Journal of Biomedical Informatics 43 (2010) 747–751

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

The utility of general purpose versus specialty clinical databases for research: Warfarin dose estimation from extracted clinical variables

Hersh Sagreiya, Russ B. Altman *

Stanford University, Stanford, CA, USA

ARTICLE INFO

Article history:
Received 11 November 2009
Available online 2 April 2010

Keywords:
Clinical
Translational
Database
Warehouse
Research
Quality
Warfarin
Dosing
STRIDE
CoagClinic

ABSTRACT

There is debate about the utility of clinical data warehouses for research. Using a clinical warfarin dosing algorithm derived from research-quality data, we evaluated the data quality of both a general-purpose database and a coagulation-specific database. We evaluated the functional utility of these repositories by using data extracted from them to predict warfarin dose. We reasoned that high-quality clinical data would predict doses nearly as accurately as research data, while poor-quality clinical data would predict doses less accurately. We evaluated the Mean Absolute Error (MAE) in predicted weekly dose as a metric of data quality. The MAE was comparable between the clinical gold standard (10.1 mg/wk) and the specialty database (10.4 mg/wk), but the MAE for the clinical warehouse was 40% greater (14.1 mg/wk). Our results indicate that the research utility of clinical data collected in focused clinical settings is greater than that of data collected during general-purpose clinical care.

1. Introduction

The value of clinical data and clinical data warehouses for research is controversial. On one hand, they integrate information from multiple sources, such as patient charts, radiology reports, and laboratory results, and offer a centralized resource for accessing multiple data sources. Compared to most large clinical studies, querying these databases does not require a substantial commitment of time or resources, making them a potentially useful tool for hypothesis testing. Simple queries can identify thousands of records, granting these studies significant statistical power. Nevertheless, the quality of this data is subject to the realities and variations in clinical practice, and is not equivalent to the data in a research clinical trial. These issues have become particularly important in the setting of many population-based efforts in genomic discovery [1,2]. In these efforts, large clinical data repositories are used to search for potential associations between clinical phenotypes and genetic markers [3]. They rest on the assumption that the quality of the clinical phenotypes derived from clinical databases will be sufficiently high to provide detectable signals. Thus, the purpose of this study was to test the accuracy of clinical data collected and stored in an enterprise-wide data warehouse with clinical data collected and stored in a specialty clinic focusing on the variables of interest. In particular, we used the performance of a clinical algorithm for estimating the dose of warfarin as a functional measure of data quality. We asked whether data extracted from the general clinical database or the specialty clinic database could be used to estimate warfarin doses with accuracies comparable to that observed in research-grade data.

Warfarin is an anticoagulant taken by 30 million Americans, with 2 million new starts per year [4]. The therapeutic range of the drug is narrow, as a high dose can lead to hemorrhage, while a low dose fails to protect against thromboembolism. As a result, it is one of the top three drugs leading to emergency department visits by the elderly, accounting for 17.3% of such visits [5]. Warfarin is dosed by trial-and-error. Patients are initially given a fixed 5 mg starting dose. On each subsequent visit, the International Normalized Ratio (INR) is used to measure clotting time, and warfarin dose is gradually adjusted using this test until a stable dose is achieved. This usually occurs when the INR value is between 2.0 and 3.0. Clinical factors, such as height, weight, age, and race, affect the final therapeutic dose. The International Warfarin Pharmacogenetics Consortium (IWPC) has developed two algorithms that predict warfarin dose, one that uses both genetic and clinical factors and another that uses only clinical factors [6]. We focus on the algorithm using clinical factors in this work because genetic data is not yet routinely stored in clinical databases. The key clinical factors are age, weight, ethnic background, use of amiodarone, and use of other drugs known to induce the metabolizing enzyme CYP2C9. Using these variables, the IWPC

* Corresponding author. Address: Department of Bioengineering, Stanford University School of Medicine, 318 Campus Drive, Clark Center S170, MC: 5444, Stanford, CA 94305-5444, USA. Fax: +1 650 723 8544.
E-mail address: russ.altman@stanford.edu (R.B. Altman).

