Contents lists available at ScienceDirect

Artificial Intelligence In Medicine

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

Causality assessment of adverse drug reaction reports using an expert-defined Bayesian network

Pedro Pereira Rodrigues[a,b,*], Daniela Ferreira-Santos[a,b], Ana Silva[a,b,c], Jorge Polónia[a,c], Inês Ribeiro-Vaz[a,c]

[a] CINTESIS – Centre for Health Technology and Services Research, Rua Dr. Plácido Costa, 4200-450 Porto, Portugal
[b] MEDCIDS-FMUP, Faculty of Medicine of the University of Porto, Alameda Prof. Hernâni Monteiro, 4200-319 Porto, Portugal
[c] UFN – Northern Pharmacovigilance Centre (INFARMED), Rua Dr. Plácido Costa, 4200-450 Porto, Portugal

ARTICLE INFO

Keywords:
Adverse drug reactions
Causality assessment
Bayesian networks

ABSTRACT

In pharmacovigilance, reported cases are considered suspected adverse drug reactions (ADR). Health authorities have thus adopted structured causality assessment methods, allowing the evaluation of the likelihood that a drug was the causal agent of an adverse reaction. The aim of this work was to develop and validate a new causality assessment support system used in a regional pharmacovigilance centre. A Bayesian network was developed, for which the structure was defined by experts while the parameters were learnt from 593 completely filled ADR reports evaluated by the Portuguese Northern Pharmacovigilance Centre medical expert between 2000 and 2012. Precision, recall and time to causality assessment (TTA) was evaluated, according to the WHO causality assessment guidelines, in a retrospective cohort of 466 reports (April–September 2014) and a prospective cohort of 1041 reports (January–December 2015). Additionally, a simplified assessment matrix was derived from the model, enabling its preliminary direct use by notifiers. Results show that the network was able to easily identify the higher levels of causality (recall above 80%), although struggling to assess reports with a lower level of causality. Nonetheless, the median (Q1:Q3) TTA was 4 (2:8) days using the network and 8 (5:14) days using global introspection, meaning the network allowed a faster time to assessment, which has a procedural deadline of 30 days, improving daily activities in the centre. The matrix expressed similar validity, allowing an immediate feedback to the notifiers, which may result in better future engagement of patients and health professionals in the pharmacovigilance system.

1. Introduction

In pharmacovigilance, most of the reported cases are considered as suspected adverse drug reactions (ADR). Health professionals and consumers are asked to report episodes they believe are related with drug intake, but in most of the cases ADR are not particular for each drug and a drug rechallenge (i.e. the suspected drug was reintroduced into the patient's therapy, or the patient has taken the same suspected drug before) rarely occurs. To solve this difficulty, health authorities have adopted structured and harmonized causality assessment methods, in order to classify the ADR reports with one of the causality degrees proposed by the *Uppsala Monitoring Centre (WHO-UMC)* causality assessment system [1]. Apart from ADR identification, where innovative methods have been proposed [2], causality assessment is an essential tool in the pharmacovigilance system, as it helps the risk-benefit evaluation of commercialized medicines, and is part of the signal detection (being a signal a “reported information on a possible causal relationship between an adverse event and a drug, the relationship being unknown or incompletely documented previously” [1]) performed by health authorities.

The Portuguese Pharmacovigilance System has adopted the method of Global Introspection [3], since its creation. During this process, an expert (or a group of experts) expresses judgement about possible drug causation, considering all available data in the ADR report. The decision is based on the expert knowledge and experience, and uses no standardized tools. Although this is the method most widely used [4], it has some limitations related to its reproducibility and validity [5–7]. Besides, this method is closely linked with the medical expert availability which not always allows meeting legal deadlines.

