Validating drug repurposing signals using electronic health records: a case study of metformin associated with reduced cancer mortality

Hua Xu,[1] Melinda C Aldrich,[2,3] Qingxia Chen,[4,5] Hongfang Liu,[6] Neeraja B Peterson,[7] Qi Dai,[3] Mia Levy,[5,7] Anushi Shah,[5] Xue Han,[4] Xiaoyang Ruan,[6] Min Jiang,[1] Ying Li,[8] Jamii St Julien,[2] Jeremy Warner,[5,7] Carol Friedman,[8] Dan M Roden,[7,9] Joshua C Denny[5,7]

▶ Additional material is published online only. To view please visit the journal online (http://dx.doi.org/10.1136/amiajnl-2014-002649).

For numbered affiliations see end of article.

Correspondence to
Dr Hua Xu, The University of Texas School of Biomedical Informatics at Houston, 7000 Fannin St, Suite 600, Houston, TX 77030, USA; hua.xu@uth.tmc.edu

MCA and HX contributed equally to the work.

Received 15 January 2014
Revised 10 June 2014
Accepted 3 July 2014

ABSTRACT

Objectives Drug repurposing, which finds new indications for existing drugs, has received great attention recently. The goal of our work is to assess the feasibility of using electronic health records (EHRs) and automated informatics methods to efficiently validate a recent drug repurposing association of metformin with reduced cancer mortality.

Methods By linking two large EHRs from Vanderbilt University Medical Center and Mayo Clinic to their tumor registries, we constructed a cohort including 32 415 adults with a cancer diagnosis at Vanderbilt and 79 258 cancer patients at Mayo from 1995 to 2010. Using automated informatics methods, we further identified type 2 diabetes patients within the cancer cohort and determined their drug exposure information, as well as other covariates such as smoking status. We then estimated HRs for all-cause mortality and their associated 95% CIs using stratified Cox proportional hazard models. HRs were estimated according to metformin exposure, adjusted for age at diagnosis, sex, race, body mass index, tobacco use, insulin use, cancer type, and non-cancer Charlson comorbidity index.

Results Among all Vanderbilt cancer patients, metformin was associated with a 22% decrease in overall mortality compared to other oral hypoglycemic medications (HR 0.78; 95% CI 0.69 to 0.88) and with a 39% decrease compared to type 2 diabetes patients on insulin only (HR 0.61; 95% CI 0.50 to 0.73). Diabetic patients on metformin also had a 23% improved survival compared with non-diabetic patients (HR 0.77; 95% CI 0.71 to 0.85). These associations were replicated using the Mayo Clinic EHR data. Many site-specific cancers including breast, colorectal, lung, and prostate demonstrated reduced mortality with metformin use in at least one EHR.

Conclusions EHR data suggested that the use of metformin was associated with decreased mortality after a cancer diagnosis compared with diabetic and non-diabetic cancer patients not on metformin, indicating its potential as a chemotherapeutic regimen. This study serves as a model for robust and inexpensive validation studies for drug repurposing signals using EHR data.

To cite: Xu H, Aldrich MC, Chen Q, *et al*. *J Am Med Inform Assoc* Published Online First: [*please include* Day Month Year] doi:10.1136/amiajnl-2014-002649

INTRODUCTION

The pharmaceutical industry faces a productivity problem to smoothly deliver new drugs to market. Current de novo drug discovery and development is costly, time-consuming, and risky.[1] Developing a new drug is estimated to cost over US$800 million and to take anywhere from 10 to 17 years,[2] with a success rate of less than 10%.[3] Therefore, pharmaceutical companies and public-sector researchers are both seeking more creative methods for drug discovery. Recently, drug repurposing (also called repositioning or re-profiling), which finds new indications for existing drugs, has received great attention.[1 4–7] Drug candidates for repurposing have often been through the pre-clinical and clinical stages and, therefore, have known safety profiles which can substantially reduce the risk, cost, and time of drug development, which offers the possibility of solving the productivity dilemma. Successful stories of drug repurposing have been reported[1] and the need for drug repurposing is well recognized by leaders in industry, academia, and government.[8] For example, The Learning Collaborative[9] aims to advance therapies for blood cancers through developing a drug repurposing framework across different organizations.

Recently, there has been a growing effort to develop computational approaches to predict drug repurposing associations.[10 11] With the availability of comprehensive compound databases containing structure, bioassay, and genomic information, such as NIH's Molecular Libraries Initiative,[12 13] new computational methods that utilize high-throughput data to predict drug repurposing signals have been developed, including structure-based virtual screening,[14] and analysis of side effect profiles,[15 16] genomic and gene expression data,[5 17 18] and the biomedical literature.[19] More and more potential drug repurposing signals are being predicted; however, how to further validate these potential signals and determine the appropriate next steps (eg, to conduct a clinical trial or not) remains challenging. Here we propose the use of large electronic health record (EHR) databases to validate potential drugs for repurposing. Over the past decade, rapid growth in the clinical implementation of large EHRs has led to an unprecedented expansion in the availability of dense longitudinal clinical datasets of large populations, which are ideal for quantifying drug outcome. Large EHRs have emerged as a valuable resource, enabling clinical and translational research,[20 21] including drug outcome studies, for instance for pharmacovigilance.[22–25] Moreover, informatics approaches that can efficiently and accurately extract and analyze

clinical information from heterogeneous data sources within EHR databases have also been developed and applied to facilitate cost-effective clinical studies using EHRs.[26–30]

As a first step to assess the use of EHRs for drug repurposing, we conducted a study to validate a recently reported association of metformin, a first-line therapy for type 2 diabetes mellitus (DM2), with reduced cancer mortality. A growing body of evidence suggests metformin improves cancer survival[31 32] and decreases cancer risk[33–36] when compared to other glucose-lowering therapies, suggesting metformin may have clinical promise as an antineoplastic agent.[33 37] A recent study of incident cancer patients from primary care clinics in the UK showed that metformin was associated with reduced mortality compared with cancer patients not exposed to metformin.[32] Specific cancers, such as pancreatic or colorectal cancer, may have improved survival with metformin use especially for early stage disease.[38 39] As a result, metformin is being evaluated for use as a cancer therapeutic agent[40 41] and requires confirmation in an independent clinical setting.

In this study, we used two state-of-the-art EHR databases at Vanderbilt University Medical Center (VUMC) and Mayo Clinic to conduct a retrospective cohort study to evaluate the association between metformin and overall mortality among incident cancer cases. The purpose of our study was twofold: (1) to validate the association between metformin use and cancer mortality using comprehensive EHRs; and (2) to demonstrate the use of informatics tools in automated data extraction tasks for EHR-based drug repurposing studies. To the best of our knowledge, this is the first study that aims to apply EHR data to drug repurposing research.

METHODS

Data sources

We conducted a retrospective cohort study from January 1, 1995 to December 31, 2010 using the EHRs at VUMC and Mayo Clinic. At VUMC, the Synthetic Derivative (SD), a comprehensive and de-identified image of the EHR at VUMC,[42] was used for this study. The SD is updated regularly as new clinical information, including inpatient and outpatient billing codes, laboratory values, laboratory reports, medication orders, and clinical notes, is accrued in the EHR. As of May 2013, the SD contained information on about 2.2 million individuals with dense electronic medical record data dating back to the early 1990s, while the Mayo Clinic EHR contained information on about 7.4 million patients.

