Application and Optimisation of the Comparison on Extreme Laboratory Tests (CERT) Algorithm for Detection of Adverse Drug Reactions: Transferability Across National Boundaries

SHORT RUNNING TITLE: Optimisation of CERT to Detect ADRs

Mun Yee Tham,[1] Qing Ye,[1,2] Pei San Ang,[1] Liza Y. Fan,[1,2] Dukyong Yoon,[3,4] Rae Woong Park,[3,4] Zheng Jye Ling,[5] James W. Yip,[5] Bee Choo Tai,[6] Stephen Evans,[7] Cynthia Sung[1,8]

[1] Vigilance and Compliance Branch, Health Sciences Authority, Singapore
[2] Genome Institute of Singapore, Agency for Science and Technology, Singapore
[3] Department of Biomedical Informatics, Ajou University School of Medicine, Suwon, Korea
[4] Department of Biomedical Sciences, Ajou University Graduate School of Medicine, Korea
[5] Academic Informatics Office, National University Health System, Singapore
[6] Saw Swee Hock School of Public Health, National University of Singapore, Singapore
[7] Department of Medical Statistics, Faculty of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, London, UK
[8] Health Services and Systems Research, Duke-NUS Medical School, Singapore

CORRESPONDING AUTHOR:

Cynthia Sung

11 Biopolis Way, #11-03, Singapore 138667|Tel: +65 8123 1495 |Fax: +65 6478 9069 | |cynthia_sung@hsa.gov.sg |

KEYWORDS:

Adverse reaction
Data mining
Pharmacovigilance
Signal detection
Laboratory abnormality
Electronic Medical Records

KEY POINTS:

1. The Comparison on Extreme Laboratory Tests (CERT) algorithm was implemented in the electronic medical records (EMR) of a major tertiary hospital in Singapore, the National University Hospital (NUH).
2. A modified version of CERT that requires a minimum of 400 cases to assess a drug-laboratory abnormality (CERT400) yielded higher positive predictive value and sensitivity.
3. CERT400 demonstrated potential in detecting drug induced hepatic and renal toxicities, but limited utility in detecting ADRs associated with haematopoiesis and coagulation.

ACKNOWLEDGEMENTS

The authors are deeply grateful to A/Prof Cheng Leng Chan, Group Director of the Health Products Regulation Group (HPRG) and Dr. Dorothy Toh, Assistant Group Director, HPRG Post-market cluster at the Health Sciences Authority for steadfast support to pursue this research

to advance the organization's pharmacovigilance mission. We also would like to thank Siew Har Tan, Belinda Foo, Sally Soh and Sowmya Rudrappa at HSA for helpful discussions, Michelle Ng for assistance with creating the reference standard, and Michael Winther for his valuable programmatic coordination for the SAPhIRE (Surveillance and Pharmacogenomics Initiative for Adverse Drug Reactions) Project.

This study was conducted under the SAPhIRE Project, funded by a Strategic Positioning Fund grant from the Biomedical Research Council of the Agency for Science, Technology and Research of Singapore (SPF2014/001).

CONFLICT OF INTEREST

The authors declare no conflict of interest.

WORD COUNT

3192

This study has neither been previously published nor presented at any conferences.

Abstract

Purpose: The Singapore regulatory agency for health products (Health Sciences Authority), in carrying out active surveillance of medicines and their potential harms, is open to new methods to achieve this goal. Laboratory tests are a potential source of data for this purpose. We have examined the performance of the Comparison on Extreme Laboratory Tests (CERT) algorithm, developed by Ajou University, Korea, as a potential tool for adverse drug reaction (ADR) detection based on the electronic medical records (EMR) of the Singapore healthcare system.

Methods: We implemented the original CERT algorithm, comparing extreme laboratory results pre- and post-drug exposure, and five variations thereof using 4.5 years of National University Hospital (NUH) EMR data (31,869,588 laboratory tests, 6,699,591 drug dispensings from 272,328 hospitalizations). We investigated six drugs from the original CERT paper and an additional 47 drugs. We benchmarked results against a reference standard we created from UpToDate® 2015.

Results: The original CERT algorithm applied to all 53 drugs and 44 laboratory abnormalities yielded a PPV and sensitivity of 50.3% and 54.1%, respectively. By raising the minimum number of cases for each drug-laboratory abnormality pair from 2 to 400, the PPV and sensitivity increased to 53.9% and 67.2%, respectively. This post-hoc variation, named CERT400, performed particularly well for drug-induced hepatic and renal toxicities.

Discussion: We have demonstrated that the CERT algorithm can be applied across national boundaries. One modification (CERT400) was able to identify ADR signals from laboratory data with reasonable PPV and sensitivity, which indicates potential utility as a supplementary pharmacovigilance tool.

Text

INTRODUCTION

Traditionally, spontaneous reporting systems (SRS) have been the predominant data source for the detection of signals of adverse reactions.[1-3] This system, usually maintained by a government agency, receives suspected adverse drug reaction (ADR) reports submitted by healthcare professionals, pharmaceutical companies and consumers.[1-3] With the expanding use of electronic medical records (EMRs) in recent years, the pharmacovigilance community has another potentially rich source of information for drug safety surveillance.[1,3] The prospect of scanning EMRs is attractive, as it overcomes some of the limitations inherent in the SRS: (1) reliance mainly on voluntary reporting from its contributors, and susceptibility to under-reporting as well as over-reporting (e.g. due to media interest), (2) incomplete or missing data, hindering causality assessment, and (3) difficulty in detecting duplicate reports.[1, 2]