Causality assessment can also be done through validated algorithms such as the Naranjo [8], Jones [9] or Karch-Lasagna [10] algorithms. Although these algorithms have better agreement rates than Global

[*] Corresponding author at: CINTESIS – Centre for Health Technology and Services Research, Rua Dr. Plácido Costa, 4200-450 Porto, Portugal.
E-mail address: pprodrigues@med.up.pt (P.P. Rodrigues).

https://doi.org/10.1016/j.artmed.2018.07.005
Received 8 November 2017; Received in revised form 10 July 2018; Accepted 27 July 2018

Introspection, they also have the disadvantage of not being flexible and, consequently, it is not possible to include more causal factors to be evaluated at the same time [4]. Moreover, in our experience, some real cases evaluated by more than one algorithm may give rise to different degrees of causality. Guidelines such as the ones used for causality assessment are several times hard to interpret and to apply, even by experienced practicioners. Furthermore, they often result in simple rules or association measures, making their application in decision support somewhat limited, especially in the context of guidelines that are to be computer-interpreted for decision support systems [11].

The definition of decision support systems (most of the times based on expert systems) is currently a major topic since it may help the diagnosis, treatment selection, prognosis of rate of mortality, prognosis of quality of life, etc. They can even be used to help on administrative tasks like the one addressed by this work. However, the complicated nature of real-world biomedical data has made it necessary to look beyond traditional biostatistics [12] without loosing the necessary formality. Hence, such systems could be implemented applying methods of machine learning [13], since new computational techniques are better at detecting patterns hidden in biomedical data, and can better represent and manipulate uncertainties [14]. In fact, the application of data mining techniques to medical knowledge discovery tasks is now a growing research area. These techniques vary widely and are based on data-driven conceptualisations, model-based definitions or on a combination of data-based knowledge with human-expert knowledge [12]. Bayesian approaches have an extreme importance in these problems as they provide a quantitative perspective and have been successfully applied in health care domains [15]. One of their strengths is that Bayesian statistical methods allow taking into account prior knowledge when analysing data, turning the data analysis into a process of updating that prior knowledge with biomedical and health-care evidence [12]. However, only after the 90's we may find evidence of a large interest on these methods, namely on Bayesian networks.

Bayesian networks can be seen as an alternative to logistic regression, where statistical dependence and independence are not hidden in approximating weights, but rather explicitly represented by links in a network of variables [15], offering a general and versatile approach to capturing and reasoning with uncertainty in medicine and health care. Moreover, they intrinsically include an evidence synthesis mechanism that is yet to be fully exploited as meta-analysis method, central piece for evidence-based guidelines [16]. Generally, a Bayesian network represents a joint distribution of one set of variables, specifying the assumption of independence between them, with the interdependence between variables being represented by a directed acyclic graph. Each variable is represented by a node in the graph, and is dependent on the set of variables represented by its ascendant nodes. This dependence is represented by a conditional probability table that describes the probability distribution of each variable, given their ascendant variables [17].

Given their successful applications in previous healthcare applications [18,19], we decided to build (and validate) a Bayesian network model to help in the process of causality assessment carried out in pharmacovigilance centres, making possible a new path for implementing such practice guidelines. Fig. 1 sketches the workflow of causality assessment process at the pharmacovigilance centre, highlighting the inclusion of our proposal to speed up the feedback to original reporters. The Bayesian network is not used by the expert. The pharmacovigilance team uses the Bayesian network to have a prediction of the causality that the expert will assign to the report. With this, they can start to prepare the global report for national and European institutions, while providing a quick response to the reporter, with a preliminary assessment degree. This not only speeds up the final report preparation (which is where most of the time is spent by the team) but also improves the feedback to the reporter, increasing engagement in the pharmacovigilance system. Then, the expert performs the usual global introspection and gives the final degree which is sent back to the reporter as well. The Bayesian network has, therefore, two main objectives: (a) predict the most probable causality degree the expert will assign to the report, and (b) be interpretable by the pharmacist team members to inspect which variables are causing the shifts in causality classification.

Although the entire process is called “causality assessment”, we do not intend to model causal relationships. In fact, the term “causality assessment” does not come from the methodological approach of causal inference or causality modelling; rather, it comes directly from the pharmacovigilance system, describing the process where an algorithm or an expert physician tries to classify whether a drug (or a combination of drugs) was the causal agent of the adverse event that was reported. In this work, we modelled associations by means of conditional probabilities structured in a Bayesian network, trying to predict the classification the expert would provide for the same data.

