathbb{I}_{s_i \in V}\right) +$$

$$\mathbb{I}_{\{p_i = p_{i-1}\}} \cdot \mathbb{I}_{\{h_i = h_{i-1}\}} \cdot \left(l_i \cdot \mathbb{I}_{f_i \in F_y} \cdot \mathbb{I}_{s_i \in V} + s_i \cdot \mathbb{I}_{s_i \notin V} + c_{i-1} \cdot \mathbb{I}_{f_i \notin F_y} \cdot \mathbb{I}_{s_i \in V}\right),$$

where $\mathbb{I}_{\{\cdot\}}$ is an indicator function:

$$\mathbb{I}_{\{a=b\}} = \begin{cases} 1, if\ a = b \\ 0, else \end{cases} \qquad \mathbb{I}_{\{a\in b\}} = \begin{cases} 1, if\ a \in b \\ 0, else \end{cases}$$

5. Replace values of end dates that falls out of the follow-up interval with last follow-up date.

$$e_i = e_i \cdot \mathbb{I}_{\{e_i \le l_i\}} + l_i \cdot \mathbb{I}_{\{e_i > l_i\}}$$

6. Delete records if start date is missing and create date is greater than stop date. Reduce the dataset to the set of patients from the cohort of interest and to set of keys of selected drug(s).

$$MF^2 = \{MF^1 : p_i \in PS \ \wedge m_i \in MS \wedge \neg(b_i \in V \ \wedge c_i > e_i), i = \overline{1,n}\}$$

7. Merge the following to the PF set by patient key:

7.1) date of birth $DOB = (db_1, db_2,...,d_k)^T$

7.2) last available follow-up date within the database $L = (l_1, l_2,...l_k)^T$. The extended PF dataset would take the following form:

$$PF^1 = \{M, P, C, B, R, DOB, L\}$$

8. Replace erroneous prescription dates ($b_i \notin V \wedge (b_i < db_i \wedge b_i > l_i)$, $i = \overline{1,k}$) with missing values.

9. If number of refills is greater than pre-defined maximal number of possible refills or negative or missing, replace it with zero.

$$r_i = r_i \cdot \mathbb{I}_{\{r_i < mx\}} \cdot \mathbb{I}_{\{r_i \ge 0\}} \cdot \mathbb{I}_{\{r_i \notin V\}}, i = \overline{1,k}$$

10. Calculate end dates $E = (e_1, e_2,...,e_k)^T$ by the following rules.

10.1) if prescription date is not missing, then end date is equals to standard duration multiplied by the number of refills plus one and added to prescription date.

10.2) if prescription date is missing, then end date is equals to standard duration multiplied by the number of refills plus one and added to record creation date.

$$e_i = (e_i + (r_i + 1) \cdot sd) \cdot \mathbb{I}_{\{e_i \notin V\}} + (c_i + (r_i + 1) \cdot sd) \cdot \mathbb{I}_{\{e_i \in V\}}$$

11. Update end dates as described in Step 5 ($e_i = e_i \cdot \mathbb{I}_{\{e_i \le l_i\}} + l_i \cdot \mathbb{I}_{\{e_i > l_i\}}$, $i = \overline{1,k}$) .

12. Reduce PF^1 to the set of patients from the cohort of interest, to the set of patients not in MF^2, and to the set of keys of selected drug(s).

$$PF^2 = \{PF^1 : p_i \in PS \wedge m_i \in MS \wedge (p_i \notin P \subset MF^2), i = \overline{1,k}\}$$

13. Append both datasets by the following values: patient key, record creation date, start / prescription date and end date, assume that the new dataset *MP* contain *n*' records.

$$MF^3 = \{P, C, B, E\} \subset MF^2$$

$$PF^3 = \{P, C, B, E\} \subset PF^2$$

$$MP = MF^3 \bigcup PF^3$$

14. Calculate the number (*cn*) of distinct record creation dates for each patient, treat missing start dates by the following rules:

14.1) if *cn* is equal to one, then delete the record.

14.2) if *cn* is greater than one, replace it with record creation date.

15. Sort by patient key ascending, start date ascending within same patient key.

$$MP: \quad a) \ \ p_i \le p_{i+1}, i = \overline{1, n'}$$

$$b) \quad b_i \le b_{i+1}, \forall i_i: p_i = p_{i+1} \ \text{ - post } a) \text{ sorting}$$

16. For each unique patient key $ps_j \in PS, j = \overline{1, u}$ reduce *MP* to the set FN^j containing only $p_i = ps_j$, $i = \overline{1, n'}$. Assume that obtained dataset FN^j has n'' rows. Set $e_0 = 0$ and calculate selected medication duration for the patient avoiding double calculations of overlapping intervals.

$$FN^j: \quad d_j = \sum_{i=1}^{n''} \left((e_i - b_i) \cdot \mathbb{I}_{\{b_i \ge e_{i-1}\}} + (e_i - e_{i-1}) \cdot \mathbb{I}_{\{b_i < e_{i-1}\}} \cdot \mathbb{I}_{\{e_i \ge e_{i-1}\}} \right)$$

17. Use medication duration $D = (d_1, d_2, \ldots, d_u)^T$ to conduct further research.

3.3. "Continuous" Method

1. Repeat steps 1 and 2 from "chaining" method, then perform step 6, and treat missing values in MF^2 as described in step 14. Assume that obtained dataset MF^2 has $\tilde{n}$ instances.