Patients were eligible for the study if: (1) they had an incident cancer diagnosis (excluding non-melanoma skin cancers because they have a much better prognosis than other types of cancers) between January 1, 1995 and December 31, 2010 identified using the Vanderbilt tumor registry which is linked to the Vanderbilt EHR; and (2) were aged 18 years or older at the time of tumor diagnosis. Cancer patients were identified using ICD-O (International Classification of Diseases for Oncology) codes and their corresponding date of diagnosis in the Vanderbilt tumor registry, which was initiated in the early 1980s and is regularly maintained by trained nurse abstractors for all cancer patients diagnosed or with their first course of treatment at Vanderbilt. We included only the first incident cancer in individuals having multiple primary tumors. We excluded patients with congestive heart failure (CHF) or chronic kidney disease (CKD) prior to tumor diagnosis, resulting in a total of 42 165 cancer patients in this study, since heart failure and kidney disease are considered contraindications for metformin use. CHF was excluded by removing patients with an ICD-9 code of 428.* at any point before the date of tumor diagnosis and CKD was excluded by removing patients with a creatinine level >1.5 mg/dL before the tumor diagnosis date. (Since CHF and CKD can both occur as complications from cancer treatment, we did not remove patients who developed these conditions after cancer diagnosis.)

From the date of their cancer diagnosis, patients were followed for overall mortality. Mortality status was assessed by linkage with the local tumor registry. For example, the Vanderbilt tumor registry follows the NAACCR (North American Association of Central Cancer Registries) Death Clearance Manual[43] when ascertaining death information for cancer patients.

Study design and data extraction

Figure 1 shows the overall design and workflow of this study. Four exposure groups were identified among the Vanderbilt cancer patients based on DM2 disease status and medication status following their cancer diagnosis. The four exposure groups were as follows: (1) DM2 patients using metformin (including patients exposed to both metformin and other DM2 drugs); (2) DM2 patients using other oral hypoglycemic medications (and never metformin); (3) DM2 patients using insulin only; and (4) non-diabetic patients with no use of diabetes drugs. Construction of the study cohort, identification of exposed/unexposed individuals, and ascertainment of covariates was done automatically by using existing or newly developed EHR selection algorithms[44] incorporating techniques such as natural language processing (NLP).

To identify DM2 patients, we used an existing algorithm[45 46] previously developed by the eMERGE (electronic Medical Records and Genomics) Network.[47] The algorithm identifies patients with and without diabetes using three types of clinical information: (1) ICD-9 codes for DM2; (2) medications for DM2; and (3) clinical laboratory results (glucose >200 mg/dL or hemoglobin A1c >6.5%). DM2 individuals met at least two of the three requirements for diagnosis, while non-diabetic patients had none of the three criteria in their records. Prior research has demonstrated that this algorithm has a positive predictive value (PPV) of 98% for DM2 and a PPV of 100% for non-diabetes.[45] Patients not meeting either DM2 or non-diabetic definitions were excluded (eg, a single ICD-9 code for diabetes without other supporting evidence; N=7452).

Diabetic individuals were divided into three exposure groups based on medication use after their cancer diagnosis. To identify metformin and other DM2 drug exposure, we used both structured (eg, electronic physician orders) and unstructured (eg, clinic visit notes) medication information in the EHR. MedEx,[48 49] an existing high performance NLP system, was used to extract medication names and signature information from unstructured clinical text. We required that subjects have two or more mentions of the diabetes medications in the EHR and at least one mention within 5 years after their cancer diagnosis date to classify subjects by medication use; subjects lacking this information were excluded (N=2298). Metformin medications included monotherapy and combination therapy, such as metformin with thiazolidinediones (eg, Actoplus Met or PrandiMet). Cancer patients without DM2 were included as an additional unexposed comparison group.

Clinical covariates were selected a priori and included patient age at cancer diagnosis, sex, race, body mass index (BMI), insulin use, tobacco use, tumor type, and tumor stage. Some covariates were found in structured fields in the EHR or the Vanderbilt tumor registry (eg, age at diagnosis, tumor type, and

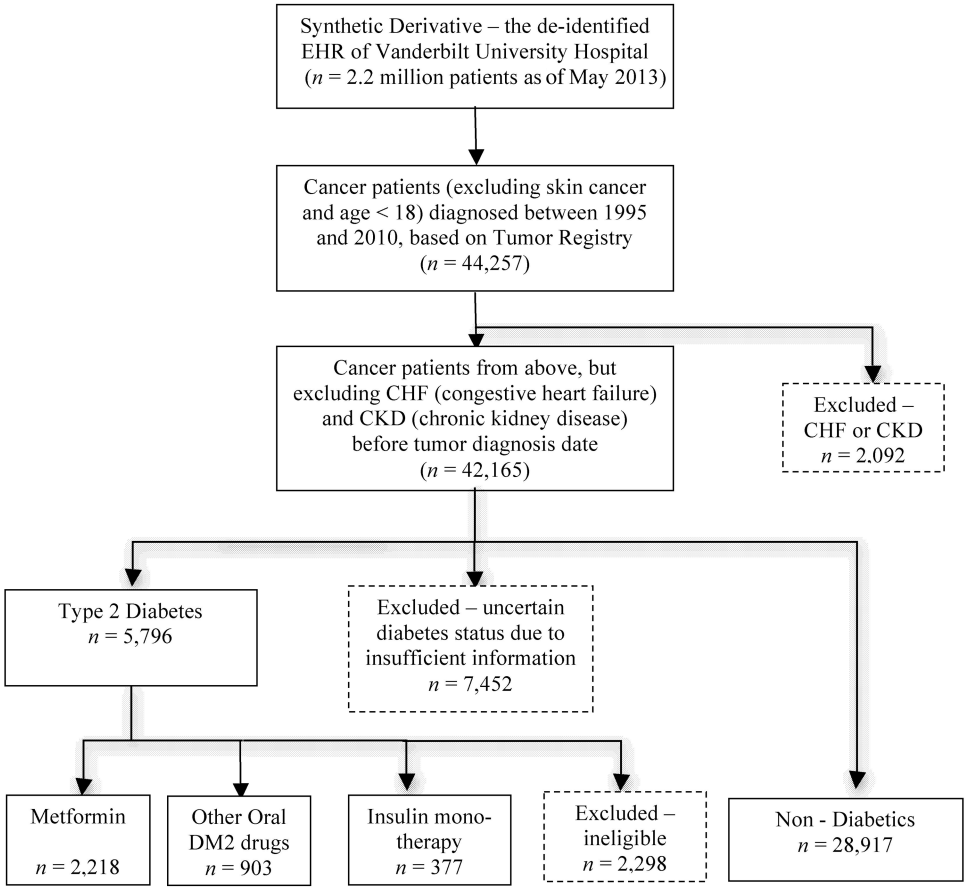

Figure 1 The study design and data extraction workflow for patients in the Vanderbilt electronic health record (EHR) system from January 1995 to December 2010.

tumor stage). For covariates that were not available in structured formats, we used NLP algorithms to extract the information from clinical narratives. To determine tobacco use, we utilized a recently developed smoking status extraction algorithm, which achieved a PPV of 93% for determining smoking status in Vanderbilt medical records.[50] Height and weight were extracted from patient records within 2 months before and 1 month after cancer diagnosis date and used to estimate BMI. Although structured fields of height and weight exist in the EHRs, these data were often missing (42% individuals were missing height information and 36% were missing weight). To reduce the number of missing values for height and weight, we developed a regular expression-based program to extract height and weight information from clinical notes. Our manual review of 100 random NLP-extracted weights and heights revealed the PPV was 100%. Use of the NLP method reduced the percentage of missing data for height and weight to 33% and 16%, respectively. In addition, Deyo adaptation of the Charlson comorbidity index was calculated using ICD-9 codes.[51] Since cancer mortality is the primary response and subcancer type was either adjusted in the regression model or was the group in the subgroup analysis, ICD-9 codes of cancer diagnoses (140–239) were excluded. All non-cancer ICD-9 codes before or within 30 days after the date of cancer diagnosis were used for the Charlson comorbidity index calculation.