1532-0464/$ - see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2010.03.014

predicted the weekly warfarin dose with an average mean absolute error of 10.1 mg/week on a diverse global population of more than 5000 patients taking warfarin.

STRIDE is a clinical data warehouse created for research purposes, but based on clinical data collected during the provision of care. The data model for STRIDE is based upon the Health Level Seven Reference Information Model, and it uses SNOMED CT and the National Library of Medicine's Unified Medical Language System, UMLS [7]. As of April 2009, STRIDE contains over 7.5 million full-text clinical documents. This database permits clinical data extraction, yielding thousands of short excerpts de-identified of personal information. STRIDE offers substantial data integration, allowing investigators to ask arbitrary questions across diseases and cohorts.

The CoagClinic database is maintained by a hematologist and pharmacists who specialize in anticoagulation [8]. Moreover, the database itself is designed with warfarin therapy in mind. This database contains the dose and INR values from every visit, ICD-9 codes for each diagnosis, target INR goal, complete pharmacy records, and the physician notes from each visit. In order to evaluate the quality of data in clinical warehouses, we queried both STRIDE and CoagClinic for all the variables used by the IWPC algorithm. We compared the accuracy of the doses predicted from STRIDE and CoagClinic with the accuracy achieved in the IWPC research cohort, which we consider our gold standard. Our goal was not to simply characterize the potential errors in each particular measurement, but instead to define a more integrative "functional" test of data quality. The clinical dosing algorithm represents such a functional test. We are thus able to compare the data quality of the research-grade clinical data with both a general-purpose database based on clinical practice (STRIDE) as well as a special-purpose database also based on clinical practice (CoagClinic). As a result, we can compare not only the functional data quality of research-grade data to clinical practice data (IWPC gold standard vs. STRIDE/CoagClinic), but also the functional data quality of general practice data to specialty practice data (STRIDE vs. CoagClinic). Our results have important implications for cohort-finding in clinical data-mining efforts.

2. Methods

2.1. Patient selection

We queried the STRIDE database and retrieved 1472 patient discharge summaries using five search terms: "height", "weight", "age", "race", and "warfarin" and/or "coumadin". The short excerpts were thoroughly examined to best ensure that the dose was therapeutic. The patient selection procedure is shown in Fig. 1. First, duplicate patient records were removed so that each patient in the analysis was unique. Patients were excluded if the text record indicated that they were recently starting or stopping warfarin treatment, if they were only going to be on warfarin for a short period of time (i.e. prior to surgery), if their warfarin dose had just changed or was unstable, if the wording of the record was ambiguous, or if no warfarin dose was available. Next, we excluded patients taking 5 mg/day, which is the standard starting dose of warfarin and hence not clearly therapeutic. We also did not include patients taking 1 mg/day, which is a standard fixed dose for indications such as PICC catheter placement. After manual examination, 357 records were selected with clearly specified warfarin doses. Of these records, 335 had all the necessary variables to predict a warfarin dose using the IWPC clinical equation (height, weight, age, and race).

For the CoagClinic database, we analyzed 104 patient records. These data are not aggregated into STRIDE and so represent a separate, non-overlapping set of data. It is possible that some patients appear in both CoagClinic and STRIDE, but our process of de-identification did not allow us to determine this. In any case, the data in STRIDE and CoagClinic was independently collected. The same variables were extracted from this clinical database.