2. Create stop date status variable $SI = (st_1, st_2, \ldots, st\tilde{n})^T$ on the basis of the following rules:

2.1) if active flag is "Y" and stop date is missing, then stop date status equals to 2.

2.2) if stop date is not missing, then stop date status equals to 1.

2.3) else 0.

$$st_i = 2 \cdot \mathbb{I}_{f_i \in F_y} \cdot \mathbb{I}_{s_i \in V} + 1 \cdot \mathbb{I}_{s_i \notin V} + 0 \cdot \mathbb{I}_{f_i \notin F_y} \cdot \mathbb{I}_{s_i \in V}$$

$$MF^3 = \{M, P, C, B, S, F, H, G, DOB, L, ST\}$$

3. Sort MF^3 by patient key ascending, start date descending within same patient key, stop date status ascending within the same start dates of the same patient:

$$MF^3: \quad a) \ p_i \le p_{i+1}, i = \overline{1, \tilde{n}}$$

$$b) \ b_i \ge b_{i+1}, \forall i: p_i = p_{i+1} \ \text{ - post } a) \text{ sorting}$$

$$c) \ st_i \le st_{i+1}, \forall i: b_i = b_{i+1} \ \wedge \ p_i = p_{i+1} \ \text{ - post } b) \text{ sorting}$$

4. Set initial value $p_0 = 0$ and approximate individual medication end dates $E = (e_1, e_2,...,e_{\tilde{n}})^T$.

4.1) if stop date is not missing, then end date equals to stop date.

4.2) else, if active flag is "Y", then end date equals to last follow-up date.

4.3) else, if first unique patient key, then end date equals to start date plus standard duration.

4.4) else end date equals to start date of previous record.

$$e_i = \mathbb{I}_{\{p_i \neq p_{i-1}\}} \cdot \left(l_i \cdot \mathbb{I}_{f_i \in F_y} \cdot \mathbb{I}_{s_i \in V} + s_i \cdot \mathbb{I}_{s_i \notin V} + (b_i + sd) \cdot \mathbb{I}_{f_i \notin F_y} \cdot \mathbb{I}_{s_i \in V} \right) +$$

$$+ \mathbb{I}_{\{p_i = p_{i-1}\}} \cdot \left(l_i \cdot \mathbb{I}_{f_i \in F_y} \cdot \mathbb{I}_{s_i \in V} + s_i \cdot \mathbb{I}_{s_i \notin V} + b_{i-1} \cdot \mathbb{I}_{f_i \notin F_y} \cdot \mathbb{I}_{s_i \in V} \right)$$

5. Perform step 5 from "chaining" method, and steps 7-11.

6. Reduce PF^1 to the set of patients from the cohort of interest, to the set of patients not in MF^3, and to the set of keys selected drug(s).

$$PF^2 = \left\{ PF^1 : p_i \in PS \wedge m_i \in MS \wedge (p_i \notin P \subset MF^3), i = \overline{1,k} \right\}$$

7. Treat missing values in PF^2 as described in step 14 of "chaining" method.

8. Append both datasets by the following values: patient key, record creation date, start/prescription date and end date, assume that the new dataset *MP* contain $\tilde{n}$ records.

$$MF^4 = \{P, C, B, E\} \subset MF^3$$

$$PF^3 = \{P, C, B, E\} \subset PF^2$$

$$MP = MF^4 \cup PF^3$$

9. Perform steps 15-17 from "chaining" method.

4. REMARKS

Identified erroneous entries are declared as missing in Steps 2, 8, and 9 of "chaining" method. In the Step 14, the algorithm counts the number of unique creation dates for selected drug(s) at patient level. If obtained number is greater than one, then missing start dates are replaced with record creation dates. In such a way, a patient is considered to take a particular drug if the medication records were entered in a systematic manner, otherwise the records with missing start dates are disregarded.

As an example, the prescription scenario for anti-diabetes drugs for a patient with type 2 diabetes is presented in Table **1**. The treatment was initiated with metformin (METFORMIN HCL) on the 6th of May 2009 and continued until the 25th of April 2011, when a switch to GLP-1RA (LIRAGLUTIDE) was made. With a stop date for GLP-1RA recorded on 14th of December 2011, data show a gap in the treatment till 26th of September 2012, when insulin therapy begun. However, a patient with diabetes using GLP-1RA is unlikely to have had a nine month long gap in the treatment. Indeed, careful data examination leads to the conclusion that insulin treatment started on 27th of February 2012, as would be estimated by the algorithm.

As it was mentioned earlier, MF was considered as primary data source, thus if at least one record for selected drug(s) at patient level is present in the MF, then both methods disregard entities in the PF. However, if there is no available data in MF table, the methods append data from PF.

Assessment of the first marketing date for a particular drug is an example of additional global consistency audit that is omitted in the methods' description. For instance, any start date of GLP-1RA drugs must not be prior to April 2005,

the date when first representative (Exenatide) was approved.