To verify the accuracy of our automated data extraction algorithms for drug exposure, a stratified random sample was selected from each exposure group (N=50 for each group) and two thoracic oncology nurses independently reviewed the medical charts to determine drug exposure. Discrepancies between the two nurse reviewers were reconciled by a third physician reviewer (JCD), thus forming a 'gold-standard' to compare with the automated algorithms. Metformin, other DM2 drug, and insulin groups achieved PPVs of 0.98, 0.95, and 0.92, respectively.

The same study design was applied to the Mayo Clinic EHR. The tumor registry at Mayo Clinic is also linked to the EHR to identify cancer patients and obtain tumor-specific information. The same algorithm was used to identify patients with and without diabetes in the Mayo EHR. The MedEx tool was used to process Mayo clinical data to identify different DM2 drug exposure groups. A locally developed program,[52] similar to the Vanderbilt algorithm, was used to identify the smoking status of patients in this study.

Statistical analysis

Characteristics of the study population were summarized using median, IQR, and percent. Kaplan–Meier plots were used to visualize the unadjusted cancer survival probabilities of the four exposure groups. To formally assess the influence of metformin on cancer mortality, we used stratified Cox regression models, stratifying on tumor stage, and adjusting for age, sex, race, BMI, tobacco use, insulin, cancer type, and non-cancer Charlson index. Similar stratified Cox regression models were created to evaluate the effect of metformin on cancer survival in the patient population with breast, colorectal, lung, or prostate cancer, although tumor stage 0 and 1 were combined for lung and prostate cancers due to the limited sample sizes in these

two stages. In all the aforementioned models, age, BMI, and Charlson index were modeled as restricted cubic spline functions with four knots. The covariate sex was removed from the analytical model when breast cancer and prostate cancer were being examined since these models were restricted to females and males, respectively. For the overall and individual cancer survival analysis, multiple imputation with 20 imputations was implemented for missing BMI measurements following the guidance described by White *et al*.[53] Two-sided p values less than 0.05 were considered statistically significant. All analyses were conducted using R 2.13.1 with the survival, Hmisc, and rms packages (http://www.r-project.org).

RESULTS

We identified 42 165 individuals with an incident cancer diagnosis (excluding skin cancer and CHF/CKD, and age ≥18 years old) between January 1, 1995 and December 31, 2010 (figure 1) at Vanderbilt. Among these cancer patients, 28 917 did not have diabetes, and 3498 had DM2 matching one of the three target medication exposure groups. Of these, 63% used metformin, 26% used other oral DM2 medications, and 11% were on insulin monotherapy. Vanderbilt cancer patients had a median age of 59 years, approximately half (57%) were male, and 93% were white (table 1). Median BMI was 27 kg/m^2 and 53% of cancer patients were ever smokers. DM2 patients had a median hemoglobin A1c of 7.6%. At Mayo Clinic, we identified 79 258 patients in four exposure groups (figure 2 and table 2). Figure 3 presents the Kaplan–Meier survival curves and associated 95% CIs for the four exposure groups at Vanderbilt University and Mayo Clinic. Cancer patients on metformin drugs had significantly improved 5-year survival compared to patients on other oral hypoglycemic agents ($p<0.001$), insulin only ($p<0.001$), or without diabetes ($p<0.001$). Adjusting for age, sex, race, BMI, tobacco use, insulin use, cancer type, and Charlson index, metformin significantly reduced overall mortality compared to diabetic patients on other oral hypoglycemic (HR 0.78; 95% CI 0.69 to 0.88) and diabetic patients on insulin only (HR 0.61; 95% CI 0.50 to 0.73). Reduced mortality was also observed for metformin compared to cancer patients without diabetes (HR 0.77; 95% CI 0.71 to 0.85) (figure 4). We replicated our findings for overall mortality after a cancer diagnosis in the Mayo Clinic cohort with HRs and 95% CIs as follows: HR 0.70 (95% CI 0.63 to 0.77), HR 0.65 (95% CI 0.58 to 0.73), and HR 0.59 (95% CI 0.54 to 0.65) (figure 4), comparing the metformin group versus other drugs, insulin only, and non-diabetic groups, respectively.

The impact of metformin on mortality varied by cancer type and also by exposure group (figure 4). In the Vanderbilt cohort, reduced mortality with metformin use was observed across all four of the most frequent cancers, specifically breast, colorectal, lung, and prostate. Among diabetic patients with breast cancer, the greatest benefit was observed with metformin use compared to use of other diabetes drugs or insulin only. A reduced but not statistically significant HR was observed when metformin diabetic patients with breast cancer were compared to non-diabetic breast cancer patients. For colorectal cancer, metformin was beneficial compared to patients with diabetes on other drugs and cancer patients without diabetes. Lung cancer and prostate cancer mortality was not significantly improved with metformin, although the HRs did show an overall trend toward reduced mortality. Associations showed a similar protective benefit of metformin in the Mayo cancer population and most associations by cancer type (breast, colorectal, lung, and prostate) were statistically significant, likely due to larger sample sizes in the Mayo cohort. Metformin was also associated with improved survival in other cancers types in at least one EHR, including bone marrow, gynecologic, genitourinary, and gastrointestinal (see online supplementary appendix figure 1). We present the adjusted overall cancer survival curves for each tumor stage in figure 5. Metformin reduced mortality irrespective of tumor stage. Predicted survival curves for colorectal, lung, breast, and prostate cancer show patients on metformin had improved survival for each specific cancer (figure 6).

Table 1 Descriptive characteristics of the Vanderbilt cohort (all cancers, 1995–2010)

	N	DM2 Metformin N=2218	DM2 Other drugs N=903	DM2 Insulin N=377	Non-diabetic patients N=28 917	Combined N=32 415
Age, years	32 415	54, 62, 69*	56, 64, 70	48, 55, 65	48, 58, 67	49, 59, 67
Male	32 413	58%	61%	54%	57%	57%
White	29 371	88%	90%	86%	93%	93%
Body mass index (kg/m^2)	27 352	27, 31, 36	26, 31, 35	25, 30, 35	23, 27, 31	24, 27, 32
Hemoglobin A1c	1 069	7.1, 7.6, 8.5	7.1, 7.6, 8.4	7.1, 7.7, 8.6	NA	7.1, 7.6, 8.5
Tobacco use (ever/never)	22 885	58%	60%	61%	53%	53%
Insulin use	32 415	27%	27%	100%	0%	4%
Tumor type	32 415					
Colorectal		8%	7%	3%	6%	6%
Breast		9%	4%	3%	10%	9%
Lung		7%	8%	5%	8%	8%
Prostate		14%	9%	2%	18%	18%
Other		63%	71%	86%	58%	59%
Tumor stage	27 017					
0		5%	4%	2%	6%	6%
1		28%	25%	22%	26%	26%
2 or 3		46%	44%	32%	47%	47%
4		21%	27%	43%	21%	22%

*Lower, median, and upper quartile for continuous variables.
N is the number of non-missing values.