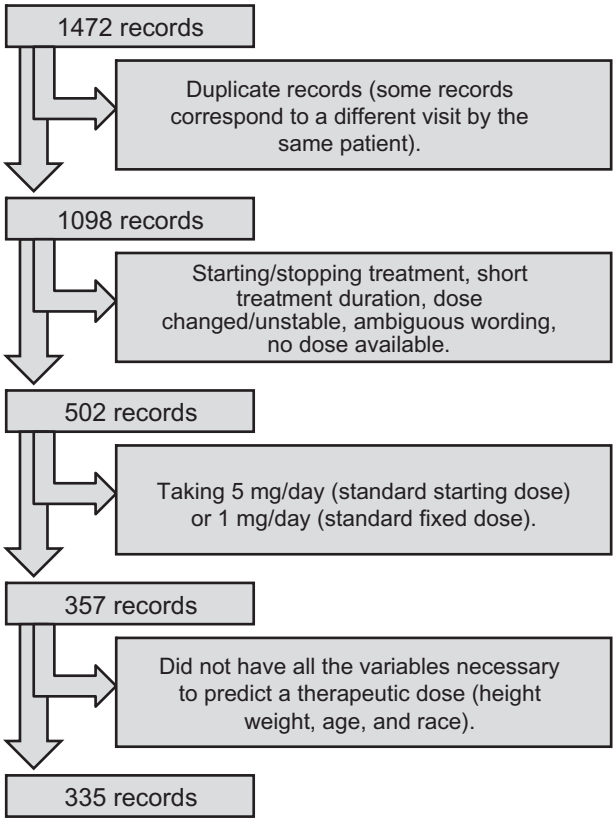


Fig. 1. Flow diagram for the patient selection procedure.

2.2. Data analysis

For each patient in the STRIDE and CoagClinic cohorts, a therapeutic warfarin dose was predicted using the IWPC clinical algorithm. Next, the mean absolute error, the average difference between the predicted dose and actual dose, was calculated for the IWPC clinical algorithm. Finally, the performance of the clinical algorithm was assessed by computing the percent of patients for whom the predicted dose was more than 1 mg/day greater than the actual dose (henceforth referred to as *high dose*), within 1 mg/day of the actual dose (*ideal dose*), and more than 1 mg/day less than the actual dose (*low dose*). The 1 mg/day cut-off was used since the IWPC had determined it to be clinically significant. In order to compare the number of patients with either a high, ideal, or low dose in the different patient populations, we used R (version 2.9.1) to perform chi square tests. The following combinations were analyzed: STRIDE vs. CoagClinic, STRIDE vs. IWPC, and CoagClinic vs. IWPC. To meet the more stringent criteria for multiple comparisons, a p-value under $0.05/3 = 0.017$ was needed.

3. Results

3.1. Results for the STRIDE database

For the STRIDE cohort, the characteristics of the 335 patients are listed in Table 1. The mean absolute error (MAE) with the IWPC algorithm was 14.1 mg/wk (Table 2). The clinical algorithm showed 29.6% of patients with an *ideal dose*, 34.3% with *high doses*, and 36.1% with *low doses* (Fig. 2).

Table 1
STRIDE, CoagClinic, and IWPC patient populations.

Variable	Measurement	STRIDE (n = 335)	CoagClinic (n = 104)	IWPC (n = 5052)
Stable dose	Mean ± SD	35.7 ± 19.5 mg/wk	35.4 ± 15.8 mg/wk	30.8 ± 16.8 mg/wk
Age	Mean ± SD	62.4 ± 13.5 years	64.0 ± 15.3 years	64.8 ± 14.5 years
Weight	Mean ± SD	82.5 ± 23.4 kg	81.0 ± 20.2 kg	78.0 ± 21.8 kg
Height	Mean ± SD	171.1 ± 11.2 cm	171.7 ± 10.4 cm	167.7 ± 10.3 cm
Race	White	79%	75%	55%
	Asian	8%	17%	30%
	Black	4%	8%	9%
	Other/unknown	8%	0%	6%

Table 2
Mean absolute error values for the STRIDE, CoagClinic, and IWPC databases.

Dataset	STRIDE (n = 335)	CoagClinic (n = 104)	IWPC (n = 5052)
MAE	14.1 mg/wk	10.4 mg/wk	10.1 mg/wk

3.2. Results for the CoagClinic database

In the CoagClinic cohort, we studied 104 patients (Table 1). The mean absolute error of the clinical equation was 10.4 mg/wk (Table 2). The clinical algorithm also showed 54.7% of patients with an *ideal dose*, 14.0% of patients with a *high dose*, and 31.4% of patients with a *low dose* (Fig. 2). Next, we compared the number of low, ideal, and high doses for the IWPC clinical algorithm using either the STRIDE cohort or the CoagClinic cohort (black vs. gray), yielding a χ^2 value of 24.3 ($p = 5.2 \times 10^{-6}$).