5. RESULTS

To evaluate the performance of described methods, we chose to focus on the estimation of the duration of treatment with two widely used anti-diabetic drugs, namely GLP-1RA and insulin. In the CEMR database 1,861,560 patients were identified as having been diagnosed with type 2 diabetes mellitus, as inferred from the assigned ICD-9 codes.

5.1. Case Study 1

As the first case study, we consider a randomly selected patient from the CEMR database, whose relevant treatment details are shown in Table **3**. The treatments with EXENATIDE and INSULIN GLARGINE started on the 18th of June 2007. The treatment with EXENATIDE was terminated on the 7th of January 2008, while INSULIN therapy continued until the last recorded follow-up date on the 24th of January 2008 (notice that the treatment is flagged as active, "Y"). In this case, the "chaining" and "continuous" methods produce the same estimates for the durations of the two treatments. Specifically, the estimates corresponding to insulin and GLP-1RA are 7.2 and 6.7 months, respectively.

Table 3. Snapshot of MF table-combining therapies. Patient's last follow-up date was identified as 24 January 2008.

GPI category 4	Medication key (M)	Patient key (P)	Create date (C)	Start date (B)	Stop date (S)	Active flag (F)	Chain ID (H)	Chain seq (G)	Enddate ("chaining")	Enddate ("continuous")
INSULIN GLARGINE	682327	15219411	18-Jun-07	18-Jun-07		N	136664321	0	20-Jun-07	20-Jun-07
EXENATIDE	12670645	15219411	18-Jun-07	18-Jun-07		N	136664552	0	17-Oct-07	15-Oct-07
INSULIN GLARGINE	1096062	15219411	20-Jun-07	20-Jun-07		N	136664321	1	7-Jan-08	7-Jan-08
EXENATIDE	12670548	15219411	17-Oct-07	15-Oct-07		N	136664552	1	7-Jan-08	7-Jan-08
INSULIN GLARGINE	1096062	15219411	7-Jan-08	7-Jan-08		Y	136664321	2	24-Jan-08	24-Jan-08
EXENATIDE	12670548	15219411	7-Jan-08	7-Jan-08	7-Jan-08	N	136664552	2	7-Jan-08	7-Jan-08

5.2. Case Study 2

As an insightful case study, we consider a patient whose relevant treatment details are shown in Table **4**. Since all of the records shown have the same chain ID it can be concluded that in the period from the 23rd of April of 2010 until the 13th of March 2013 the patient was alternating between two therapies, namely with GLP-1RA (EXENATIDE) and insulin (INSULIN GLARGINE). This example illustrates the importance of chain ID information, as readily corroborated by comparing the predicted therapy end dates using the "chaining" and "continuous" methods (per record estimates are shown in the two rightmost columns of Table **4**). The latter disregards chain ID information, it implicitly assumes that EXENATIDE was taken continuously from the 23rd of April 2010 until the 27th of April 2011, with the last prescription date being the 28th of March 2011. However, treatment with EXENATIDE was terminated on the 29th of December 2010 when a switch to insulin was made. Treatment with insulin continued until the 28th of March 2011 when a switch back to EXENATIDE appeared. This complex and frequent pattern of therapy alteration leads to vastly different treatment duration estimates when chain ID information is used ("chaining") and when it is not ("continuous"). For example, in this particular case, "continuous" approach estimates the total duration of insulin/ EXENATIDE treatment to be 5.7/ 28.9 months, compared to 26.5/ 12.1 months estimated by "chaining" method.

Table 4. Snapshot of MF table-switching between therapies. Patient's last follow up date was identified as 13 March 2013.

GPI category 4	Medication key (M)	Patient key (P)	Create date (C)	Start date (B)	Stop date (S)	Active flag (F)	Chain ID (H)	Chain seq (G)	Enddate ("chaining")	Enddate ("continuous")
EXENATIDE	1523512	64832053	23-Apr-10	23-Apr-10		N	1002923273	0	29-Dec-10	28-Mar-11
INSULIN GLARGINE	682327	64832053	29-Dec-10	29-Dec-10		N	1002923273	1	06-Jan-11	06-Jan-11
INSULIN GLARGINE	682327	64832053	06-Jan-11	06-Jan-11		N	1002923273	2	28-Mar-11	18-Dec-12

(Table 4) contd.....

GPI category 4	Medication key (M)	Patient key (P)	Create date (C)	Start date (B)	Stop date (S)	Active flag (F)	Chain ID (H)	Chain seq (G)	Enddate ("chaining")	Enddate ("continuous")
EXENATIDE	1523512	64832053	28-Mar-11	28-Mar-11		N	1002923273	3	18-Dec-12	27-Apr-11
INSULIN GLARGINE	682327	64832053	18-Dec-12	18-Dec-12		N	1002923273	4	13-Mar-13	13-Mar-13
INSULIN GLARGINE	682327	64832053	13-Mar-13	13-Mar-13		Y	1002923273	5	13-Mar-13	13-Mar-13

5.3. General Analysis

Given our focus on GLP-1RA and insulin, to facilitate further analysis, from the cohort of all T2DM patients we selected those who at any point in their medical history received treatment with either of the two drugs of interest. Text mining of drug names in MD table revealed various insulin regimens as well as related devices (e.g. insulin syringe). To quantify the result, we found that approximately 30% of the patients in the T2DM cohort received at least one prescription for insulin drug. Interestingly, a large number of patients (~25,000) were found to have received prescriptions for insulin devices but not for insulin therapy itself. Further exploration on these patients revealed that the average duration of use of these devices in this patient group was 21 months (Table **5**), strongly suggesting that there was an accompanying insulin therapy which was not recorded in the stored EMRs. This conclusion is further corroborated by the finding that the mean glycated haemoglobin (HbA1c) level for these patients was measured to be 7.8% on the date of the first record associated with the device.