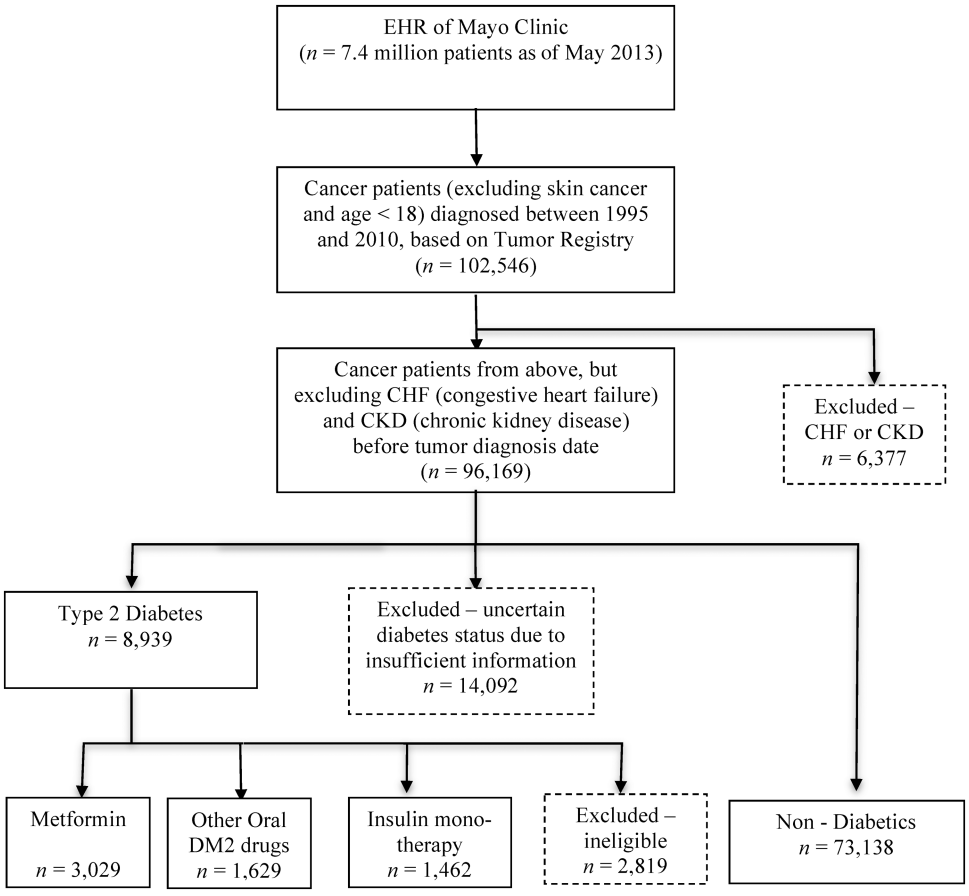

Figure 2 The study design and data extraction workflow for patients in the Mayo Clinic electronic health record (EHR) system from January 1995 to December 2010.

DISCUSSION

Using two independent study populations, we validated the recently reported drug repurposing association of metformin with cancer survival. Our data demonstrate that metformin improves overall cancer survival compared to other hypoglycemic therapies in patients with DM2 and compared to patients

Table 2 Descriptive characteristics of the Mayo Clinic cohort (all cancers, 1995–2010)

	N	DM2 Metformin N=3029	DM2 Other drugs N=1629	DM2 Insulin N=1462	Non-diabetic patients N=73 138	Combined N=79 258
Age, years	79 258	58, 65, 72*	62, 69, 75	57, 65, 72	53, 62, 71	54, 62, 71
Male	79 258	60%	68%	61%	57%	58%
White	70 411	99%	98%	99%	99%	99%
Body mass index (kg/m^2)	57 513	28, 32, 36	27, 30, 34	26, 29, 33	24, 27, 30	24, 27, 30
Tobacco use (ever/never)	67 680	52%	50%	46%	37%	38%
Insulin use	79 258	45%	36%	100%	0%	5%
Tumor type	79 258					
Colorectal		7%	7%	7%	6%	6%
Breast		12%	7%	7%	11%	11%
Lung		7%	11%	6%	10%	10%
Prostate		19%	17%	7%	22%	21%
Other		55%	59%	74%	50%	47%
Tumor stage	73 224					
0		7%	5%	3%	5%	5%
1		30%	26%	31%	26%	27%
2 or 3		46%	48%	42%	49%	48%
4		17%	21%	24%	20%	20%

*Lower, median, and upper quartile for continuous variables.
N is the number of non-missing values.

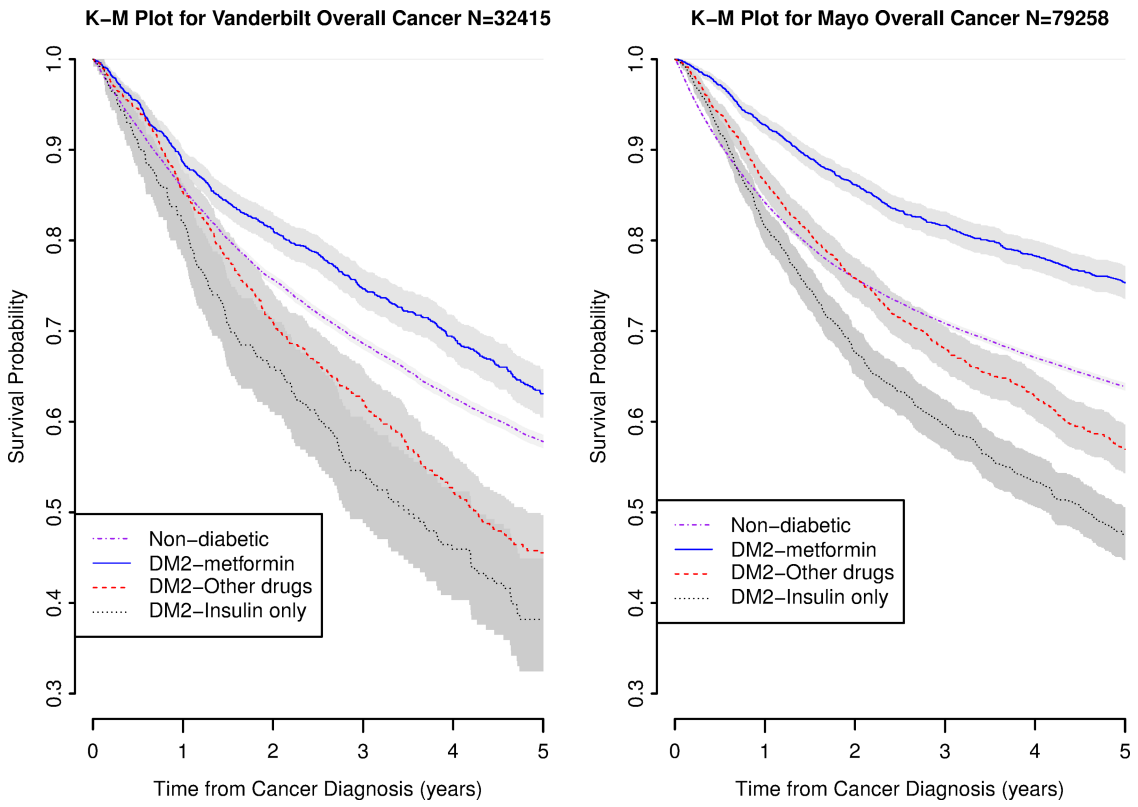

Figure 3 Kaplan–Meier (K–M) plot of overall cancer survival for the Vanderbilt and Mayo Clinic cohorts. DM2, type 2 diabetes mellitus.

without diabetes. These findings included a total of 111 673 patients and demonstrated a metformin survival benefit for individuals with breast and colorectal cancer in both the Vanderbilt and Mayo cohorts. Evidence for lung and prostate cancer showed a reduced mortality in both the Vanderbilt and Mayo populations, which was statistically significant only in the Mayo