3.3. Comparison with IWPC cohort

In the original IWPC cohort of 5052 patients (Table 1), the mean absolute error of the clinical equation was 10.1 mg/wk (Table 2). The clinical algorithm demonstrated 46.6% of patients with an *ideal dose*, 25.7% of patients with a *high dose*, and 27.7% of patients with a *low dose* (Fig. 2). We compared the number of low, ideal, and high doses for the IWPC clinical algorithm using either the STRIDE cohort or the IWPC cohort (black vs. white), yielding a χ^2 value of 36.9 ($p = 9.6 \times 10^{-9}$). Comparing low, ideal, and high doses between the CoagClinic cohort and the IWPC cohort (gray vs. white) yielded a χ^2 value of 6.8 ($p = 0.033$), which was not under the threshold for multiple comparisons ($p = 0.017$).

4. Discussion

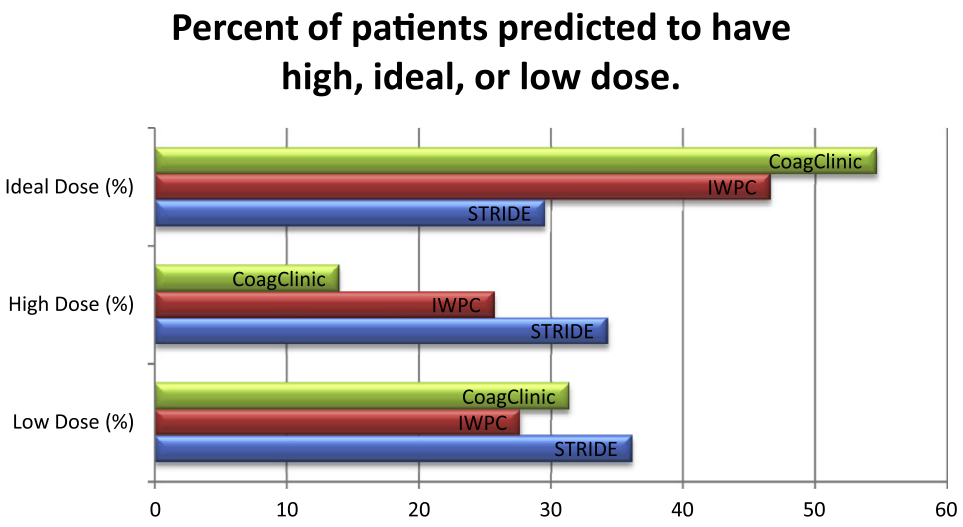
4.1. Utility of the clinical databases for exploratory analyses

Our results demonstrate that clinical warehouses can have data of high "functional" quality, in the sense that the CoagClinic data nearly matched IWPC research-grade data in its ability to predict warfarin dose. In addition, the distribution of low, ideal, and high doses was significantly different between STRIDE and the CoagClinic, but far closer between CoagClinic and the IWPC gold standard research cohort. Thus, although the general clinical database had data that could be used for estimating warfarin dosing, the noise in this data was considerably greater than the noise in another clinical practice database that was more closely focused on warfarin. In many ways, this is not surprising, but it provides a quantitative estimate of how much noise is introduced by clinical practice under two scenarios: close clinical attention to the variables of interest (almost no additional noise over research-grade data) and general clinical practice without specific focus (considerable additional noise).

When the results from STRIDE are compared to those from the CoagClinic, it is obvious that the signal is weaker. CoagClinic had lower mean absolute errors than STRIDE after applying the clinical dosing algorithm (Table 2). In addition, the MAE value for the CoagClinic cohort was close to the corresponding value for the IWPC cohort, but the MAE for the IWPC clinical algorithm was 40% *greater* for the STRIDE cohort compared to the IWPC cohort. Compared to CoagClinic, STRIDE had fewer patients within 1 mg/day of the actual dose, defined as having an ideal dose, and thus had more outliers (Fig. 2). When the distribution of low, ideal, and high doses is compared between the STRIDE and CoagClinic databases using the IWPC clinical algorithm, thus comparing the same algorithm on the two patient populations, the difference is substantial and significant ($p = 5.2 \times 10^{-6}$).