Table 5. Summary statistics on the estimated duration in months of treatment with specific medications in T2DM cohort (n=1,861,560) by "chaining" and "continuous" methods, and the difference in the estimated duration between "chaining" and "continuous" methods.

	"Chaining" method				"Continuous" method				"Chaining" - "continuous"			
	n (%)	Mean (sd)	(min, max)	Median (IQR)	n (%)	Mean (sd)	(min, max)	Median (IQR)	n (%)	Mean (sd)	(min, max)	Median (IQR)
Insulin + device	588923 (32)	32.5 (35)	(0, 657.8)	21.6 (6.5, 46.8)	591441 (32)	32.7 (34.9)	(0, 657.8)	21.8 (6.3, 47.4)	588923 (32)	-0.2 (4.8)	(-167.8, 183.4)	0 (0, 0)
Insulin only	563293 (30)	32.0 (34.9)	(0, 657.8)	20.8 (6.1, 45.8)	566014 (30)	32.2 (34.8)	(0, 657.8)	21.0 (6, 46.5)	563293 (30)	-0.3 (5)	(-167.8, 176.9)	0 (0, 0)
no Insulin, but device	25536 (1)	21.2 (21.5)	(0, 196.8)	14.3 (4.8, 30.9)	25386 (1)	21.2 (21.9)	(0, 190.7)	14.1 (4.4, 31.1)	24910 (1)	-0.2 (5)	(-131.8, 183.4)	0 (0, 0)
GLP1RA	113416 (6)	18.3 (19.4)	(0, 110.7)	11.7 (3.9, 26)	114316 (6)	19.2 (21.0)	(0, 111.7)	11.7 (3.5, 27.4)	113416 (6)	-1.0 (7.6)	(-103.9, 95.4)	0 (0, 0)
Exenatide	73326 (4)	18.8 (20.2)	(0, 110.7)	11.6 (3.9, 26.5)	74060 (4)	18.8 (21.4)	(0, 111.2)	10.6 (3.1, 26.7)	73326 (4)	-0.2 (8.4)	(-97.0, 95.4)	0 (0, 0)
Liraglutide	56406 (3)	12.5 (11.9)	(0, 56.2)	8.6 (3, 19)	56907 (3)	12.7 (12.4)	(0, 56.2)	8.3 (2.5, 19.5)	56406 (3)	-0.3 (4.0)	(-49.5, 47.5)	0 (0, 0)
Albiglutide	14 (0)	1.3 (0.5)	(1, 2.4)	1 (1, 1.9)	15 (0)	1.3 (0.5)	(1, 2.4)	1.0 (1, 1.9)	14 (0)	0 (0)	(0, 0)	0 (0, 0)
In patients with treatment duration ≥2 Months												
Insulin + device	518000 (28)	36.8 (35.2)	(2, 657.8)	26.4 (11.1, 51.6)	518318 (28)	37.1 (35.0)	(2, 657.8)	26.8 (11.2, 52.3)	516808 (28)	-0.3 (4.9)	(-167.8, 176.9)	0 (0, 0)
Insulin only	492992 (26)	36.4 (35.2)	(2, 657.8)	25.8 (10.7, 50.8)	493494 (27)	36.7 (35.1)	(2, 657.8)	26.3 (10.9, 51.6)	491847 (26)	-0.4 (5.2)	(-167.8, 176.9)	0 (0, 0)
no Insulin, but device	22085 (1)	24.3 (21.5)	(2, 196.8)	17.8 (8, 34.1)	21628 (1)	24.7 (21.9)	(2, 190.7)	18.0 (8, 34.8)	21342 (1)	-0.5 (4.1)	(-131.8, 65.3)	0 (0, 0)
GLP1RA	96458 (5)	21.3 (19.6)	(2, 110.7)	14.9 (6.8, 29.3)	94972 (5)	22.9 (21.3)	(2, 111.7)	15.7 (6.9, 31.8)	94372 (5)	-1.5 (7.8)	(-103.9, 95.4)	0 (0, 0)
Exenatide	62538 (3)	21.8 (20.4)	(2, 110.7)	14.7 (6.6, 30.4)	60228 (3)	22.9 (21.7)	(2, 111.2)	15.0 (6.5, 32.1)	59812 (3)	-0.8 (8.0)	(-97.0, 95.4)	0 (0, 0)
Liraglutide	45432 (2)	15.3 (11.6)	(2, 56.2)	12 (5.8, 22.1)	44344 (2)	16 (12.2)	(2, 56.2)	12.5 (5.9, 23.4)	43991 (2)	-0.6 (3.9)	(-49.5, 43.9)	0 (0, 0)
Albiglutide	2 (0)	2.2 (0.2)	(2.1, 2.4)	2.2 (2.1, 2.4)	2 (0)	2.2 (0.2)	(2.1, 2.4)	2.2 (2.1, 2.4)	2 (0)	0 (0)	(0, 0)	0 (0, 0)