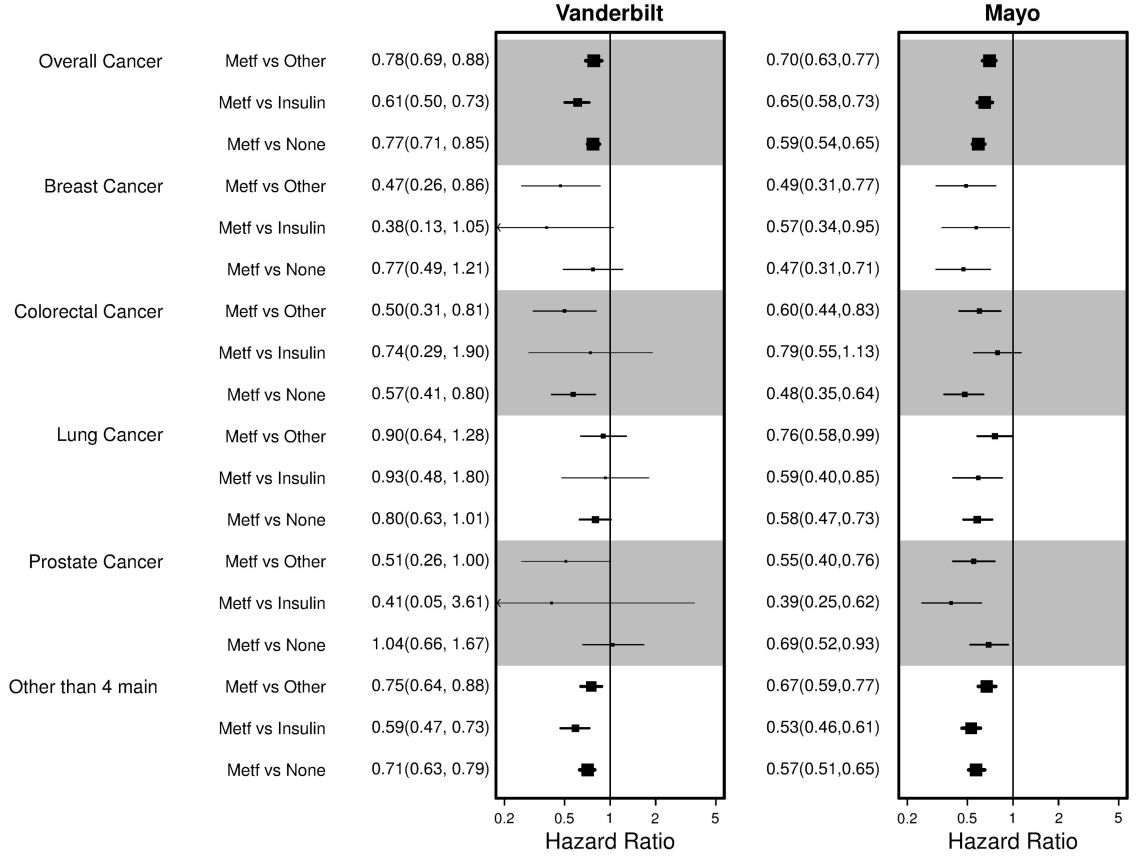

Figure 4 Adjusted HRs by cancer type for the Vanderbilt and Mayo cohorts. Other, DM2 cancer patients on other drugs; Insulin, DM2 cancer patients on insulin only; Metf, DM2 cancer patients on metformin; None, cancer patients without DM2.

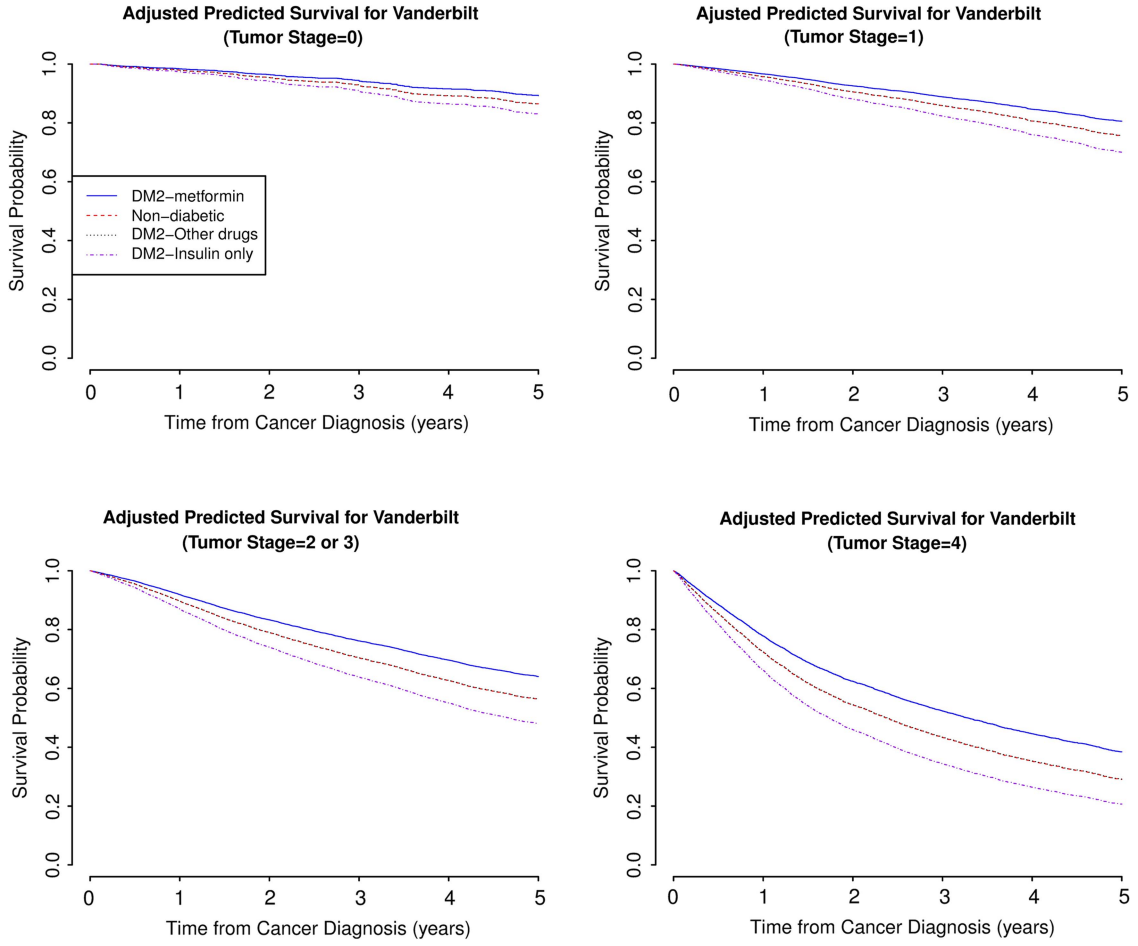

Figure 5 Adjusted Cox proportional hazards model stratified by tumor stage for the Vanderbilt cohort. All models are based on cancer survival in a smoking white male, age 58 years, body mass index 27 kg/m^2, with a cancer other than the four most common tumor types, and not using insulin. DM2, type 2 diabetes mellitus.

cohort, likely due to its larger sample size. Mortality improvements were also seen for a number of other cancers and for all cancer stages. Thus, our data support a broad role for metformin in many cancer types and, potentially, for patients with and without diabetes.

We leveraged study site-maintained tumor registries combined with advanced informatics techniques examining the full text of the EHR. Prior studies have shown such methods lead to more accurate results than use of administrative data alone, as has been used in previous studies.[44 46] These informatics methods, applied at both study sites, interrogated patient records to provide information on detailed medication exposures and important cancer risk factors such as smoking histories and BMI —detail not commonly afforded in retrospective claims data. With the future ubiquity of available EHR data, such data mining may provide an important tool for drug repurposing, pharmacovigilance, and comparative effectiveness research.