When the aforementioned distributions are compared between the STRIDE and IWPC cohorts, the differences are again highly significant ($p = 9.6 \times 10^{-9}$), with a greater proportion of patients having an ideal dose using the IWPC cohort. When the same distributions are compared between the CoagClinic and IWPC cohorts, the differences are only nominally significant ($p = 0.033$) and not under the threshold for multiple comparisons ($p = 0.017$). Interestingly, the CoagClinic cohort appears to have *slightly more* patients with an ideal dose even compared to the IWPC (Fig. 2). One potential reason why the IWPC algorithm does not work as well in the IWPC population is that algorithms tend to be more accurate within cohorts, rather than across cohorts. The CoagClinic represented one patient population using a single database, while the IWPC study included 23 sites from around the world with different data collection procedures. In addition, since the IWPC study contained a greater proportion of Asian patients, who on average have a lower warfarin dose [9], this could have impacted dose prediction, as the average dose in the IWPC population was slightly lower (Table 1). In any case, these data indicate that the distribution of dosing patterns is far more similar between the IWPC and CoagClinic cohorts ($p > 0.017$) compared to the STRIDE and CoagClinic cohorts ($p < 10^{-5}$).

Overall, the findings of this study are more likely due to differences in the quality of data collection rather than underlying differences in the patient populations from which the data was collected. The patient populations for STRIDE and CoagClinic were highly similar (Table 1). The mean stable dose for STRIDE was 35.7 ± 19.5 mg/wk, and the mean stable dose for CoagClinic was 35.4 ± 15.8 mg/wk, which was not statistically different using a t-test ($p = 0.89$). Additional statistical comparisons between CoagClinic and STRIDE included the following: age ($p = 0.33$), weight ($p = 0.56$), height ($p = 0.59$), and race ($p = 0.025$, not significant after adjusting for multiple comparisons). Despite these similarities, STRIDE and CoagClinic showed large differences in MAE and dosing patterns (Table 2, Fig. 2). Comparing the IWPC and CoagClinic (Table 1), the mean stable dose for the IWPC was 30.8 ± 16.8 mg/wk and the mean stable dose for CoagClinic was 35.4 ± 15.8 mg/wk, which was statistically different ($p = 0.0055$). Additional comparisons between the IWPC and CoagClinic included the following: age ($p = 0.56$), weight ($p = 0.16$), height ($p < 0.0001$), and race ($p = 0.0025$). Despite these differences, the IWPC and CoagClinic demonstrated very similar MAE values (Table 2), while the CoagClinic had somewhat better performance when measuring ideal dosing patterns (Fig. 2).



Percent:	Low Dose (%)	High Dose (%)	Ideal Dose (%)
CoagClinic	31.4	14.0	54.7
IWPC	27.7	25.7	46.6
STRIDE	36.1	34.3	29.6

Fig. 2. IWPC clinical algorithm applied to the STRIDE, CoagClinic, and IWPC databases.

Using STRIDE, it was challenging to determine whether patients had reached a stable dose based upon short clinical excerpts, especially without a history of INR values, so this was certainly a source of noise. CoagClinic, which was designed with hematology in mind, provides the dose and INR value associated with every visit, making it easy to determine whether patients had reached a stable dose. Moreover, the discharge summaries written for STRIDE were likely written by physicians who had less experience with anticoagulation than the staff at the Stanford Oral Anticoagulation Clinic. While an additional method for validating these results would involve manual chart abstraction of the de-identified records from STRIDE, doing this would require more extensive IRB approval.

This analysis excluded patients who were taking exactly 5 mg/day of warfarin, as it was far more likely that those discharge summaries reflected patients who had just started on warfarin and hence had not reached a stable therapeutic dose. However, some patients may by chance happen to be therapeutic at 5 mg/day, so they would have been excluded by this analysis. According to the data in the CoagClinic database, only 5.8% of the patients were therapeutic at 5 mg/day, so it is unlikely that this adjustment introduced large bias. Finally, patients taking 1 mg/day were excluded, as this is a standard fixed dose for certain indications, such as catheter placement, that is not adjusted on subsequent visits. For comparison, in the CoagClinic database there were no patients taking 1 mg/day. It should also be noted, however, that our results are particularly striking because the CoagClinic most likely serves patients that are generally more difficult to dose, so the good performance of the extracted data is even more impressive and potentially an underestimate of its data quality.