2. Material and methods

The study is framed as the development of a diagnostic test, where the comparison (gold standard) is the method of medical expert's global introspection.

2.1. Cohorts of adverse drug reaction reports

Three cohorts of suspected ADR were used in the building and assessment of the Bayesian model: a derivation cohort, consisting of the registries of suspected ADR evaluated by a medical expert in a regional pharmacovigilance centre between 2000 and 2012; a retrospective validation cohort consisting of all reports of suspected ADR received in the same centre within the initial 6 months of implantation of the system in the centre (in 2014); and a prospective cohort, consisting of all reports of suspected ADR received in the same centre during the year of 2015. Additionally, a cohort of reports from the first semester of 2016 was used to validate the derived simplified assessment matrix (explained further in the text).

2.2. Relevant variables

Each suspected ADR was evaluated by the medical expert of the Northern Pharmacovigilance Centre using global introspection, for causality categories of *Definite*, *Probable*, *Possible* or *Conditional*, according to the WHO causality assessment guidelines. The variables used to develop the network were the usual data needed for common causality assessment algorithms [8–10] and for global instrospection, as explained below:DescribedIf the ADR was previously reported in other patients so that this event is descibed in the summary of product characteristics (SPC), it enhances the likelihood that a drug is the cause of the observed event; this variable was slightly enhanced for the prospective cohort, including descriptions also on other sources of published literature. Variable *Described* can take (yes/no) values.Reintroduced / ReappearedData on drug rechallenge is mostly absent, because it is not likely that a patient who has suffered ADR receive the suspected drug again. This data is available when the patient uses the drug for the second time by mistake, or when the first ADR episode has not been interpreted as such. When this data is available, it is a very useful variable on causality assessment, because it provides a confirmation of the previous suspicion. Since uncertainty exist during adverse event reporting, two separate variables are modeled: *Reintroduced* (yes/no) and *Reappeared* (yes/no/notapplicable).Suspended / ImprovedA favorable evolution of the ADR after drug withdrawal increases the likelihood that the suspected drug was the cause of the ADR. Since uncertainty exist during adverse event reporting, two separate variables are modeled: *Suspended* (yes/no) and *Improved* (yes/no/notapplicable).ConcomitantThe presence of other drugs can represent alternative causes (other than the suspected drug) that could on their own cause the ADR. Variable modeled with (yes/no) values.Suspected