As EMRs are used for the clinical management of patients, they constitute an information-rich database[3] of patients' demographics, medications, past medical history, laboratory results, etc, which are commonly missing from ADR reports. The records reflect actual clinical practice, allowing for a more complete benefit-risk assessment. For specific ADRs, mining of EMRs has the added advantage of applying a consistent phenotype definition, thus overcoming variations in diagnostic criteria by different clinicians. However, unlike in SRS where a clinician has made a connection between the drug and an adverse event and files a report in a standardized format, much of EMR data are unstructured and housed in different databases. Pre-processing and data cleaning are required to extract and collate critical elements, such as drug exposure, concomitant medications, laboratory results, temporal relationships, and possible confounders.[1]

The Comparison of Extreme Laboratory Test (CERT) algorithm was developed by Korean researchers who applied it to 10 different drugs over 10 years of EMR data from Ajou University Hospital.[4] For each patient exposed to a particular drug, the algorithm selects the extreme laboratory test result (minimum or maximum) among multiple laboratory values from each of the pre-drug and post-drug exposure periods. CERT then determines whether a cohort of exposed patients demonstrates a significant change in abnormal laboratory values after drug exposure. As a regulatory agency seeking to build a toolkit of methods for active surveillance, the Health Sciences Authority (HSA), Singapore, sought to investigate the potential applicability of the CERT algorithm on the EMR in the Singapore healthcare system. The CERT algorithm had many desirable features that we were seeking, namely a temporal relation between drug exposure and a laboratory abnormality, the flexibility to evaluate any drug and laboratory test, and good performance metrics. Utilisation of numerical laboratory values before and after drug exposure made it potentially more portable across national boundaries, regardless of the language of the country. The objectives of this work were to implement and test CERT on the EMRs of the National University of Hospital (NUH), examine variations that could improve predictive performance, and assess its potential utility as a pharmacovigilance tool.

METHODS

Data source

De-identified EMRs were obtained from NUH, a 1,230-bed tertiary hospital, following approval of the study by the National Healthcare Group Domain Specific Review Board. The information retrieved included patient demographics, admission and discharge dates, inpatient drug dispensings, and laboratory test results over a 4.5 year period from July 2009 to Dec 2013. The data comprised over 31 million (31,869,588) laboratory tests and 6 million (6,699,591) inpatient drug dispensed orders from 272,327 hospitalization visits for 158,096 patients.

Selection of drugs for evaluation

Among the ten drugs analysed in the original CERT paper, one drug (ketorolac) was not used at NUH, while three oncologic drugs (etoposide, fluorouracil and methotrexate) were incompletely captured because oncologic drugs are mainly ordered and recorded in another database. In order to have a direct head-to-head comparison of algorithm performance from the EMRs of two different healthcare institutions, we first analysed only six drugs described in the original CERT publication (Round 1, Table 1).

In Round 2, we investigated an additional 47 drugs (Table 2). Factors considered in drug selection were drug usage volume and the likelihood of the drug being started during hospitalisation. Drugs with high usage were prioritised to provide sufficient number of cases for analysis. We also included negative controls (chlorpheniramine, metronidazole and risedronic acid) with no ADRs detectable by abnormal laboratory test results in the reference standard.

The CERT algorithm and variations

The original CERT algorithm paper examined 41 laboratory tests and 51 laboratory abnormalities. Six laboratory tests and seven laboratory abnormalities were not included in our evaluation because the laboratory results were infrequently ordered by clinicians in NUH (most of the 53 drugs had zero cases). Consequently, our evaluation included 35 laboratory tests and 44 laboratory abnormalities (Supplementary Table S1).

A common issue in assessing EMR data mining is the need for a benchmark reference standard to evaluate algorithm performance.[1,4] The original CERT publication used the 2010 version of UpToDate® Drug Information Database (UpToDate Inc, Waltham, MA, USA) to create a reference standard. We used the 2010 version of UpToDate® to directly compare our results with those previously published. To evaluate the performance of the original CERT algorithm and variations for all 53 drugs, two pharmacists constructed an updated reference standard from UpToDate® 2015 (Supplementary Table S3). As CERT utilises laboratory abnormalities as a surrogate of ADRs, the ADRs were mapped to their respective laboratory abnormalities using the mapping table described in the original CERT paper.[4]

In the original CERT algorithm, a case is defined when (1) the patient was prescribed the study drug at least once, and (2) at least one laboratory test result exists in each of the pre-drug and

post-drug periods. A minimum of two cases was required for CERT to run the statistical tests and generate output. If either the paired t-test or McNemar's test had $P<0.05$, the drug-laboratory-abnormality pair would be considered a positive signal.

Four variations of the original CERT algorithm were assessed on the set of 53 drugs (Table 3): (1) Limiting the period of observation to a defined period after the start of drug exposure, (2) Limiting the post-drug exposure observation period to a defined number of laboratory tests, (3) Taking an average of the two most extreme values instead of using only one extreme pre- and post-drug value, and (4) Using the paired t-test and non-parametric Wilcoxon's signed-rank test instead of paired t-test and McNemar's test. A fifth variation was also assessed post-hoc in which only drug-AE pairs with a minimum of 400 cases were included. The rationale for these variations is presented in the Discussion section.

Evaluation metrics

To evaluate the performance of CERT, we compared the drug-laboratory-abnormality pairs detected as significant signals by CERT with those identified in the reference standard. We then calculated the average positive predictive value (PPV), negative predictive value (NPV), sensitivity and specificity for each drug based on laboratory abnormalities. The F-score, which is the harmonic mean of PPV and sensitivity, is also reported (Supplementary Table S2). To contrast the results from different variations of the algorithm and get a pooled point estimate of the performance metrics and the 95% confidence interval, a random effects meta-analysis was performed to summarise a particular measurement of interest.