The number of patients receiving insulin and GLP-1RA, and the corresponding treatment duration estimates (in months) produced by our algorithms ("chaining" and "continuous"), are summarized in Table **5**. Different insulin regimens were treated jointly, as we found that any finer level of detail is poorly recorded in the database. As regards to

GLP-1RA treatment, only three different GLP-1RA drugs (namely, Exenatide, Liraglutide, and Albiglutide) have been used. Being new to the market (introduced in 2014), only limited data was available for Albiglutide treatment.

The estimate of the proportion of patients identified as having received specific individual drugs was found to be very similar using both the "chaining" approach, as well as the non-chain ID based alternative "continuous" approach, as shown in Table **5**. The corresponding values of the key statistics – namely the mean, standard deviation (SD), median, and the interquartile range (IQR)- of the respective estimates of the duration of treatment with individual drugs were also similar. The average differences in the estimated duration of treatment with insulin only and GLP-1RA drugs were 0.3 month and 1 month respectively. There were no differences at the median levels. Separate analyses for patients with minimum 2 months of treatment duration with individual therapies also revealed the same results. However, it is important to note that although the cumulative statistics of the estimated treatment durations with different therapies were not significantly different, we did find notable differences in the minimum and maximum duration estimates for specific patient subgroups, as evident from (Table **5**).

6. DISCUSSION

In this work we addressed a number of challenging data mining related issues while extracting patient-level longitudinal information on prescription patterns and medication usages from large relational databases (our data set comprises more than a billion records). There are several key contributions of note. Firstly we identified the specific challenges which automatic methods must deal with in the processing of this complex voluminous data. We corroborated our arguments using analysis of real-world EMRs and discussed the importance and the implications of being able to handle erroneous and incomplete longitudinal information. Secondly, we introduced two methods for the estimation of the duration of treatment with specific drug(s) in the presence of the aforementioned challenges. Developed sequentially ordered case by case rules were presented mathematically. To the best of our knowledge, no robust algorithmic approach has yet been reported to evaluate treatment duration with individual medications in multiple treatment scenario [22, 27].

We have described two algorithmic approaches to estimate treatment duration on the individual record level. First method ("chaining") relies on specific chaining fields of medication information, while second approach ("continuous") does not use chain related information and employs only chronological record information instead. Our results on the large Centricity EMR database show that the two approaches do not produce significantly different results on average at population level. However, when examined in detail, the "chaining" method could identify the treatment alterations longitudinally and was shown to be more robust at individual patient level. Furthermore, treatment duration estimates from the "continuous" approach are more sensible to the set of selected medications. The difference between methods is particularly prominent in studies involving multiple drugs as opposed to single drug therapies or focusing on the order of treatment initiation [48, 49].

Our study highlighted the potential risk of underestimating the duration of treatment when EMR data is used directly, due to erroneous or incomplete data emerging from omissions in the data entry process, appointments missed by patients, typographical errors, or numerous others. Both proposed algorithms robustly handle these challenges whenever is possible, estimating values of the missing or erroneous entries. Importantly, being rule based, the decisions of our algorithms are readily interpretable by humans and lend themselves to effortless use by medical professionals not necessarily proficient in data mining and related disciplines. Both approaches implement two fact datasets available in the Centricity EMRs, however algorithms are easily adjusted in case of only one available dataset.

CONCLUSION

This study discusses the challenges in exploring the prescription / medication patterns for individual patients in large primary / ambulatory care electronic databases, and introduces two algorithmic approaches for robust estimation of treatment duration with individual drug(s). We have demonstrated that implementing chaining fields of medication information additionally improve the quality of estimates. Given the importance of extracting medication information appropriately in pharmaco-epidemiological studies based on real world data, the proposed algorithms has the potential to significantly contribute to the analytical quality aspects in the future EMR based clinical and epidemiological studies.

LIST OF ABBREVIATIONS

EMR	=	Electronic Medical Rerecords
CEMR	=	Centricity Electronic Medical Rerecords

T2DM	=	Type 2 Diabetes
MD	=	Medication Dimension
MF	=	Medication Fact
PF	=	Prescription Fact
GPI	=	Generic Product Identifier
ID	=	Identification
DOB	=	Date of Birth
SD	=	Standard Deviation
IQR	=	Interquartile Range
GLP-1RA	=	Glucagon-Like Peptide-1 Receptor Agonist

ETHICS APPROVAL AND CONSENT TO PARTICIPATE

Not applicable.

HUMAN AND ANIMAL RIGHTS

No Animals/Humans were used for studies that are base of this research.

CONSENT FOR PUBLICATION

Not applicable.