Our findings add to a growing body of knowledge supporting a role for metformin in reducing cancer mortality.[31 32 54] A strength of this study is that the same study population was used to evaluate multiple cancers. Metformin was also statistically associated with improved survival for less common cancers (see online supplementary appendix figure 1), suggesting future studies with greater statistical power should evaluate these less frequently observed cancers. Moreover, most prior epidemiologic studies have used DM2 registries[31] or patient surveys[38] to assess the association between metformin and cancer risk and survival. We were able to utilize two densely populated EHR-based cohorts in the USA with longitudinal follow-up and linkage with tumor registries. We were also able to incorporate smoking status into our analyses, an important consideration for many cancers but not assessed in some other retrospective studies.[38 39] Using NLP for data extraction is an efficient design for hospital-based epidemiologic studies, significantly reducing the time and cost to conduct and replicate the study since no follow-up of participants is needed. In addition, our study was replicated in another independent large EHR (Mayo Clinic), demonstrating the generalizability of both our findings and the informatics tools used in this study.

The mechanism by which metformin improves cancer survival either directly (insulin-independent) or indirectly (insulin-dependent) remains unknown[37 40 55 56] but may be related to mTOR inhibition.[57 58] The broad-based effect on multiple cancers seen in this study suggests a generalized anticancer effect. Future studies are needed to unravel the exact mechanism by which metformin acts and whether metformin should be targeted to particular patients. Currently, large efforts are underway to link EHRs across institutions and to standardize the definition of phenotypes for large-scale clinical and genomics studies of disease and treatment.[59–61] Informatics approaches, such as NLP technologies that are able to extract standardized clinical information from unstructured clinical text, offer an

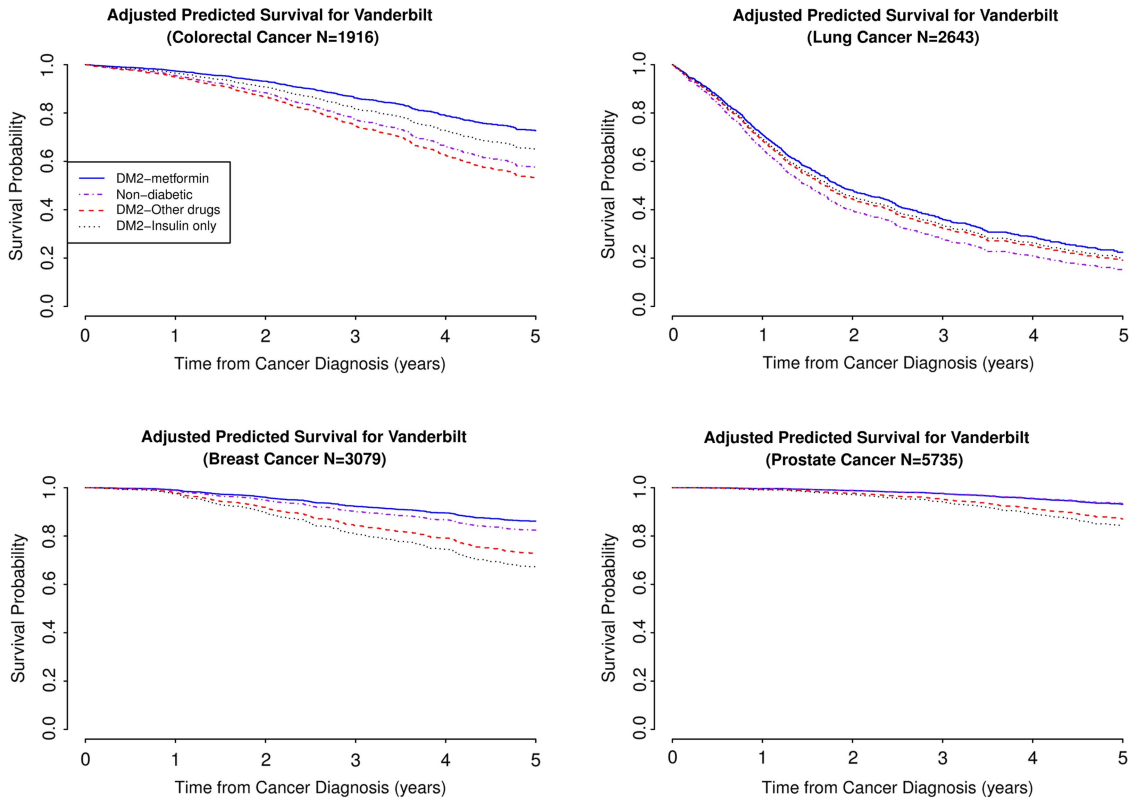

Figure 6 Adjusted Cox proportional hazards model stratified by tumor type for the Vanderbilt cohort. All models are based on cancer survival in a white smoker, age 58 years, body mass index 27 kg/m^2, and not using insulin. DM2, type 2 diabetes mellitus.

approach to automate the data extraction process from EHRs.[62] Successful progress has been made in applying informatics approaches to clinical and translational research, ranging from identifying patient safety occurrences[29] and biosurveillance[28] to facilitating genomics research such as genetic epidemiology and pharmacogenomic studies.[63 64] In this study, we further demonstrated the value of NLP tools and electronic phenotyping algorithms in epidemiologic studies based on large-scale observational clinical practice data. To improve the efficiency of EHR-based epidemiologic research, more informatics tools to record and/or accurately extract broad types of epidemiologic information such as environmental variables (eg, exercise, diet, and other lifestyle data) are highly desirable.

Limitations caution interpretation of our findings. Our medication exposures were derived from EHRs instead of pharmacy fill records. However, we have previously shown that these methods have both high sensitivity and high PPV and the ability to replicate known pharmacogenetic signals that require accurate knowledge of the timelines of medications exposures.[65 66] Moreover, in comparison to claims data, they are not subject to biases from low-cost generic prescriptions, for which insurance claims are often not filed.[67] The potential imperfect sensitivity of our algorithms leads to an inability to classify every patient as either diabetic or not, or to fully determine their medication exposures, primarily due to lack of data captured in the EHR. For example, we cannot exclude remote exposures to particular antidiabetic medications occurring prior to cancer diagnosis (eg, a patient with diabetes may have been treatment with metformin prior to the cancer diagnosis at an outside hospital). However, these exposures should have limited effect on cancer prognosis. Our study may be subject to immortal time bias due to misclassification of exposure time, since we are unable to discern whether erroneous exposure time was assigned between cohort entry and mention of medication in the clinical record.[68 69] Excluding CHF and CKD patients could be a potential limitation of this study as well, as some physicians use metformin for these patients despite FDA warnings in these populations. We were also unable to stratify by histologic subtype within each cancer type due to small sample sizes within each cancer. We did not adjust for chemotherapy treatment regimens due to the lack of treatment information beyond first-line therapy in the tumor registry. This is a common limitation of epidemiologic studies using tumor registry or SEER data for cancer treatment information. However, it is likely that diabetic patients using metformin receive the same cancer treatment as those not using metformin, thus biasing our results towards the null. There is no published evidence, to our knowledge, of disparities in the treatment of diabetic patients with cancer, although the dosages of steroid pre-medications are often reduced in an effort to reduce incident hyperglycemia. Future classes of antineoplastics, for example, phosphoinositide 3-kinase inhibitors, may be specifically contraindicated for diabetic patients, but these medications are not yet approved for general clinical use. Diabetic patients with cancer may have greater co-morbidities than non-diabetic patients with cancer and we would expect diabetic patients to have worse survival after a cancer diagnosis than non-diabetic patients.[70] However, we found in most comparisons for the four major cancers that diabetic patients on metformin had a better survival compared to non-diabetic patients, although non-diabetic patients had a better survival than diabetic patients using other drugs or insulin only. This observation is consistent with that from a recent study conducted in the UK.[32] One possible interpretation for this finding is that metformin use significantly improved survival among

diabetic patients despite higher prevalence of co-morbidities. Thus, it is possible that metformin use may be able to improve survival among non-diabetic cancer patients. Further studies are needed to address this important issue.