Creation of a reference standard

Supplementary Table S3 presents the reference standard created by mapping ADRs in UpToDate® 2015 for the 53 drugs to laboratory abnormalities. Researchers may find this a useful resource for benchmarking other algorithms intended to identify ADRs from laboratory abnormalities. However, it is worthwhile to note that this reference standard is not a list of confirmed ADRs, and is constantly being updated, and hence some may consider it a "silver standard" rather than a "gold standard". While UpToDate® contains information from multiple sources about a drug's safety profile, ADRs that occur in specific population could be overlooked, and it may be incomplete for drugs that have only been recently approved.

RESULTS

Evaluation of CERT performance

The PPV, NPV, sensitivity and specificity for Round 1 (6 drugs) are summarised in Table 4A. When comparing the same drugs between NUH and Ajou University, our results showed similar PPV (55.6% vs 58%) and better specificity (64.3% vs 52.2%). We had lower NPV (56.2% vs 66.7%) and sensitivity (48.9% vs 71.3%).

When the CERT algorithm was evaluated on a larger set of 47 drugs (Round 2) and benchmarked against an updated reference standard, PPV decreased to 48.5%, specificity was similar (65.2%), and sensitivity increased to 54.7%. Combining all 53 drugs evaluated in both

Rounds, overall PPV was 50.3%, specificity was 65.4%, and sensitivity was 54.1% for an F-score of 52.1% (Table 4B).

Performance across different laboratory panels

Consistent with Ajou University's findings, the majority of the signals (93.6%) detected by CERT were from "haematopoiesis and coagulation", "hepatobiliary enzymes" and "renal function and urine tests", and these panels were associated with higher F-scores compared to the remaining laboratory panels. However, decreases in red blood cells, white blood cells, neutrophils, haematocrit, as well as haemoglobin were found for all of the drugs (with the exception of alendronic acid). Therefore, CERT may not be particularly discriminating for drug effects on those laboratory tests.

In "hepatobiliary enzymes", ALT and AST showed good PPV (92%, 87%) and specificity (83%, 83%), suggesting potential utility in detecting hepatotoxicity signals. The trade-off is the lower sensitivity (59%, 32%), potentially missing some valid signals. For the "renal function and urine tests", creatinine showed good PPV (80%) and specificity (93%) but very low sensitivity (11%). BUN had good PPV (77%) and sensitivity (62%). Many signals for increased potassium were detected which were not reported in UpToDate®.[5] The lipid and metabolism, hormones and other panels also had high specificity (>87%) but low sensitivity (16-17%). PPV was also low, presumably because ADRs related to these abnormalities are rarer.

Performance across different variations

Among the four initial variations, Variations 1, 3 and 4 generally did not perform better than the original algorithm (Table 6, Figure 2). Variation 2 had the best specificity (76.8%, Table 6). However, this was accompanied by a large drop in sensitivity (38.3%). When we examined the evaluation metrics as a function of number of cases, we noted that sensitivity increased above 50% when there were 400 or more cases (Figure 1). Increasing the minimum number of cases from two to 400 cases for each drug-laboratory abnormality pair appears to better control the number of false negatives, as expected from increased power of a larger sample size. Hence, we performed a post-hoc analysis by imposing a threshold of 400 cases (Variation 5). Variation 5, not surprisingly because of its post-hoc nature, gave the best overall performance (PPV 53.9%, sensitivity 67.2%, F-score 59.8%), and hereafter is referred to as CERT400.

Negative controls

We tested CERT on three negative controls: chlorpheniramine, metronidazole and risedronic acid. These drugs have no signals in the reference standard that would be indicative of laboratory abnormalities. Yet, for all three drugs, CERT detected decreases in red blood cell, white blood cell, neutrophil, haematocrit, haemoglobin, and protein, as well as increases in platelets and alkaline phosphatase (ALP). As noted above, most haematopoeisis tests returned positive results for all drugs, therefore these tests are of limited utility for ADR signal detection using CERT.

DISCUSSION

As a drug regulatory authority responsible for monitoring the post-market safety of drugs, HSA has been interested in supplementing SRS with other methodologies to strengthen the system for identifying drug safety signals. Knowledge of the full safety profile of a drug, particularly for rarer adverse reactions, only becomes available through post-marketing surveillance from drug usage in actual clinical practice across a broad population.[6] With EMRs, new opportunities have arisen to mine these information-rich resources for safety signals. Here, we have shown that the CERT algorithm, which utilises laboratory test data in a temporal relationship with drug exposure, can be implemented on EMR data in a healthcare institution from another country with a different population. We have examined the performance of CERT for 53 drugs, of which direct comparison could be performed for 6 drugs in both countries. We also investigated 5 variations of the original CERT algorithm, and identified two that improve specificity and/or sensitivity.

The PPV of CERT was high for the liver enzymes ALT and AST and renal tests serum BUN and creatinine (92%, 87%, 77% and 80%, respectively), thus a positive signal from CERT is likely to signify hepatic and renal toxicity. However, the aminoglycosides gentamicin and amikacin, which are known to be nephrotoxic, did not show a positive signal with either increased creatinine or BUN. This could be a result of close monitoring of renal function and/or therapeutic drug monitoring by clinicians to prevent any acute renal injury, thereby dampening the incidence of a well-known signal. CERT did detect a signal of raised creatinine for other drugs with known nephrotoxic potential (e.g. hydrochlorothiazide, ranitidine, cefazolin), but sensitivity of the serum creatinine test was low (11%). However, sensitivity for BUN was much higher at 62%.