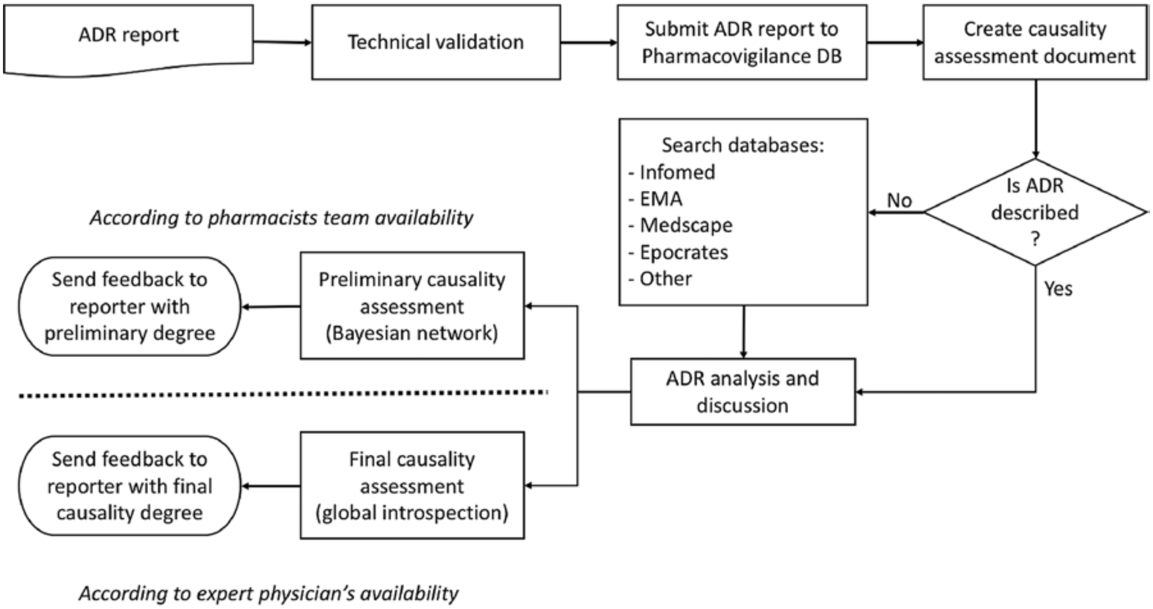

ADR report
Technical validation
Submit ADR report to Pharmacovigilance DB
Create causality assessment document
Is ADR described ?
No
Yes
Search databases:
- Infomed
- EMA
- Medscape
- Epocrates
- Other
ADR analysis and discussion
According to pharmacists team availability
Send feedback to reporter with preliminary degree
Preliminary causality assessment (Bayesian network)
Send feedback to reporter with final causality degree
Final causality assessment (global introspection)
According to expert physician's availability

Fig. 1. Causality assessment workflow at a regional pharmacovigilance centre, including administrative steps, causality assessment by two different methods, and feedback to the original creator of the adverse drug report (ADR).

interactionIf there is suspicion that an interaction with other drug existed, then the cause of the ADR is less clear. Variable modeled with (yes/no) values.Route of administrationSome ADR are more likely to occur when the drug is administered intravenously, and others when it is administered orally or topically, for example. Variable modeled with (oral/topic/injectable) values.NotifierThis variable represents a proxy for quality of information. In most cases, physicians report their ADR suspicions more completely and precisely than other health professionals. Variable modeled with (physician/pharmacist/nurse/other) values.Pharmacotherapeutic groupAlthough ADR are not specific to drug classes, some events are more related to certain pharmacotherapeutic group(s). The nomenclature used for drug classification was the one adopted by Portuguese Authority of Medicines and Health Products (INFARMED, IP) according to national legislation *(Despacho n^{o} 21844/2004, de 12 de outubro)* which includes a correspondence with the international ATC. Variable modeled with (anti-infectious/central nervous system/cardiovascular system/blood/respiratory system/gastrointestinal system/genitourinary system/hormones and drugs used to treat endocrine diseases/loco-motor system/anti-allergic medication/nutrition/corrective agents in blood volume and electrolyte disturbances/ drugs for skin disorders/drugs used in otorhinolaryngological disorders/drugs for eye disorders/antineoplastic drugs and immune-modulators/drugs used to treat poisoning/vaccines and immunoglobulins/diagnosis media) values.IneffectivenessDrug ineffectiveness means the drug did not act as expected (example: an analgesic that does not relieve a headache). This is an inherent condition to any drug, because it is accepted that none drug is 100% effective. Until February 2015, the expert interpretation of this kind of ADR (lack of effect) assumed that they should all receive the causality degree of *Conditional*, as medical experts needed further information to a complete the assessment. After February 2015, this interpretation has changed into assuming lack of effect is described for all drugs. Variable modeled with (yes/no) values.

The decision to include variables in the Bayesian network was expert-oriented, following international guidelines for "causality assessment" in pharmacovigilance. Some of those are not queried in the original report, as they require expert-knowledge from the pharmacovigilance pharmacist team (e.g. drug's pharmaceutical group). However, none requires the intervention of the expert (medical doctor) who proceeds with the assessment a posteriori.

2.3. Data pre-processing

Data was collected from the official adverse drug report (ADR) database, including completely filled ADR reports evaluated by the Portuguese Northern Pharmacovigilance Centre. In case of duplicate reports, an evaluation was performed and the duplicate dismissed. All the reports inserted in the database were cross-checked with the original paper reports:

- in the cases were we had two or more drugs that had different properties and characteristics in the same report, the classification on the database was unknown for variables *Suspended*, *Administration* and *PharmaGroup*;
- all reports where there was no indication of the reaction being described in the literature were classified as not being described;
- also, if the same report described two levels of causality, we considered it having the lowest one, following the recommendations of the experts;
- variables *Suspended* and *ImprovedAfterSuspension* (and likewise *Reintroduced* and *Reappeared*) have an intimate connection, with the former imposing a "NotApplicable" status to the latter in cases where the former were negative; if we do not know what happened, then both variables are left missing.

