Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220401

Accelerating High-Dimensional Temporal Modelling Using Graphics Processing Units for Pharmacovigilance Signal Detection on Real-Life Data

Pierre SABATIER[a,b,1] Jean FEYDY[b] and Anne-Sophie JANNOT[a]
[a] *AP-HP.Centre, Université de Paris, Paris, France*
[b] *INSERM INRIA HeKA team, Cordeliers Research Center, Paris, France*

Abstract. Adverse drug reaction is a major public health issue. The increasing availability of medico-administrative databases offers major opportunities to detect real-life pharmacovigilance signals. We have recently adapted a pharmaco-epidemiological method to the large dimension, the WCE (Weigthed Cumulative Exposure) statistical model, which makes it possible to model the temporal relationship between the prescription of a drug and the appearance of a side effect without any a priori hypothesis. Unfortunately, this method faces a computational time problem. The objective of this paper is to describe the implementation of the WCE statistical model using Graphics Processing Unit (GPU) programming as a tool to obtain the spectrum of adverse drug reactions from medico-administrative databases. The process is divided into three steps: pre-processing of care pathways using the Python library Panda, calculation of temporal co-variables using the Python library "KeOps", estimation of the model parameters using the Python library "PyTorch" - standard in deep learning. Programming the WCE method by distributing the heaviest portions (notably spline calculation) on the GPU makes it possible to accelerate the time required for this method by 1000 times using a computer graphics card and up to 10,000 times with a GPU server. This implementation makes it possible to use WCE on all the drugs on the market to study their spectrum of adverse effects, to highlight new vigilance signals and thus to have a global vigilance tool on medico-administrative database. This is a proof of concept for the use of this technology in epidemiology.

Keywords. Graphics Processing Unit, data-driven approach, signal detection, adverse drug reactions.

1. Introduction

Adverse drug reaction is a major public health issue. A recent study by Makary [1], in 2016, places drug-related iatrogeny as the 3rd leading cause of death in the United States. Currently, post-marketing surveillance of drugs is carried out by a pharmacovigilance system, which is based on spontaneous reports from healthcare professionals, pharmaceutical companies and patients. However, this system suffers from under-reporting, with less than 10% of serious adverse reactions being reported [2], and is

[1] Corresponding Author, Pierre Sabatier, Hôpital Européen Georges Pompidou, 20 rue Leblanc, 75015 Paris, France ; E-mail: Pierre.sabatier@aphp.fr

subject to bias due to selective reporting. Most of the reported cases are considered as suspected adverse reactions. The increasing availability of medico-administrative databases offers major opportunities to detect real-life pharmacovigilance signals [3]. However, these databases have mainly been used to confirm or refute potential signals reported [4]. It would be very appropriate to use these databases not to confirm a suspected adverse reaction but to generate adverse reaction signals. The challenge in developing a real-life pharmacovigilance signal detection method presents two difficulties: (a) The choice of a powerful method that can detect a signal among a very large number of drugs and therefore face the problem of multiple tests and therefore the generation of potential false positives, but also face problems of computer resources. (b) Taking into account the temporal heterogeneity of the link between drug intake and adverse effects, which may appear in the short or medium term depending on the pathophysiological mechanism causing the adverse effect.

Over the last decade, several international initiatives have worked on the development of pharmacoepidemiology tools: Mini-Sentinel [5], Observational Medical Outcomes Partnership [6], Pharmaco-epidemiological Research on Outcomes of Therapeutics by a European Consortium [7], Exploring and Understanding Adverse Drug Reactions [8], and the Asian Pharmacoepidemiology Network [9]. The methods compared in these working groups are methods using pharmacovigilance feedback databases (not exhaustive) or so-called "single-candidate" methods, i.e. methods that seek to prove or disprove a hypothesis. These methods do not address our signal detection problem. We have recently adapted a pharmaco-epidemiological method to the large dimension, the WCE (Weigthed Cumulative Exposure) statistical model [10], which makes it possible to model the temporal relationship between the prescription of a drug and the appearance of a side effect without any a priori hypothesis [11], taking into account the temporal heterogeneity described below. Unfortunately, this method faces a computational time problem.

Since 2007, Nvidia has been developing CUDA (originally an acronym for Compute Unified Device Architecture), which is a technology that uses a graphics processing unit (GPU) to perform general purpose computations in place of the processor (CPU), thus speeding up calculations. No methods in epidemiology have been implemented with this technology yet, although it has become widely used in other sectors [12].

The objective of this paper is to describe the implementation of the WCE statistical model using GPU programming as a tool to obtain the spectrum of adverse drug reactions from medico-administrative databases.