CERT appears to be less discriminating for the hematopoiesis and coagulation panel, as nearly every drug had one or more signals in this panel. This may be more a reflection of the course of the disease or treatment. Similarly, the high number of false positives with potassium may be due to the high incidence of hyperkalaemia (up to 10%) in hospitalised patients, as many conditions can affect potassium levels (e.g. transcellular shifts, impaired excretion, or increase in potassium intake).[5] Indeed, a major limitation of the CERT algorithm is the lack of adjustment for confounding factors. The CLEAR algorithm[7] , also developed by the Ajou University group, controls for confounder effects with the use of matched controls having the same admitting department and diagnosis but who had not taken the drug, but CLEAR is much more computationally intensive. Another limitation is the lack of an actual gold standard for ADRs. Even though we created a reference standard using UpToDate 2015, we cannot confirm that the ADRs listed are indeed true ADRs. In addition, for chronic medications (e.g. simvastatin, enalapril), patients might already be taking them prior to admission. As such, the pre-exposure laboratory test results retrieved by CERT for these cases may not be true, potentially diluting any positive signals.

In the original CERT algorithm, all tests in the pre- and post-drug exposure period were included. However, it was often the case that many more tests were ordered during the post-drug exposure period, which tends to inflate the chance finding of an abnormal result in the post-drug exposure period. By limiting the number of tests in the post-drug exposure period to two more than the number in the pre-drug exposure period (Variation 2), we observed an

increase in the specificity of CERT from 65.4% to 76.8%. The original CERT algorithm counted a drug-laboratory abnormality pair if there were at least two cases. We found that raising the minimum to 400 cases for each drug-laboratory abnormality pair (CERT400) helped to reduce the false negative rate, increasing the sensitivity from 54.1% to 67.2%. However, since the choice of 400 is based on results in these data, these estimates may be biased upwards. Although these performance metrics are not sufficiently high to solely rely on CERT400 for active surveillance, it promises to be a valuable addition to the toolkit for postmarket surveillance. A drug-laboratory abnormality pair identified by an automated CERT400 algorithm could then be further evaluated by other methodologies such as text mining of discharge summaries[8-10] to determine the validity of the signal. In the case of infrequently ordered laboratory tests, rarely used drugs, or newly approved drugs where usage has yet to pick up, the use of CERT400 may hinder detection of safety signals, since there may be insufficient cases to meet the minimum. With the growth of electronic data in healthcare databases and linkages across multiple institutions, however, we anticipate that the rising volume of data will overcome the limitation of having this threshold of cases for evaluating the algorithm.

Other groups have been investigating a variety of data mining methodologies to query health records for identification of ADRs based primarily on clinical features.[11-16] One notable effort is the Sentinel Initiative funded by the United States Food and Drug Administration. Specific queries of interest are submitted to the Sentinel coordinating center, which sends computer programs to data partners to extract and aggregate data on administrative and insurance claims data of over 180 million subjects.[17] The Sentinel group successfully identified intussusception after rotavirus vaccination in infants[13] and risk of coeliac disease in patients on long-term therapy with olmesartan[18, 19] from algorithms applied to their extensive databases. The Observational Medical Outcomes Partnership investigated methods that relied primarily on diagnosis codes in administrative databases.[11, 20, 21] Our dataset had limited structured diagnostic coding, which made it challenging for us to explore algorithms that rely on codes such as ICD-9.

Methods using abnormal laboratory values for identifying ADRs have been investigated previously. In a study by Levy et al[22] , automatic laboratory signals were generated when a specific laboratory value met a pre-defined criteria and tested on a prospective cohort of 192 patients. A list of cases was generated for further manual review. The false positive rate throughout the entire study period was 83%, which is a likely barrier to implementation. Ramirez et al implemented a prospective program based on automatic laboratory signals (ALS) for 54,525 hospitalisations in Spain.[23] The algorithm flagged patients whose laboratory values met the criteria specified for six serious ADRs, but did not include a temporal relationship with drug intake, hence the cases needed to be manually reviewed to determine if the timing of drug intake could account for the abnormal laboratory values. The authors concluded that this was an intensive manual process requiring considerable effort.

Liu et al[3] aimed to have a more automated methodology that incorporated a temporal relationship with drug exposure. Abnormal laboratory results were correlated with specific drug administration by comparing the outcomes of drug-exposed and a matched unexposed

group; higher thresholds for categorizing a laboratory result as abnormal were used and a minimum of 25 cases was required. When benchmarked with two reference datasets (the same 10 drugs evaluated by Yoon et al. or 9 other drugs), the reporting odds ratio (ROR) method performed best when applied to an EMR database containing four times more unique patients than the NUH database (PPV 77% and 58% respectively, sensitivity 61% and 67%, respectively).

In summary, we have demonstrated the transferability of the CERT algorithm to a health care institution of another country. We have developed a reference standard of drug-laboratory abnormalities for 53 drugs based on the 2015 version of UpToDate® to evaluate CERT on our data, and which can also be used to benchmark other published algorithms. CERT400, a modification of CERT which only accepts results generated from more than 400 drug-laboratory abnormality cases, gave the best overall performance with a PPV of 53.9% and sensitivity of 67.2% (F-score 59.8%). High PPV for increased AST and ALT enzyme levels, BUN and serum creatinine suggests that CERT would be particularly useful for identifying drug-induced hepatic and renal toxicities. The ability of CERT400 to sift through a large volume of laboratory tests obtained before and after drug exposure and identify potential signals with reasonable positive predictive value and sensitivity indicates that it will be a useful tool to add to a pharmacovigilance programme.