2.4. Bayesian network model definition

A Bayesian network was developed where the structure of causal dependence was defined towards implementing the current guidelines for causality assessment, in cooperation with the medical expert, whereas the conditional probabilities were induced from the derivation cohort.

Fig. 2 presents a graphical representation of the development process. ADR reports are sent to pharmacovigilance centres using any available channel (e.g. web service, online form[1], phone, email, etc.) where a medical expert assesses the causality by the process of global introspection, assigning a causality degree (*Definite*, *Probable*, *Possible* or *Conditional*) to the report, which is then stored in a relational database.

[1] http://newdbserver.med.up.pt/web.care/UFN/notificacao/notificacao.php

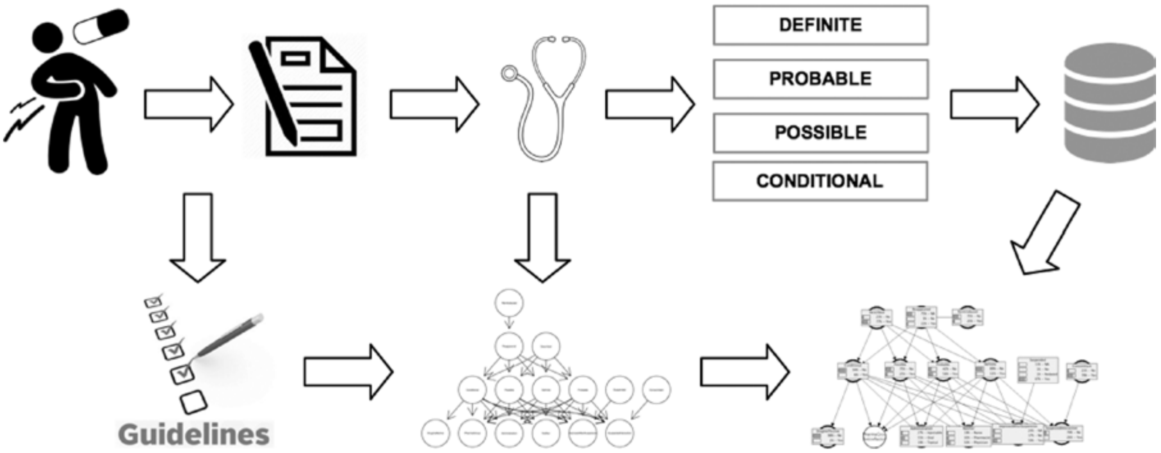


Fig. 2. Definition of the expert-informed Bayesian network for causality assessment of adverse drug reaction reports. Both the guidelines and the medical expert opinion were used to define the structures, while parameters were learnt from historical assessment.