2. Methods

2.1. Data-driven WCE

The time-weighted cumulative exposure (WCE) model is used to estimate the cumulative effect of duration, dose and date of prescription of the study drug. The WCE approach estimates the effect of past exposure using a time-weighted sum of all previous exposure events. The weights depend on the time since exposure and the dose. The WCE model is based on a risk function modelled by a spline function and a Cox model with a time-dependent variable to model exposure to the drug of interest. This allows for adverse events that may occur shortly after the start of the study drug or after prolonged exposure. Adjustment variables such as age and gender can be introduced.

The method for estimating the parameters of the WCE model works in two steps. First, the parameters of the risk function are estimated and then the parameters of the Cox model are estimated. The risk function is estimated from the set of patients exposed to the study drug and will be different depending on the drug and/or the event. From this risk function, the WCE model, in a second step, calculates, based on a Cox model, the risk of having the event being exposed to the study drug. The event in our model can be the prescription of another drug, a hospitalization, an illness or a hybrid drug-disease event.

We adapted WCE to an approach without a priori hypothesis by proposing to take into account the multiple tests performed via a bootstrap approach. In the framework of the data-driven WCE extension, the precise consideration of the temporal aspect in the WCE model allows for a good ratio of true positives to false positives in the framework of an approach without a priori hypothesis. The current limitation of this method is its computation time.

In comparison, the most commonly used pharmacovigilance signal detection methods (without a priori hypothesis) are signal disproportionality methods. These methods roughly take into account the temporal aspect of the prescription of the drug (before/after the studied event) and generate many false positives.

2.2. WCE implementation on GPU

The process is divided into three steps:

- Pre-processing of care pathways using the Python library "Panda" - standard for processing tabular data (https://pandas.pydata.org/). Our raw data describes the set of prescriptions associated with a patient population. We extract a list of "times of interest" which correspond to the first intake of selected drugs, used as proxies for adverse side effects.
- Calculation of temporal co-variables using the Python library "KeOps" (https://www.kernel-operations.io). For each time of interest, the prior exposure to the study drug ("suspect") is quantified using correlations with temporal filters - in the example of the original WCE model, these are B-spline functions. The computation of these "temporal co-variables" is efficiently performed using the KeOps library's "block" reduction operations, which distribute the patients over the thousands of cores of our graphics cards.

- Estimation of the model parameters using the Python library "PyTorch" - standard in deep learning (https://pytorch.org/). Given the temporal co-variates computed in the previous step, estimating the maximum likelihood for the Cox/WCE model corresponds to minimizing a convex function with about ten free parameters. We find the solution to this problem using a quasi-Newton method (L-BFGS solver with "strong Wolfe" search), which converges in about ten iterations. We quantify the uncertainty of our estimator using a bootstrap method (N=1,000 to N=10,000 replicates), which is very efficiently implemented with PyTorch using a batch dimension.

We can express the run time of our algorithm as a big O(Its * Drugs * Events * Survivors), where Its is the number of iterations in the convex optimizer for the Cox-PH problem (at most 100), Drugs is the number of drugs of interest, Events is the number of “Death” events - in practice, the prescription of a precise drug that we use as a proxy for

the onset of an adverse effect and Survivors is the typical number of patients that do not develop the adverse effect.

This means that the total run time of our method is at most quadratic in the number of patients (Events and Survivors are both smaller than the total number of patients) and linear in the number of drugs of interest (in practice, 10 to 500). Our GPU implementation does not cut this theoretical time complexity since we still perform the exact same computations, without any approximation. However, they dramatically "improve the constant" by three orders of magnitude. Modern GPUs are made up of thousands of cores.

In practice, the number of patients in our datasets range from 100k to 100M - with an understanding that we always extract subsets that correspond to prescription subgroups. As a consequence, the number of operations required to train our model ranges from about 100 * 100 * 100k * 100k = 10^14 to at most 100 * 100 * 100M * 100M = 10^20 for a hypothetical ubiquitous drug on a very large population. Modern GPUs have a throughput that exceeds 10^12 operations per second: we expect that the run times for our method will range from a few seconds on a single GPU (for small datasets) to a few days on a GPU cluster (for nation-wide studies on very common drugs).

3. Results

In a preliminary work on hydroxychloroquine, the calculation time required for all the prescriptions of 2010 patients over 11 years is 5 days (parallelized calculation on a 2.6 GHz Intel Core i7 6-core processor). To do the same work on 400 ATC classes would take us more than 5 years. This method, as it stands, cannot be used, in real time and on a routine basis, for pharmacovigilance signal detection.

Programming the WCE method by distributing the heaviest portions (notably spline calculation) on the GPU makes it possible to accelerate the time required for this method by 1000 times using a computer graphics card and up to 10,000 times with a GPU server. The calculation to identify the spectrum of side effects associated with a drug on a cohort of about 2000 patients went from 5 days on CPU to 8 minutes (classic PC) or less than 48 seconds (GPU server).

The data driven WCE enabled the identification of eight ATC classes associated with the prescription of HCQ. The most relevant drugs were hydrocortisone, alendronic acid cholecalciferol, valsartan and chlormadinone.