Tables and Figures

Table 1: Database size and number of patient visits used for CERT comparison between Ajou University Hospital and NUH

	Ajou University Hospital	NUH
Number of patient visits*	1,011,055	272,328
Drug	Number of Visits	
Ciprofloxacin	16,706	17,576
Clopidogrel	19,188	28,672
Levofloxacin	9,059	4,673
Ranitidine	68,995	7,474
Rosuvastatin	4,811	2,252
Valproic acid	11,523	4,300

*Number of hospitalizations during which the patient received at least one dispensing of the drug.

Table 2: Number of patient visits for 47 drugs used in Round 2

Drug	No. of patient visits	Drug	No. of patient visits
Aciclovir	5,353	Ezetimibe	2,809
Alendronic acid	2,009	Famotidine	23,590
Allopurinol	8,694	Fenofibrate	6,552
Amikacin	1,629	Fluconazole	3,566
Amoxicillin-Clavulanic Acid	116,511	Gentamicin	13,512
Ampicillin	4,850	Gliclazide	8,622
Azithromycin	5,812	Hydrochlorothiazide	6,811
Carbamazepine	1,373	Imipenem-cilastatin	3,509
Carvedilol	40,013	Isoniazid	1,610
Cefazolin	26,750	Itraconazole	629
Ceftazidime	7,228	Levetiracetam	7,955
Ceftriaxone	74,834	Losartan	19,158
Celecoxib	3,266	Meropenem	26,273
Chlorpheniramine	26,790	Metformin	55,881
Clarithromycin	17,382	Metronidazole	22,212
Clindamycin	3,869	Omeprazole	163,999
Cloxacillin	6,505	Phenytoin	5,136
Cotrimoxazole	14,969	Piperacillin-Tazobactam	14,843
Digoxin	17,338	Pyrazinamide	1,061
Domperidone	10,004	Rifampicin	1,758
Doxycycline	1,813	Risedronic acid	1,551
Enalapril	57,621	Simvastatin	81,590
Entecavir	1,368	Vancomycin	34,028
Ethambutol	1,328		

353 **Table 3: Variations of the original CERT algorithm**

Algorithm	Input (laboratory test results)	Observation period after drug exposure	Minimum no. cases required	Statistical test
Original	Extreme values	Till discharge	2	Paired t-test & McNemar's test
Variation 1	Extreme values	12 days	2	Paired t-test & McNemar's test
Variation 2	Extreme values	(n+2) laboratory test results	2	Paired t-test & McNemar's test
Variation 3	Average of the two most extreme values	Till discharge	2	Paired t-test & McNemar's test
Variation 4	Extreme values	Till discharge	2	Paired t-test &Wilcoxon's signed-rank test
Variation 5 (CERT400)	Extreme values	Till discharge	400	Paired t-test & McNemar's test

Variation 1: *Limit the unit of observation to 12 days after drug exposure, which is the mean (7.1 days) plus 2 times the standard deviation (2.6 days) of the time from drug administration to the time the extreme post-drug laboratory value was taken for all true positive cases.*
Variation 2: *Limit the unit of observation to a maximum of (n+2) laboratory test results for each encounter, where n is the number of laboratory tests before the drug was started.*
Variation 3: *Take the mean of the two most extreme values for both the pre- and post-drug periods, to minimize the possibility that a single extreme value, such as one caused by measurement error, would unduly influence the result.*
Variation 4: *Use the non-parametric Wilcoxon's signed-rank test to replace the McNemar's test.*
Variation 5 (Post-hoc variation): *Impose a minimum of 400 cases for each drug-laboratory test pair based on results shown in Figure 1.*

354

355

356

357

Table 4A: Performance Metrics of CERT: Comparison between NUH and Ajou*

	PPV (%)	NPV (%)	Sensitivity (%)	Specificity (%)	F-score (%)	Avg. No. of signals detected per drug
NUH	57.0	58.1	48.6	63.9	52.5	113/6 = 18.8
Ajou University Hospital	57.4	66.6	69.4	52.1	62.8	155/6 = 25.8

Average performance for the 6 drugs from original paper

**Results are benchmarked according to the 2010 Version of UpToDate used in original CERT paper. Six drugs included in the comparison are ciprofloxacin, clopidogrel, levofloxacin, ranitidine, rosuvastatin, and valproic acid*

Table 4B: Performance Metrics of CERT: Comparison between Rounds 1 and 2 based on NUH EMR data*

	No. of evaluated drugs	PPV (%)	NPV (%)	Sensitivity (%)	Specificity (%)	F-score (%)	Avg. No. of signals detected per drug
Round 1	6	63.4	55.1	49.8	66.8	55.8	113/6 = 18.8
Round 2	47	45.6	70.6	55.5	64.4	50.1	885/47 = 18.8
Total	53	47.6	68.9	54.8	64.6	51.0	998/53 = 18.8

**Results are benchmarked according to 2015 Version of UpToDate*

Table 5: Performance Metrics of CERT algorithm for 53 drugs – Comparison across Laboratory Panels*

A) Analysis according to the original CERT algorithm for laboratory panels or selected tests

Laboratory Panel or Test	Avg. no. of cases	No. of positive Signals	PPV (%)	NPV (%)	Sensitivity (%)	Specificity (%)	F-score (%)
Hemato-poiesis and coagulation	2634	600	44.7	71.7	67.3	49.8	53.7
Hepatobiliary enzymes	1507	228	63.6	31.5	59.7	35.2	61.6
AST	1532	15	86.7	26.3	31.7	83.3	46.4
ALT	1533	26	92.3	37.0	58.5	83.3	71.6
Renal function and urine tests	3867	75	61.3	35.7	46.0	50.8	52.6
BUN	3835	30	76.7	39.1	62.2	56.3	68.7
CRE	3848	5	80.0	29.2	10.5	93.3	18.6
K	3917	40	47.5	53.8	76	25	58.5
Lipids and metabolism	80	32	25.0	82.8	16.7	88.9	20.0
Hormones	104	14	14.3	89.1	16.7	87.2	15.4
Others	1158	49	34.7	71.1	15.5	87.7	21.4