CONFLICT OF INTEREST

Sanjay K. Paul (SKP) has acted as a consultant and/or speaker for Novartis, GI Dynamics, Roche, AstraZeneca, Guangzhou Zhongyi Pharmaceutical and Amylin Pharmaceuticals LLC. He has received grants in support of investigator and investigator initiated clinical studies from Merck, Novo Nordisk, AstraZeneca, Hospira, Amylin Pharmaceuticals, Sanofi-Avensis and Pfizer. Olga Montvida (OM) and Ognjen Arandjelovic (OA) has no conflict of interest to declare. Edward Reiner (ER) was an employee of Quintiles and was responsible for the strategic development of the Centricity EMR database.

ACKNOWLEDGEMENTS

Olga Montvida (OM) and Sanjay K. Paul (SKP) conceived the idea and were responsible for the primary design of the study and the methodological developments. Ognjen Arandjelovic (OA) and Edward Reiner (ER) evaluated the methodological approach. Olga Montvida (OM) conducted the data extraction and statistical analyses. The first draft of the manuscript was developed by Sanjay K. Paul (SKP) and Olga Montvida (OM), and all authors contributed to the finalization of the manuscript. Sanjay K. Paul (SKP) had full access to all the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

Melbourne EpiCentre gratefully acknowledges the support from the Australian Government's National Collaborative Research Infrastructure Strategy (NCRIS) initiative through Therapeutic Innovation Australia. No separate funding was obtained for this study.

REFERENCES

[1] Paul SK, Klein K, Thorsted BL, Wolden ML, Khunti K. Delay in treatment intensification increases the risks of cardiovascular events in patients with type 2 diabetes. Cardiovasc Diabetol 2015; 14: 100.
[http://dx.doi.org/10.1186/s12933-015-0260-x] [PMID: 26249018]

[2] Bhatnagar P, Wickramasinghe K, Williams J, Rayner M, Townsend N. The epidemiology of cardiovascular disease in the UK 2014. Heart 2015; 101(15): 1182-9.
[http://dx.doi.org/10.1136/heartjnl-2015-307516] [PMID: 26041770]

[3] Crawford AG, Cote C, Couto J, *et al.* Comparison of GE Centricity Electronic Medical Record database and National Ambulatory Medical Care Survey findings on the prevalence of major conditions in the United States. Popul Health Manag 2010; 13(3): 139-50.
[http://dx.doi.org/10.1089/pop.2009.0036] [PMID: 20568974]

[4] Wettermark B, Zoëga H, Furu K, *et al.* The Nordic prescription databases as a resource for pharmacoepidemiological research--a literature review. Pharmacoepidemiol Drug Saf 2013; 22(7): 691-9.
[http://dx.doi.org/10.1002/pds.3457] [PMID: 23703712]

[5] Lau EC, Mowat FS, Kelsh MA, *et al.* Use of electronic medical records (EMR) for oncology outcomes research: assessing the comparability of EMR information to patient registry and health claims data. Clin Epidemiol 2011; 3: 259-72.
[PMID: 22135501]

[6] Paul SK, Klein K, Maggs D, Best JH. The association of the treatment with glucagon-like peptide-1 receptor agonist exenatide or insulin with cardiovascular outcomes in patients with type 2 diabetes: A retrospective observational study. Cardiovasc Diabetol 2015; 14: 10.
[http://dx.doi.org/10.1186/s12933-015-0178-3] [PMID: 25616979]

[7] Nadkarni PM. Drug safety surveillance using de-identified EMR and claims data: issues and challenges. J Am Med Inform Assoc 2010; 17(6): 671-4.
[http://dx.doi.org/10.1136/jamia.2010.008607] [PMID: 20962129]

[8] Liu M, McPeek Hinz ER, Matheny ME, *et al.* Comparative analysis of pharmacovigilance methods in the detection of adverse drug reactions using electronic medical records. J Am Med Inform Assoc 2013; 20(3): 420-6.
[http://dx.doi.org/10.1136/amiajnl-2012-001119] [PMID: 23161894]

[9] Coloma PM, Trifirò G, Patadia V, Sturkenboom M. Postmarketing safety surveillance: where does signal detection using electronic healthcare records fit into the big picture? Drug Saf 2013; 36(3): 183-97.
[http://dx.doi.org/10.1007/s40264-013-0018-x] [PMID: 23377696]

[10] Lin J, Jiao T, Biskupiak JE, McAdam-Marx C. Application of electronic medical record data for health outcomes research: a review of recent literature. Expert Rev Pharmacoecon Outcomes Res 2013; 13(2): 191-200.
[http://dx.doi.org/10.1586/erp.13.7] [PMID: 23570430]

[11] Belletti D, Zacker C, Mullins CD. Perspectives on electronic medical records adoption: Electronic Medical Records (EMR) in outcomes research. Patient Relat Outcome Meas 2010; 1: 29-37.
[http://dx.doi.org/10.2147/PROM.S8896] [PMID: 22915950]