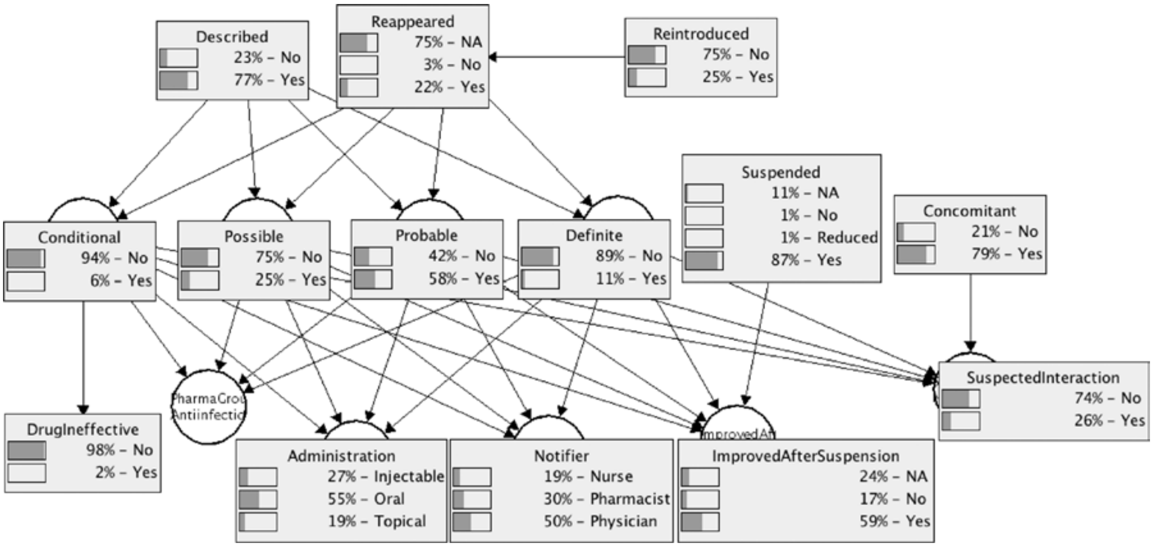


Fig. 3. Expert-informed Bayesian network for causality assessment of adverse drug reaction (ADR) reports. Monitors show the marginal probabilities for all nodes, except for drug pharmacotherapeutical group, which was hidden for presentation space reasons.

Using published guidelines for the causality assessment, and the experienced opinion of the medical expert, a Bayesian network was built trying to capture the causal interdependences of the ADR. Then, using the historical assessment done by the medical expert, the network parameters were learnt from data of the derivation cohort, defining its quantitative model.

2.5. Bayesian network model structure

Following the guidelines and expert opinion, we separated the relevant variables into four groups of nodes for the network: (a) factors that generally influence the occurrence of ADR for the drug in question (i.e. *Described*, *Reintroduced* and *Reappeared*); (b) factors which are related with the particular report in question (i.e. *Suspended*, *ImprovedAfterSupension*, *Concomitant*, *SuspectedInteraction*, *Notifier* and *Administration*); (c) special cases, i.e. *PharmaGroup* which, given the number of possible states, would make the network too complex if modeled in causal ways, and also, *Ineffectiveness* was considered to only influence the *Conditional* degree; and (d) given that we have observed the expert using the same information in different ways for different causality degrees, and the limited number of reports for the modeling (which hinder the possibility of bigger conditional probability tables), we decided to model each causality degree in a separate node, with the final degree being decided by the team member, assigning the degree with higher a posteriori probability. Fig. 3 presents the final structure of the model.

2.6. Bayesian network model parameters

Since we are trying to model the assessment done by the expert, the model's parameters were learnt from the actual reports and assessment in the derivation cohort. Given the limited quality of the electronic reporting of suspected ADR in years prior to 2012, we only considered complete reports for this step. An exception is noted for node *Ineffectiveness*, as this info was not registered up until 2014; thus, to model the uncertainty described by the expert, for this node, the conditional probability table was defined as

$$P(Yes|Conditional) = 1.0 \text{ and } P(Yes| \sim Conditional) = .375$$

Fig. 3 presents the marginal probabilities for each node after the conditional probability table fitting procedure.

2.7. Preliminary causality assessment matrix

The preliminary application of the model by the notifier can be approached by an appropriately defined matrix. In order to choose

Table 1
Descriptive analysis of the three cohorts used in the derivation and validation of the model.