We successfully detected the signal of metformin improving cancer survival using EHR data and informatics approaches. However, conducting large-scale drug repurposing studies using EHRs remains challenging. One of the problems is related to sample size. We had enough power in this study because both DM2 and cancer are high prevalence diseases and metformin is a first-line therapy for DM2. But the lack of power would be an issue for low prevalence drugs and indications. In our stratified analysis for individual cancers (see online supplementary appendix figure 1), we noticed larger CIs for less frequent cancers such as thyroid, most likely due to small sample size. This problem may be ameliorated by combining EHRs and/or complementing EHR data with data provided by drug manufacturers, drug monitoring agencies (eg, the FDA), and other ancillary data sources.

CONCLUSION

In this study, we have demonstrated that large EHRs are valuable sources for drug repurposing studies. Our findings validate the beneficial effects of metformin for cancer survival. Ongoing and future clinical trials of metformin for specific subtypes of cancer may lead to new opportunities for chemotherapy. This study serves as a model for using EHRs and informatics approaches to robustly and inexpensively validate drugs for repurposing.

Author affiliations
[1]The University of Texas School of Biomedical Informatics at Houston, Houston, Texas, USA
[2]Department of Thoracic Surgery, Vanderbilt University School of Medicine, Nashville, Tennessee, USA
[3]Division of Epidemiology, Vanderbilt University School of Medicine, Nashville, Tennessee, USA
[4]Department of Biostatistics, Vanderbilt University School of Medicine, Nashville, Tennessee, USA
[5]Department of Biomedical Informatics, Vanderbilt University School of Medicine, Nashville, Tennessee, USA
[6]Division of Biomedical Statistics and Informatics, Mayo Clinic, Rochester, Minnesota, USA
[7]Department of Medicine, Vanderbilt University School of Medicine, Nashville, Tennessee, USA
[8]Department of Biomedical Informatics, Columbia University, New York, New York, USA
[9]Department of Pharmacology, Vanderbilt University School of Medicine, Nashville, Tennessee, USA

Acknowledgements We thank Rhonda Pinkerman and Breanne Osborne for their careful chart review.

Contributors HX, MCA, and JCD designed and performed the research, guided the data analysis, and wrote the manuscript. QC and XH analyzed the data and wrote the manuscript. AS, MJ, XR, JSJ, and YL developed informatics tools, and extracted and processed data. HL, NBP, QD, ML, JW, CF, and DMR provided domain expertise, designed research, and wrote the manuscript. All authors read, revised, and approved the manuscript.

Funding This study was supported in part by National Cancer Institute grant R01CA141307, K07CA172294, a Vanderbilt Lung SPORE (P50 CA090949) Career Development Award, R21HL097334, and CPRIT (Cancer Prevention and Research Institute of Texas) R1307. Vanderbilt University Medical Center's Synthetic Derivative is supported by institutional funding and by Vanderbilt CTSA grant 1UL1RR024975 from NCRR/NIH. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests None.

Ethics approval Vanderbilt and Mayo Clinic IRBs approved this study.

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES

1 Ashburn TT, Thor KB. Drug repositioning: identifying and developing new uses for existing drugs. *Nat Rev Drug Discov* 2004;3:673–83.
2 Reichert JM. Trends in development and approval times for new therapeutics in the United States. *Nat Rev Drug Discov* 2003;2:695–702.
3 Gilbert J, Henske P, Singh A. Rebuilding big pharma's business model. *In Vivo* 2003;21:73–80.
4 Tobinick EL. The value of drug repositioning in the current pharmaceutical market. *Drug News Perspect* 2009;22:119–25.
5 Sanseau P, Agarwal P, Barnes MR, *et al*. Use of genome-wide association studies for drug repositioning. *Nat Biotechnol* 2012;30:317–20.
6 O'Connor KA, Roth BL. Finding new tricks for old drugs: an efficient route for public-sector drug discovery. *Nat Rev Drug Discov* 2005;4:1005–14.
7 Harrison C. Signatures for drug repositioning. *Nat Rev Genet* 2011;12:668.
8 Collins FS. Reengineering translational science: the time is right. *Sci Transl Med* 2011;3:90cm17.
9 Weir SJ, DeGennaro LJ, Austin CP. Repurposing approved and abandoned drugs for the treatment and prevention of cancer through public-private partnership. *Cancer Res* 2012;72:1055–8.
10 Hurle MR, Yang L, Xie Q, *et al*. Computational drug repositioning: from data to therapeutics. *Clin Pharmacol Ther* 2013;93:335–41.
11 Swamidass SJ, Lu Z, Agarwal P, *et al*. Computational approaches to drug repurposing and pharmacology - session introduction. *Pac Symp Biocomput* 2014;19:110–3.
12 Huang R, Southall N, Wang Y, *et al*. The NCGC pharmaceutical collection: a comprehensive resource of clinically approved drugs enabling repurposing and chemical genomics. *Sci Transl Med* 2011;3:80ps16.
13 Austin CP, Brady LS, Insel TR, *et al*. NIH Molecular Libraries Initiative. *Science* 2004;306:1138–9.
14 Ma DL, Chan DS, Leung CH. Drug repositioning by structure-based virtual screening. *Chem Soc Rev* 2013;42:2130–41.
15 Campillos M, Kuhn M, Gavin AC, *et al*. Drug target identification using side-effect similarity. *Science* 2008;321:263–6.
16 Lounkine E, Keiser MJ, Whitebread S, *et al*. Large-scale prediction and testing of drug activity on side-effect targets. *Nature* 2012;486:361–7.
17 Wang ZY, Zhang HY. Rational drug repositioning by medical genetics. *Nat Biotechnol* 2013;31:1080–2.
18 Sirota M, Dudley JT, Kim J, *et al*. Discovery and preclinical validation of drug indications using compendia of public gene expression data. *Sci Transl Med* 2011;3:96ra77.
19 Andronis C, Sharma A, Virvilis V, *et al*. Literature mining, ontologies and information visualization for drug repurposing. *Brief Bioinform* 2011;12:357–68.
20 Shea S, Hripcsak G. Accelerating the use of electronic health records in physician practices. *N Engl J Med* 2010;362:192–5.
21 Jha AK, DesRoches CM, Campbell EG, *et al*. Use of electronic health records in US hospitals. *N Engl J Med* 2009;360:1628–38.
22 Strom BL. *Pharmacoepidemiology*. 4th edn. Chichester; Hoboken, NJ: J. Wiley, 2005.
23 Haerian K, Varn D, Vaidya S, *et al*. Detection of pharmacovigilance-related adverse events using electronic health records and automated methods. *Clin Pharmacol Ther* 2012;92:228–34.
24 LePendu P, Iyer SV, Bauer-Mehren A, *et al*. Pharmacovigilance using clinical notes. *Clin Pharmacol Ther* 2013;93:547–55.
25 Schildcrout JS, Denny JC, Bowton E, *et al*. Optimizing drug outcomes through pharmacogenetics: a case for preemptive genotyping. *Clin Pharmacol Ther* 2012;92:235–42.
26 Jiang M, Chen Y, Liu M, *et al*. A study of machine-learning-based approaches to extract clinical entities and their assertions from discharge summaries. *J Am Med Inform Assoc* 2011;18:601–6.
27 Xu H, Fu Z, Shah A, *et al*. Extracting and integrating data from entire electronic health records for detecting colorectal cancer cases. *AMIA Annu Symp Proc* 2011;2011:1564–72.
28 Elkin PL, Froehling DA, Wahner-Roedler DL, *et al*. Comparison of natural language processing biosurveillance methods for identifying influenza from encounter notes. *Ann Intern Med* 2012;156(1 Pt 1):11–8.
29 Murff HJ, FitzHenry F, Matheny ME, *et al*. Automated identification of postoperative complications within an electronic medical record using natural language processing. *JAMA* 2011;306:848–55.
30 Tatonetti NP, Patrick PY, Daneshjou R, *et al*. Data-driven prediction of drug effects and interactions. *Sci Transl Med* 2012;4:125ra131.