[12] Khunti K, Davies M, Majeed A, Thorsted BL, Wolden ML, Paul SK. Hypoglycemia and risk of cardiovascular disease and all-cause mortality in insulin-treated people with type 1 and type 2 diabetes: a cohort study. Diabetes Care 2015; 38(2): 316-22.
[http://dx.doi.org/10.2337/dc14-0920] [PMID: 25492401]

[13] Canavan C, West J, Card T. Calculating Total Health Service Utilisation and Costs from Routinely Collected Electronic Health Records Using the Example of Patients with Irritable Bowel Syndrome Before and After Their First Gastroenterology Appointment. Pharmacoeconomics 2016; 34(2): 181-94.
[PMID: 26497004]

[14] Bessou A, Guelfucci F, Aballea S, Toumi M, Poole C. Comparison of comorbidity measures to predict economic outcomes in a large UK primary care database. Value Health. 2015; 18(7): A691.
[http://dx.doi.org/10.1016/j.jval.2015.09.2565]

[15] Birkhead GS, Klompas M, Shah NR. Uses of electronic health records for public health surveillance to advance public health. Annu Rev Public Health 2015; 36: 345-59.
[http://dx.doi.org/10.1146/annurev-publhealth-031914-122747] [PMID: 25581157]

[16] Paul MM, Greene CM, Newton-Dame R, *et al.* The state of population health surveillance using electronic health records: a narrative review. Popul Health Manag 2015; 18(3): 209-16.
[http://dx.doi.org/10.1089/pop.2014.0093] [PMID: 25608033]

[17] Kukafka R, Ancker JS, Chan C, *et al.* Redesigning electronic health record systems to support public health. J Biomed Inform 2007; 40(4): 398-409.
[http://dx.doi.org/10.1016/j.jbi.2007.07.001] [PMID: 17632039]

[18] Menachemi N, Collum TH. Benefits and drawbacks of electronic health record systems. Risk Manag Healthc Policy 2011; 4: 47-55.
[http://dx.doi.org/10.2147/RMHP.S12985] [PMID: 22312227]

[19] Crapo J. Big data in gealthcare: separating the hype from the reality. HealthCatalyst 2015; p. 5.

[20] Grabenbauer L, Skinner A, Windle J. Electronic Health Record Adoption - Maybe It's not about the Money: Physician Super-Users, Electronic Health Records and Patient Care. Appl Clin Inform 2011; 2(4): 460-71.
[http://dx.doi.org/10.4338/ACI-2011-05-RA-0033] [PMID: 23616888]

[21] Paul SK, Klein K, Majeed A, Khunti K. Association of smoking and concomitant metformin use with cardiovascular events and mortality in people newly diagnosed with type 2 diabetes. J Diabetes 2016; 8(3): 354-62.
[http://dx.doi.org/10.1111/1753-0407.12302] [PMID: 25929583]

[22] Gaitanou P, Garoufallou E, Balatsoukas P. The effectiveness of big data in health care: a systematic review. Commun Comput Inf Sci 2014; 141-53.
[http://dx.doi.org/10.1007/978-3-319-13674-5_14]

[23] Svensson MK, Cederholm J, Eliasson B, Zethelius B, Gudbjörnsdottir S. Albuminuria and renal function as predictors of cardiovascular events and mortality in a general population of patients with type 2 diabetes: a nationwide observational study from the Swedish National Diabetes Register. Diab Vasc Dis Res 2013; 10(6): 520-9.
[http://dx.doi.org/10.1177/1479164113500798] [PMID: 24002670]

[24] Dean BB, Lam J, Natoli JL, Butler Q, Aguilar D, Nordyke RJ. Review: use of electronic medical records for health outcomes research: a

literature review. Med Care Res Rev 2009; 66(6): 611-38.
[http://dx.doi.org/10.1177/1077558709332440] [PMID: 19279318]

[25] Wei WQ, Denny JC. Extracting research-quality phenotypes from electronic health records to support precision medicine. Genome Med 2015; 7(1): 41.
[http://dx.doi.org/10.1186/s13073-015-0166-y] [PMID: 25937834]

[26] Denny JC. Chapter 13: Mining electronic health records in the genomics era. PLOS Comput Biol 2012; 8(12): e1002823.
[http://dx.doi.org/10.1371/journal.pcbi.1002823] [PMID: 23300414]

[27] Torre C, Martins AP. Overview of Pharmacoepidemiological Databases in the Assessment of Medicines Under real-life Conditions. In: Lunet N, Eds. Epidemiolgy-current perspective on Research and practical Intech open publishers contributers 2012; pp.131-54.
[http://dx.doi.org/10.5772/35318]

[28] Centricity Electronic Medical Record Brochure. GE Healthcare 2011.

[29] Lin J, Jiao T, Biskupiak JE, McAdam-Marx C. Application of electronic medical record data for health outcomes research: a review of recent literature. Expert Rev Pharmacoecon Outcomes Res 2013; 13(2): 191-200.
[http://dx.doi.org/10.1586/erp.13.7] [PMID: 23570430]

[30] Jermyn P, Dixon M, Read BJ. Preparing clean views of data for data mining. ERCIM Work on Database Res 1999; pp. 1-15.