	Derivation *n* (%)	Validation *n* (%)	Prospective *n* (%)	Total *n* (% [95%CI])
Period	**2000–2012**	**April–September 2014**	**January–December 2015**	
Described	**593 (28.3)**	**464 (22.1)**	**1041 (49.6)**	**2098 (100)**
Yes	459 (77.4)	383 (82.5)	932 (89.5)	1774 (84.6 [82.9,86.1])
No	134 (22.6)	81 (17.5)	109 (10.5)	324 (15.4 [13.9,17.1])
Reintroduced	**593 (55.6)**	**189 (17.7)**	**284 (26.6)**	**1066 (100)**
Yes	148 (25)	59 (31.2)	101 (35.6)	308 (28.9 [26.2,31.7])
No	445 (75)	48 (25.4)	183 (64.4)	676 (63.4 [60.4,66.3])
Not applicable	0 (0)	82 (43.4)	0 (0)	82 (7.7 [6.2,9.5])
Reappeared after reintroduction	**593 (56.6)**	**183 (17.5)**	**271 (25.9)**	**1047 (100)**
Yes	129 (21.8)	44 (24)	56 (20.7)	229 (21.9 [19.4,24.5])
No	18 (3)	9 (4.9)	32 (11.8)	59 (5.6 [4.4,7.3])
Not applicable	446 (75.2)	130 (71)	183 (67.5)	759 (72.5 [69.7,75.2])
Suspended	**593 (30.4)**	**418 (21.5)**	**937 (48.1)**	**1948 (100)**
Yes	518 (87.4)	319 (76.3)	628 (67)	1465 (75.2 [73.2,77.1])
Reduced	3 (0.5)	5 (1.2)	24 (2.6)	32 (1.6 [1.1,2.3])
No	8 (1.3)	28 (6.7)	152 (16.2)	188 (9.7 [8.4,11.1])
Not applicable	64 (10.8)	66 (15.8)	133 (14.2)	263 (13.5 [12,15.1])
Improved after suspension	**593 (31.7)**	**399 (21.3)**	**879 (47)**	**1871 (100)**
Yes	486 (82)	293 (73.4)	567 (64.5)	1346 (71.9 [69.8,74])
No	29 (4.9)	7 (1.8)	27 (3.1)	63 (3.4 [2.6,4.3])
Not applicable	78 (13.2)	99 (24.8)	285 (32.4)	462 (24.7 [22.8,26.7])
Concomitant medication	**593 (47.7)**	**189 (15.2)**	**460 (37)**	**1242 (100)**
Yes	466 (78.6)	180 (95.2)	445 (96.7)	1091 (87.8 [85.9,89.6])
No	127 (21.4)	9 (4.8)	15 (3.3)	151 (12.2 [10.4,14.1])
Suspected interaction	**593 (34.7)**	**75 (4.4)**	**1041 (60.9)**	**1709 (100)**
Yes	37 (6.2)	10 (13.3)	21 (2)	68 (4 [3.1,5])
No	556 (93.8)	65 (86.7)	1020 (98)	1641 (96 [95,96.9])
Route of administration	**593 (29)**	**443 (21.7)**	**1006 (49.3)**	**2042 (100)**
Oral	429 (72.3)	267 (60.3)	654 (65)	1350 (66.1 [64,68.2])
Injectable	123 (20.7)	167 (37.7)	346 (34.4)	636 (31.1 [29.2,33.2])
Topical	41 (6.9)	9 (2)	6 (0.6)	56 (2.7 [2.1,3.6])
Notifier	**593 (29.3)**	**466 (23)**	**966 (47.7)**	**2025 (100)**
Physician	372 (62.7)	295 (63.3)	565 (58.5)	1232 (60.8 [58.7,63])
Pharmacist	175 (29.5)	91 (19.5)	283 (29.3)	549 (27.1 [25.2,29.1])
Nurse	46 (7.8)	57 (12.2)	118 (12.2)	221 (10.9 [9.6,12.4])
Other	0 (0)	23 (4.9)	0 (0)	23 (1.1 [0.7,1.7])
Pharmacotherapeutical group	**593 (28.8)**	**466 (22.6)**	**1003 (48.6)**	**2062 (100)**
AntiallergicMedication	11 (1.9)	4 (0.9)	8 (0.8)	23 (1.1 [0.7,1.7])
Antiinfectious	136 (22.9)	103 (22.1)	264 (26.3)	503 (24.4 [22.6,26.3])
AntineoplasticDrugsImmunemodulators	35 (5.9)	82 (17.6)	212 (21.1)	329 (16 [14.4,17.6])
Blood	7 (1.2)	9 (1.9)	25 (2.5)	41 (2 [1.4,2.7])
CardiovascularSystem	72 (12.1)	29 (6.2)	42 (4.2)	143 (6.9 [5.9,8.1])
CentralNervousSystem	91 (15.3)	51 (10.9)	162 (16.2)	304 (14.7 [13.3,16.4])
DiagnosisMedia	1 (0.2)	9 (1.9)	20 (2)	30 (1.5 [1,2.1])
DrugsForEyeDisorders	4 (0.7)	1 (0.2)	8 (0.8)	13 (0.6 [0.4,1.1])
DrugsForSkinDisorders	23 (3.9)	3 (0.6)	3 (0.3)	29 (1.4 [1,2])
DrugsToTreatPoisoning	1 (0.2)	0 (0)	0 (0)	1 (0 [0,0.3])
GastrointestinalSystem	28 (4.7)	11 (2.4)	16 (1.6)	55 (2.7 [2,3.5])
GenitourinarySystem	13 (2.2)	1 (0.2)	10 (1)	24 (1.2 [0.8,1.8])
Hormones	17 (2.9)	14 (3)	33 (3.3)	64 (3.1 [2.4,4])
LocomotorSystem	101 (17)	77 (16.5)	84 (8.4)	262 (12.7 [11.3,14.2])
Nutrition	3 (0.5)	2 (0.4)	2 (0.2)	7 (0.3 [0.1,0.7])
Otorhinolaryngology	0 (0)	2 (0.4)	0 (0)	2 (0.1 [0,0.4])
RespiratorySystem	10 (1.7)	8 (1.7)	12 (1.2)	30 (1.5 [1,2.1])
VaccinesImmunoglobulins	40 (6.7)	58 (12.4)	102 (10.2)	200 (9.7 [8.5,11.1])
Volaemia	0 (0)	2 (0.4)	0 (0)	2 (0.1 [0,0.4])