31 Landman GW, Kleefstra N, van Hateren KJ, *et al*. Metformin associated with lower cancer mortality in type 2 diabetes: ZODIAC-16. *Diabetes Care* 2010;33:322–6.
32 Currie CJ, Poole CD, Jenkins-Jones S, *et al*. Mortality after incident cancer in people with and without type 2 diabetes: impact of metformin on survival. *Diabetes Care* 2012;35:299–304.
33 Evans JM, Donnelly LA, Emslie-Smith AM, *et al*. Metformin and reduced risk of cancer in diabetic patients. *BMJ* 2005;330:1304–5.
34 Libby G, Donnelly LA, Donnan PT, *et al*. New users of metformin are at low risk of incident cancer: a cohort study among people with type 2 diabetes. *Diabetes Care* 2009;32:1620–5.
35 Currie CJ, Poole CD, Gale EA. The influence of glucose-lowering therapies on cancer risk in type 2 diabetes. *Diabetologia* 2009;52:1766–77.
36 Ruiter R, Visser LE, van Herk-Sukel MP, *et al*. Lower risk of cancer in patients on metformin in comparison with those on sulfonylurea derivatives: results from a large population-based follow-up study. *Diabetes Care* 2012;35:119–24.
37 Sahra IB, Le Marchand-Brustel Y, Tanti JF, *et al*. Metformin in cancer therapy: a new perspective for an old antidiabetic drug? *Mol Cancer Ther* 2010;9:1092–9.
38 Sadeghi N, Abbruzzese JL, Yeung SC, *et al*. Metformin use is associated with better survival of diabetic patients with pancreatic cancer. *Clin Cancer Res* 2012;18:2905–12.
39 Garrett CR, Hassabo HM, Bhadkamkar NA, *et al*. Survival advantage observed with the use of metformin in patients with type II diabetes and colorectal cancer. *Br J Cancer* 2012;106:1374–8.
40 Jalving M, Gietema JA, Lefrandt JD, *et al*. Metformin: taking away the candy for cancer? *Eur J Cancer* 2010;46:2369–80.
41 Gallagher EJ, LeRoith D. Diabetes, cancer, and metformin: connections of metabolism and cell proliferation. *Ann N Y Acad Sci* 2011;1243:54–68.
42 Roden DM, Pulley JM, Basford MA, *et al*. Development of a large-scale de-identified DNA biobank to enable personalized medicine. *Clin Pharmacol Ther* 2008;84:362–9.
43 NAACCR. Death Clearance Manual. 2009. http://www.naaccr.org/LinkClick.aspx?fileticket=RD1FxWlmC24%3D&tabid=130&mid=470 (accessed 5 May 2013).
44 Denny JC. Chapter 13: Mining electronic health records in the genomics era. *PLoS Comput Biol* 2012;8:e1002823.
45 Kho AN, Hayes MG, Rasmussen-Torvik L, *et al*. Use of diverse electronic medical record systems to identify genetic risk for type 2 diabetes within a genome-wide association study. *J Am Med Inform Assoc* 2012;19:212–8.
46 Ritchie MD, Denny JC, Crawford DC, *et al*. Robust replication of genotype-phenotype associations across multiple diseases in an electronic medical record. *Am J Hum Genet* 2010;86:560–72.
47 Gottesman O, Kuivaniemi H, Tromp G, *et al*. The Electronic Medical Records and Genomics (eMERGE) Network: past, present, and future. *Genet Med* 2013;15:761–71.
48 Xu H, Stenner SP, Doan S, *et al*. MedEx: a medication information extraction system for clinical narratives. *J Am Med Inform Assoc* 2010;17:19–24.
49 Doan S, Bastarache L, Klimkowski S, *et al*. Integrating existing natural language processing tools for medication extraction from discharge summaries. *J Am Med Inform Assoc* 2010;17:528–31.
50 Liu M, Shah A, Jiang M, *et al*. A study of transportability of an existing smoking status detection module across institutions. *AMIA Annu Symp Proc* 2012;2012:577–86.
51 Deyo RA, Cherkin DC, Ciol MA. Adapting a clinical comorbidity index for use with ICD-9-CM administrative databases. *J Clin Epidemiol* 1992;45:613–9.
52 Sohn S, Savova GK. Mayo Clinic smoking status classification system: extensions and improvements. *AMIA Annu Symp Proc* 2009;2009:619–23.
53 White IR, Royston P, Wood AM. Multiple imputation using chained equations: issues and guidance for practice. *Stat Med* 2011;30:377–99.
54 Lega IC, Austin PC, Gruneir A, *et al*. Association between metformin therapy and mortality after breast cancer: a population-based study. *Diabetes Care* 2013;36:3018–26.
55 Dowling RJ, Niraula S, Stambolic V, *et al*. Metformin in cancer: translational challenges. *J Mol Endocrinol* 2012;48:R31–43.
56 Giovannucci E, Harlan DM, Archer MC, *et al*. Diabetes and cancer: a consensus report. *CA Cancer J Clin* 2010;60:207–21.
57 Lamming DW, Ye L, Sabatini DM, *et al*. Rapalogs and mTOR inhibitors as anti-aging therapeutics. *J Clin Invest* 2013;123:980–9.
58 Sinnett-Smith J, Kisfalvi K, Kui R, *et al*. Metformin inhibition of mTORC1 activation, DNA synthesis and proliferation in pancreatic cancer cells: dependence on glucose concentration and role of AMPK. *Biochem Biophys Res Commun* 2013;430:352–7.
59 McCarty CA, Wilke RA. Biobanking and pharmacogenomics. *Pharmacogenomics* 2010;11:637–41.
60 Pace WD, Cifuentes M, Valuck RJ, *et al*. An electronic practice-based network for observational comparative effectiveness research. *Ann Intern Med* 2009;151:338.
61 Kho AN, Pacheco JA, Peissig PL, *et al*. Electronic medical records for genetic research: results of the eMERGE consortium. *Sci Transl Med* 2011;3:79re71.
62 Meystre SM, Savova GK, Kipper-Schuler KC, *et al*. Extracting information from textual documents in the electronic health record: a review of recent research. *Yearb Med Inform* 2008:128–44.
63 Denny JC, Ritchie MD, Basford MA, *et al*. PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene–disease associations. *Bioinformatics* 2010;26:1205.
64 Xu H, Jiang M, Oetjens M, *et al*. Facilitating pharmacogenetic studies using electronic health records and natural-language processing: a case study of warfarin. *J Am Med Inform Assoc* 2011;18:387–91.
65 Delaney JT, Ramirez AH, Bowton E, *et al*. Predicting clopidogrel response using DNA samples linked to an electronic health record. *Clin Pharmacol Ther* 2012;91:257–63.
66 Ramirez AH, Shi Y, Schildcrout JS, *et al*. Predicting warfarin dosage in European-Americans and African-Americans using DNA samples linked to an electronic health record. *Pharmacogenomics* 2012;13:407–18.
67 Choudhry NK, Shrank WH. Four-dollar generics--increased accessibility, impaired quality assurance. *N Engl J Med* 2010;363:1885–7.
68 Levesque LE, Hanley JA, Kezouh A, *et al*. Problem of immortal time bias in cohort studies: example using statins for preventing progression of diabetes. *BMJ* 2010;340:b5087.
69 Suissa S, Azoulay L. Metformin and the risk of cancer: time-related biases in observational studies. *Diabetes Care* 2012;35:2665–73.
70 Seshasai SR, Kaptoge S, Thompson A, *et al*. Diabetes mellitus, fasting glucose, and risk of cause-specific death. *N Engl J Med* 2011;364:829–41.