B) Analysis according to the CERT400 algorithm for laboratory panels or selected tests

Laboratory Panel or Test	Avg. no. of cases	No. of positive Signals	PPV (%)	NPV (%)	Sensitivity (%)	Specificity (%)	F-score (%)
Hemato-poiesis and coagulation	3663	512	48.8	64.9	74.9	37.2	59.1
Hepatobiliary enzymes	2255	177	64.4	33.9	73.5	25.0	68.7
AST	2042	12	91.7	26.9	36.7	87.5	52.4
ALT	2000	23	91.3	37.5	67.7	75.0	77.8
Renal function and urine tests	4408	69	62.3	36.2	49.4	49.0	55.1
BUN	4372	27	74.1	36.8	62.5	50.0	67.8
CRE	4386	5	80.0	29.3	12.1	92.3	21.1
K	4466	37	51.4	66.7	86.4	25	64.4
Lipids and metabolism	639	5	40	40	40	40	40
Hormones	567	1	0	100	Not valid	50	Not valid
Others	2682	45	33.3	73.6	34.1	73.0	33.7

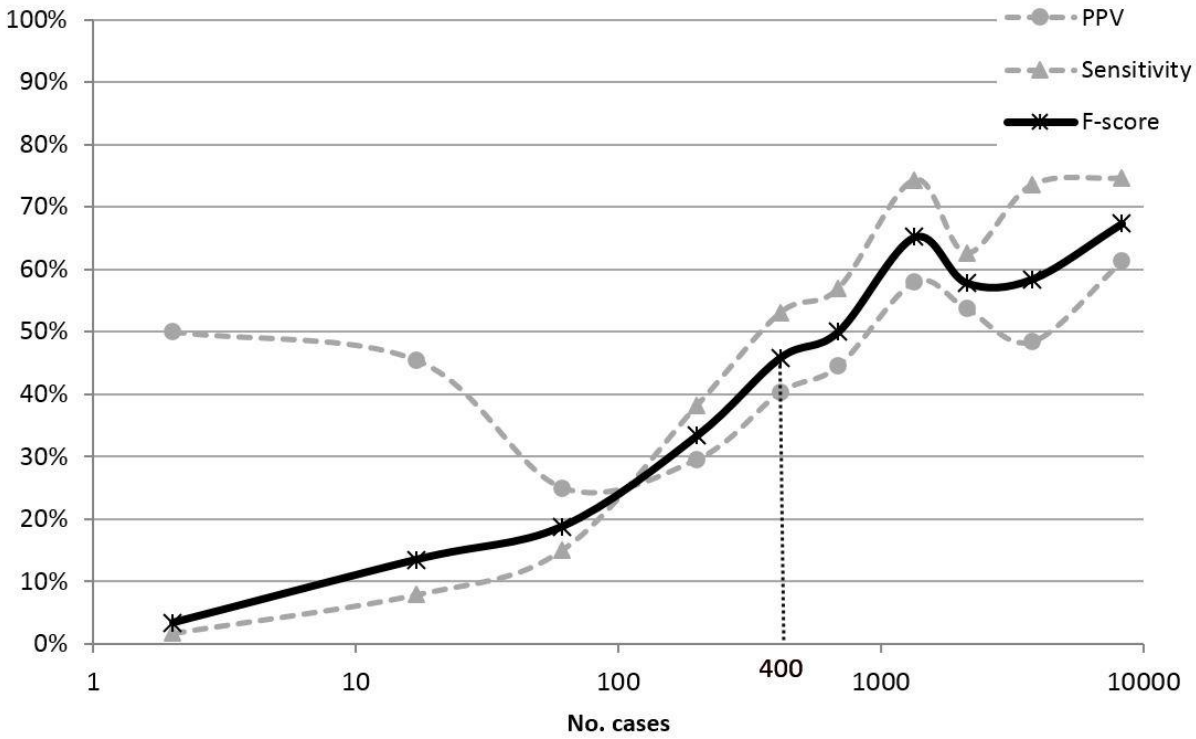
**Results are benchmarked according to 2015 Version of UpToDate*

Table 6: Performance Metrics for the Original CERT Algorithm and 5 Variations*

Variation	No. of cases (Avg)	No. of positive Signals	PPV (%)	NPV (%)	Sensitivity (%)	Specificity (%)	F-score (%)
Original	1899	998	50.3	67.2	54.1	65.4	52.1
Variation 1	1853	761	50.6	64.5	40.3	74.1	44.9
Variation 2	1899	707	51.6	64.6	38.3	76.8	44.0
Variation 3	1833	905	49.1	65.6	48.4	68.2	48.7
Variation 4	1898	1385	44.8	64.4	66.9	45.0	53.7
Variation 5 (CERT400)	2969	792	53.9	53.5	67.2	42.3	59.8