The experimental dataset composed of five tables: patient table with 2010 rows, 3 columns (id_patient, age, sex), size 775KB ; hydroxychloroquine prescription: 50,425 rows, 3 columns (id_patient, date, ATC_classes), size 1.4MB ; all drug prescriptions: 1,175.507 rows, 3 columns (id_patient, date, ATC_classes), size 36.2MB ; all medical procedures: 190,712 rows, 3 columns (id_patient, date, Code_procedures), size 4.8MB ; all medical diagnoses: 48,470 rows, 3 columns (id_patient, date, Code_CIM10), size 1.5MB.

4. Discussion

This paper proposed an implementation of the WCE method using GPU programming allowing to considerably accelerate it. This makes it possible to use this method on all the drugs on the market to study their spectrum of adverse effects, to highlight new

vigilance signals and thus to have a global vigilance tool on medico-administrative database. This is a proof of concept for the use of this technology in epidemiology.

Drug consumption data is remarkably compact and fit without any problem in the main "RAM" memory of our GPUs, which have ~24Gb of usable space. As a consequence, disk latency is much less of a problem than for e.g. medical imaging studies. Going forward, we intend to scale up to full nation-wide studies (~70M patients for the full French population, at most ~500M for the US or the European Union), with full files that may weigh up to 100Gb. These will fit without problem in the RAM of our computers (124Gb) and may be cut into smaller pieces (~1Gb each) to be loaded sequentially on the GPU.

Beyond scaling issues alone, we will seek to extend the WCE model to meet the needs of our pharmaco-vigilance analysis. In particular, we will focus on:

- Patient age management. We will seek to compare standard approaches to Cox/WCE models (which treat age as an additional covariate) with an interval censoring method (already supported by our Python prototype on the GPU).
- Handling more advanced assumptions about the shape of risk functions. Beyond B-Splines (piecewise polynomial), we will evaluate the influence of positivity or convexity assumptions on the relevance of the estimated risk functions.
- Simultaneous management of several drugs. By exploiting the structure of the ATC hierarchical classification, we will seek to go beyond a "parallel" treatment of the different drugs chosen.

References

[1] Makary MA, Daniel M. Medical error—the third leading cause of death in the US. BMJ. 2016 May 3;353:i2139.

[2] Alatawi YM, Hansen RA. Empirical estimation of under-reporting in the U.S. Food and Drug Administration Adverse Event Reporting System (FAERS). Expert Opin Drug Saf. 2017 Jul;16(7):761–7.

[3] Schneeweiss S, Avorn J. A review of uses of health care utilization databases for epidemiologic research on therapeutics. J Clin Epidemiol. 2005 Apr;58(4):323–37.

[4] Weill A, Nguyen P, Labidi M, Cadier B, Passeri T, Duranteau L, et al. Use of high dose cyproterone acetate and risk of intracranial meningioma in women: cohort study. BMJ. 2021 Feb 3;372:n37.

[5] Platt R, Carnahan RM, Brown JS, Chrischilles E, Curtis LH, Hennessy S, et al. The U.S. Food and Drug Administration's Mini-Sentinel program: status and direction. Pharmacoepidemiol Drug Saf. 2012

[6] Stang PE, Ryan PB, Racoosin JA, Overhage JM, Hartzema AG, Reich C, et al. Advancing the science for active surveillance: rationale and design for the Observational Medical Outcomes Partnership. Ann Intern Med. 2010 Nov 2;153(9):600–6.

[7] Trifiro G, Fourrier-Reglat A, Sturkenboom MCJM, Díaz Acedo C, Van Der Lei J, EU-ADR Group. The EU-ADR project: preliminary results and perspective. Stud Health Technol Inform. 2009;148:43–9.

[8] Wisniewski AFZ, Bate A, Bousquet C, Brueckner A, Candore G, Juhlin K, et al. Good Signal Detection Practices: Evidence from IMI PROTECT. Drug Saf. 2016;39(6):469–90.

[9] AsPEN collaborators, Andersen M, Bergman U, Choi N-K, Gerhard T, Huang C, et al. The Asian Pharmacoepidemiology Network (AsPEN): promoting multi-national collaboration for pharmacoepidemiologic research in Asia. Pharmacoepidemiol Drug Saf. 2013 Jul;22(7):700–4.

[10] Sylvestre M-P, Beauchamp M-E, Abrahamowicz M. WCE package: weighted cumulative exposure models. :10.

[11] A data-driven pipeline to extract potential side effects through co-prescription analysis: application to a cohort study of 2,010 patients taking hydroxychloroquine with an 11-year follow-up. Available from: https://www.medrxiv.org/content/10.1101/2021.04.14.21255406v1

[12] Feydy J, Glaunès J, Charlier B, Bronstein M. Fast geometric learning with symbolic matrices, NeurIPS 2020.