[31] Zhang S, Zhang C, Yang Q. Data preparation for data mining. Appl Artif Intell 2003; 17(5-6): 375-81.
[http://dx.doi.org/10.1080/713827180]

[32] Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: towards better research applications and clinical care. Nat Rev Genet 2012; 13(6): 395-405.
[http://dx.doi.org/10.1038/nrg3208] [PMID: 22549152]

[33] Benchimol EI, Smeeth L, Guttmann A, *et al.* The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) statement. PLoS Med 2015; 12(10): e1001885.
[http://dx.doi.org/10.1371/journal.pmed.1001885] [PMID: 26440803]

[34] PLOS Medicine Editors. From checklists to tools: Lowering the barrier to better research reporting. PLoS Med 2015; 12(11): e1001910.
[http://dx.doi.org/10.1371/journal.pmed.1001910] [PMID: 26600090]

[35] Yao L, Zhang Y, Li Y, Sanseau P, Agarwal P. Electronic health records: Implications for drug discovery. Drug Discov Today 2011; 16(13-14): 594-9.
[http://dx.doi.org/10.1016/j.drudis.2011.05.009] [PMID: 21624499]

[36] Hall GC, McMahon AD, Dain M-P, Home PD. A comparison of duration of first prescribed insulin therapy in uncontrolled type 2 diabetes. Diabetes Res Clin Pract 2011; 94(3): 442-8.
[http://dx.doi.org/10.1016/j.diabres.2011.09.003] [PMID: 21963105]

[37] Hansen RA, Farley JF, Maciejewski ML, Ye X, Qian C, Powers B. Real-world utilization patterns and outcomes of colesevelam hcl in the ge electronic medical record. BMC Endocr Disord 2013; 13(1): 24.
[http://dx.doi.org/10.1186/1472-6823-13-24] [PMID: 23866087]

[38] Hippisley-Cox J, Coupland C. 2016.

[39] Fardet L, Petersen I, Nazareth I. Prevalence of long-term oral glucocorticoid prescriptions in the UK over the past 20 years. Rheumatology (Oxford) 2011; 50(11): 1982-90.
[http://dx.doi.org/10.1093/rheumatology/ker017] [PMID: 21393338]

[40] Davis KL, Tangirala M, Meyers JL, Wei W. Real-world comparative outcomes of US type 2 diabetes patients initiating analog basal insulin therapy. Curr Med Res Opin 2013; 29(9): 1083-91.
[http://dx.doi.org/10.1185/03007995.2013.811403] [PMID: 23734906]

[41] Xie L, Wei W, Pan C, Du J, Baser O. A real-world study of patients with type 2 diabetes initiating basal insulins *via* disposable pens. Adv Ther 2011; 28(11): 1000-11.
[http://dx.doi.org/10.1007/s12325-011-0074-5] [PMID: 22038703]

[42] Deléger L, Grouin C, Zweigenbaum P. Extracting medical information from narrative patient records: the case of medication-related information. J Am Med Inform Assoc 2010; 17(5): 555-8.
[http://dx.doi.org/10.1136/jamia.2010.003962] [PMID: 20819863]

[43] Etminan M. Reporting guidelines for pharmacoepidemiological studies are urgently needed. BMJ 2014; 349: g5511.
[http://dx.doi.org/10.1136/bmj.g5511] [PMID: 25231185]

[44] Kamal KM, Chopra I, Elliott JP, Mattei TJ. Use of electronic medical records for clinical research in the management of type 2 diabetes. Res Social Adm Pharm 2014; 10(6): 877-84.
[http://dx.doi.org/10.1016/j.sapharm.2014.01.001] [PMID: 24556384]

[45] Herrin J, da Graca B, Nicewander D, *et al.* The effectiveness of implementing an electronic health record on diabetes care and outcomes. Health Serv Res 2012; 47(4): 1522-40.
[http://dx.doi.org/10.1111/j.1475-6773.2011.01370.x] [PMID: 22250953]

[46] Holt TA, Stables D, Hippisley-Cox J, O'Hanlon S, Majeed A. Identifying undiagnosed diabetes: cross-sectional survey of 3.6 million patients'

electronic records. Br J Gen Pract 2008; 58(548): 192-6.
[http://dx.doi.org/10.3399/bjgp08X277302] [PMID: 18318973]

[47] Davis KL, Tangirala M, Meyers JL, Wei W. Real-world comparative outcomes of US type 2 diabetes patients initiating analog basal insulin therapy. Curr Med Res Opin 2013; 29(9): 1083-91.
[http://dx.doi.org/10.1185/03007995.2013.811403] [PMID: 23734906]

[48] Paul SK, Klein K, Majeed A, Khunti K. Association of smoking and concomitant use of metformin with cardiovascular events and mortality in people newly diagnosed with type 2 diabetes. J Diabetes 2015; 8(3): 354-62.
[PMID: 25929583]

[49] Paul SK, Klein K, Maggs D, Best JH. The association of the treatment with glucagon-like peptide-1 receptor agonist exenatide or insulin with cardiovascular outcomes in patients with type 2 diabetes: a retrospective observational study. Cardiovasc Diabetol 2015; 14(1): 10.
[http://dx.doi.org/10.1186/s12933-015-0178-3] [PMID: 25616979]