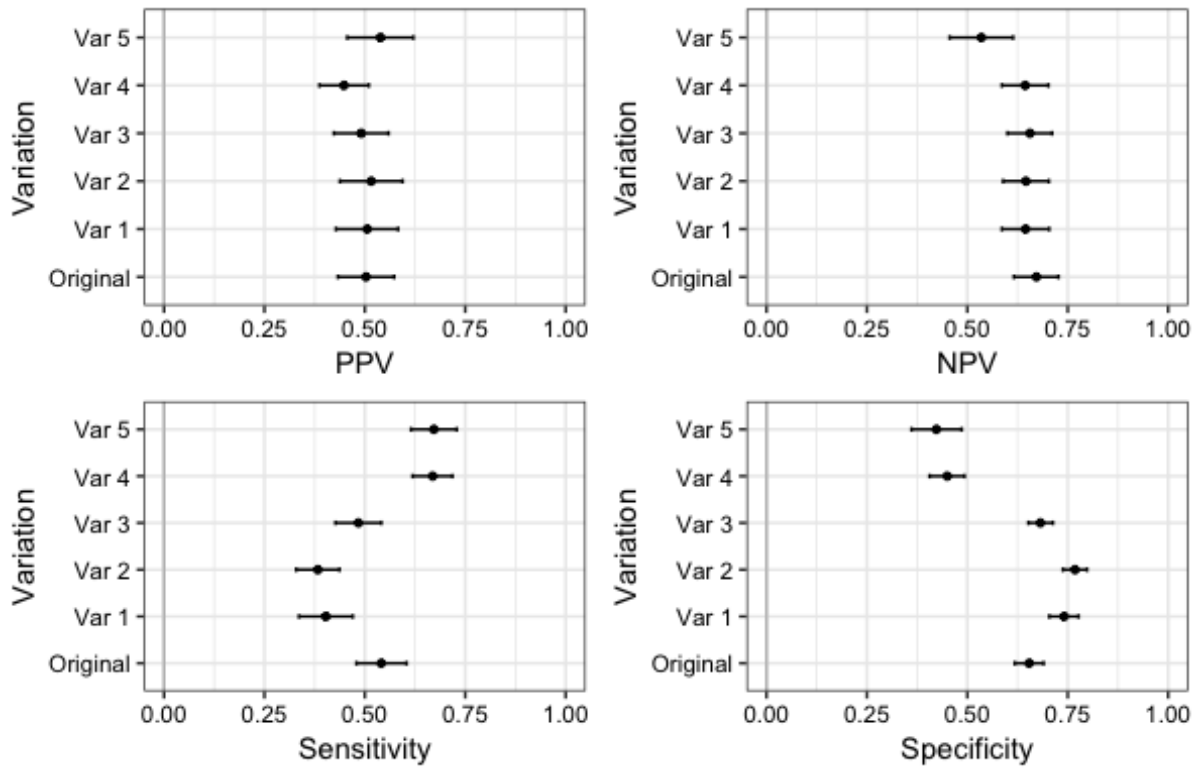
**Results are benchmarked according to 2015 Version of UpToDate for 53 drugs and 44 laboratory abnormalities*

Figure 1: CERT Algorithm: Dependence of PPV, Sensitivity and F-score on Number of Cases*



53 drugs, 44 laboratory abnormalities

Figure 2: Performance metrics and 95% confidence interval based on random effects meta-analysis for all the 53 drugs.



Supplementary tables: (Recommended to put as additional supporting information online)

Table S1: List of the 44 selected laboratory test abnormalities according to the laboratory panels.

Hematopoiesis and coagulation	
Activated partial thromboplastin time	Increased
	Decreased
Basophil	Decreased
Eosinophil	Increased
	Decreased
Fibrinogen	Decreased
Hematocrit	Decreased
Hemoglobin	Increased
	Decreased
Lymphocyte	Increased
Neutrophil	Decreased
Platelet	Increased
	Decreased
Prothrombin time	Increased
	Decreased
Red blood cell	Decreased
Reticulocyte	Increased
	Decreased
White blood cell	Increased
	Decreased
Hepatobiliary enzymes	
Alanine transaminase	Increased
Alkaline phosphatase	Increased
Aspartate transaminase	Increased
Direct bilirubin	Increased
Gamma-glutamyl transpeptidase	Increased
Protein	Decreased
Total bilirubin	Increased

Renal function and urine tests	
Blood urea nitrogen	Increased
Creatinine	Increased
Potassium	Increased
Lipids and metabolism	
Cholesterol	Increased
Glucose	Increased
	Decreased
LDL cholesterol	Increased
Triglyceride	Increased
Hormones	
Free thyroxine	Increased
	Decreased
Others	
Ammonemia	Increased
Amylase	Increased
Creatine kinase	Increased
Lactate dehydrogenase	Increased
Lipase	Increased
Sodium	Decreased
Uric acid	Increased

Table S2 (a): Evaluation metrics PPV, NPV, sensitivity (sens), specificity (spec) and F-score.

Algorithm Result	2010/2015 Version of UpToDate		
	Present	Absent	
Positive	TP	FP	$PPV = \frac{TP}{TP + FP}$
Negative	FN	TN	$NPV = \frac{TN}{FN + TN}$
	$sens = \frac{TP}{TP + FN}$	$spec = \frac{TN}{FP + TN}$	$F - score = \frac{2 \times PPV\ x\ Sensitivity}{PPV + Sensitivity}$

Table S2 (b): Example to show the performance metrics for ALT laboratory test abnormality on the 53 drugs we tested.

Algorithm Result (for ALT)	2015 Version of UpToDate (for ALT)		
	Present	Absent	
Positive	24	2	$PPV = \frac{24}{26} = 92\%$
Negative	17	10	$NPV = \frac{10}{27} = 37\%$
	$sens = \frac{24}{41} = 59\%$	$spec = \frac{10}{12} = 83\%$	$F - score = 72\%$

425 Table S3: Reference Standard of drug ADRs developed from UpToDate® 2015*

426


Table S3 - Ref Standard 1.pdf


Table S3 - Ref Standard 2.pdf

427

428 *This table is uploaded separately online as well.