4.2. Opportunities for the improvement of clinical databases for research

This study offers a number of suggestions for the improvement of clinical databases, especially as a tool for research. Of course, data should be explicitly coded with controlled, searchable terminologies whenever feasible. In this case, height, weight, and dose had to be manually extracted from the text, which was time-consuming for 1472 records. This was problematic since many of the query results did not yield an explicit numerical value for these variables. For instance, a record might have contained the word "weight" in it without listing a numerical value, and the query would have interpreted the record as if it had a value for weight. On the other hand, values for age and race were automatically provided by the query, so these values were never missing. In addition, there should be options for specifically searching pharmacy data (i.e. drugs, doses, and frequency) as well as laboratory values (i.e. INR).

A long term advantage of a warehouse like STRIDE is the ability to gather a much wider range of potentially relevant data than is available in a specialty-focused clinical database. The focused database only collects variables known to be useful, and thus is limited in its utility for discovery. Our results suggest that large scale data-mining efforts in general-purpose clinical databases may see a degradation of signal of as much as 40%. However, it is important to note that the amount of signal degradation is also affected by differences in the capture and representation of data, as some databases contain free text, while others have a more structured model for capturing data. In our study, the STRIDE database was a better example of the free-text model, and the structured organization of data in the CoagClinic database likely improved its performance. This issue is particularly important for efforts such as those within the eMerge Network, a consortium that seeks to combine genetic data and data from electronic medical records, and dbGaP, a public repository of genotypic and phenotypic data, which are examples of resources that could be used for such studies in the future [1,2].

5. Conclusions

Ultimately, this study found that clinical warehouses can certainly provide data that is comparable to research-grade data, but high-quality data is more likely when extracted from cohorts who are in clinical settings where the data of interest is clinically critical. It is not surprising that the quality of warfarin-related data in a warfarin clinical database is much better than the quality of warfarin-related data in a general-purpose clinical

database! These results suggest, however, that cohort-finding in large clinical data warehouses should take special care to filter for the clinical setting in which the data was collected in order to best find reliable data, using the clinical practice relevance as a proxy for data quality.

Our results can be extended by assessing the quality of data retrieved by other clinical databases, either by studying warfarin dose or any other clinical outcome. The American Recovery and Reinvestment Act has authorized approximately $38 billion for health-related information technology between 2009 and 2019, and provides an ideal opportunity to build and validate clinical data collections for research and discovery [10].

References

[1] Manolio TA. Collaborative genome-wide association studies of diverse diseases: programs of the NHGRI's office of population genomics. Pharmacogenomics 2009;10:235–41.
[2] Mailman MD, Feolo M, Jin Y, Kimura M, Tryka K, Bagoutdinov R, et al. The NCBI dbGaP database of genotypes and phenotypes. Nat Genet 2007;39:1181–6.
[3] Butte AJ. Medicine. The ultimate model organism. Science 2008;320:325–7.
[4] Elias DJ, Topol EJ. Warfarin pharmacogenomics: a big step forward for individualized medicine: enlightened dosing of warfarin. Eur J Hum Genet 2008;16:532–4.
[5] Budnitz DS, Shehab N, Kegler SR, Richards CL. Medication use leading to emergency department visits for adverse drug events in older adults. Ann Intern Med 2007;147:755–65.
[6] Klein TE, Altman RB, Eriksson N, Gage BF, Kimmel SE, Lee MT, et al. Estimation of the warfarin dose with clinical and pharmacogenetic data. N Engl J Med 2009;360:753–64.
[7] The STRIDE Research Data Repository. Available at: http://clinicalinformatics.stanford.edu/STRIDE/.
[8] The CoagClinic Database. Available at: http://www.standingstoneinc.com/ehealth_coag.html.
[9] Rieder MJ, Reiner AP, Gage BF, Nickerson DA, Eby CS, McLeod HL, et al. Effect of VKORC1 haplotypes on transcriptional regulation and warfarin dose. N Engl J Med 2005;352:2285–93.
[10] Cunningham R. Stimulus bill implementation: expanding meaningful use of health IT. NHPF Issue Brief 2009:1–16.