429

REFERENCES:

1. Harpaz R, Dumouchel W, Shah NH, et al. Novel data-mining methodologies for adverse drug event discovery and analysis. *Clin Pharmacol Ther* 2012; **91**: 1010-1021. DOI: 10.1038/clpt.2012.50
2. Harpaz R, Vilar S, Dumouchel W, et al. Combing signals from spontaneous reports and electronic health records for detection of adverse drug reactions. *J Am Med Inform Assoc* 2013; **20**: 413-419. DOI: 10.1136/amiajnl-2012-000930
3. 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**: 420-426. DOI: 10.1136/amiajnl-2012-001119
4. Park MY, Yoon D, Lee K, et al. A novel algorithm for detection of adverse drug reaction signals using a hospital electronic medical record database. *Pharmacoepidemiol Drug Saf* 2011; **20**: 598-607. DOI: 10.1002/pds.2139
5. Viera AJ , Wouk N. Potassium Disorders: Hypokalemia and Hyperkalemia. *Am Fam Physician* 2015; **92**: 487-495.
6. Eichler HG, Abadie E, Breckenridge A, et al. Bridging the efficacy-effectiveness gap: a regulator's perspective on addressing variability of drug response. *Nat Rev Drug Discov* 2011; **10**: 495-506. DOI: 10.1038/nrd3501
7. Yoon D, Park MY, Choi NK, et al. Detection of adverse drug reaction signals using an electronic health records database: Comparison of the Laboratory Extreme Abnormality Ratio (CLEAR) algorithm. *Clin Pharmacol Ther* 2012; **91**: 467-474. DOI: 10.1038/clpt.2011.248
8. Wang X, Hripcsak G, Markatou M, et al. Active computerized pharmacovigilance using natural language processing, statistics, and electronic health records: a feasibility study. *J Am Med Inform Assoc* 2009; **16**: 328-337. DOI: 10.1197/jamia.M3028
9. Sohn S, Kocher JP, Chute CG, et al. Drug side effect extraction from clinical narratives of psychiatry and psychology patients. *J Am Med Inform Assoc* 2011; **18 Suppl 1**: i144-149. DOI: 10.1136/amiajnl-2011-000351
10. Kholghi M, Sitbon L, Zuccon G, et al. Active learning: a step towards automating medical concept extraction. *J Am Med Inform Assoc* 2016; **23**: 289-296. DOI: 10.1093/jamia/ocv069
11. Ryan PB, Stang PE, Overhage JM, et al. A comparison of the empirical performance of methods for a risk identification system. *Drug Saf* 2013; **36 Suppl 1**: S143-158. DOI: 10.1007/s40264-013-0108-9
12. Southworth MR, Reichman ME , Unger EF. Dabigatran and postmarketing reports of bleeding. *N Engl J Med* 2013; **368**: 1272-1274. DOI: 10.1056/NEJMp1302834
13. Yih WK, Lieu TA, Kulldorff M, et al. Intussusception risk after rotavirus vaccination in U.S. infants. *N Engl J Med* 2014; **370**: 503-512. DOI: 10.1056/NEJMoa1303164
14. Arfe A, Scotti L, Varas-Lorenzo C, et al. Non-steroidal anti-inflammatory drugs and risk of heart failure in four European countries: nested case-control study. *BMJ* 2016; **354**: i4857. DOI: 10.1136/bmj.i4857
15. Toh S, Hampp C, Reichman ME, et al. Risk for Hospitalized Heart Failure Among New Users of Saxagliptin, Sitagliptin, and Other Antihyperglycemic Drugs: A Retrospective Cohort Study. *Ann Intern Med* 2016; **164**: 705-714. DOI: 10.7326/M15-2568
16. Trifiro G, De Ridder M, Sultana J, et al. Use of azithromycin and risk of ventricular arrhythmia. *CMAJ* 2017; **189**: E560-E568. DOI: 10.1503/cmaj.160355
17. Curtis LH, Weiner MG, Boudreau DM, et al. Design considerations, architecture, and use of the Mini-Sentinel distributed data system. *Pharmacoepidemiol Drug Saf* 2012; **21 Suppl 1**: 23-31. DOI: 10.1002/pds.2336
18. Sentinel. (2013). FDA Safety Communication: FDA approves label changes to include intestinal problems (sprue-like enteropathy) linked to blood pressure medicine olmesartan medoxomil. *FDA Safety Communications.* Retrieved from https://www.sentinelsystem.org/communications/fda-safety-communications/296
19. Sentinel. (2013). Angiotensin Receptor Blockers (ARBs), hydrochlorothiazide, atenolol, amlodipine use & celiac disease. Retrieved from

https://www.sentinelsystem.org/drugs/assessments/angiotensin-receptor-blockers-arbs-hydrochlorothiazide-atenolol-amlodipine-use
20. Reich CG, Ryan PB , Suchard MA. The impact of drug and outcome prevalence on the feasibility and performance of analytical methods for a risk identification and analysis system. *Drug Saf* 2013; **36 Suppl 1**: S195-204. DOI: 10.1007/s40264-013-0112-0
21. Suchard MA, Zorych I, Simpson SE, et al. Empirical performance of the self-controlled case series design: lessons for developing a risk identification and analysis system. *Drug Saf* 2013; **36 Suppl 1**: S83-93. DOI: 10.1007/s40264-013-0100-4
22. Levy M, Azaz-Livshits T, Sadan B, et al. Computerized surveillance of adverse drug reactions in hospital: implementation. *Eur J Clin Pharmacol* 1999; **54**: 887-892.
23. Ramirez E, Carcas AJ, Borobia AM, et al. A pharmacovigilance program from laboratory signals for the detection and reporting of serious adverse drug reactions in hospitalized patients. *Clin Pharmacol Ther* 2010; **87**: 74-86. DOI: 10.1038/clpt.2009.185