which variables should be included in the matrix, we applied logistic regression with all independent variables using the enter method. Variables with statistical significance, which were available to the notifier (and for which enough data was available), were chosen as factors for the matrix. Each cell of the matrix represents the marginal posterior degree probability estimate for that subgroup of patients. The values in each cell of the matrix represent the expected degree for a report in that subgroup.

To assess the discriminative ability of the matrix, specific cut-off values were chosen after performing a ROC analysis of the derivation cohort. Given the expected application of the matrix, we only label the report as *Definite* if probability of that causality degree is higher than that of *Probable*, and we label them as *Possible* in all cases where the probability of that causality degree is higher than 40%; all other cases are labeled *Probable*, in the matrix.

Table 2
Causality assessment by the medical expert (gold standard using global introspection) and the Bayesian network model.

	Derivation *n* (%)	Validation *n* (%)	Prospective *n* (%)	Total *n* (% [95%CI])
Expert assessment	**593 (28.2)**	**466 (22.2)**	**1041 (49.6)**	**2100 (100)**
Definite	60 (10.1)	37 (7.9)	36 (3.5)	133 (6.3 [5.3,7.5])
Probable	346 (58.3)	372 (79.8)	833 (80)	1551 (73.9 [71.9,75.7])
Possible	152 (25.6)	44 (9.4)	131 (12.6)	327 (15.6 [14.1,17.2])
Conditional	35 (5.9)	13 (2.8)	41 (3.9)	89 (4.2 [3.4,5.2])
Bayesian net assessment	**593 (28.2)**	**466 (22.2)**	**1041 (49.6)**	**2100 (100)**
Definite	77 (13)	36 (7.7)	47 (4.5)	160 (7.6 [6.5,8.9])
Probable	331 (55.8)	388 (83.3)	945 (90.8)	1664 (79.2 [77.4,80.9])
Possible	185 (31.2)	38 (8.2)	47 (4.5)	270 (12.9 [11.5,14.4])
Conditional	0 (0)	4 (0.9)	2 (0.2)	6 (0.3 [0.1,0.7])

Table 3
Validity assessment for the 2015 prospective cohort of adverse drug reaction (ADR) reports. Columns represent the assessment done by the expert using global introspection, while lines represent the Bayesian network most probable causality degree.

	Definite	Probable	Possible	Conditional	Precision % [95%CI]	Recall % [95%CI]	Node AUC % [95%CI]
Definite	30	16	1	0	63.8 [48.5,76.9]	83.3 [66.5,93.0]	91.7 [84.8,98.5]
Probable	4	792	117	32	83.8 [81.3,86.1]	95.1 [93.3,96.4]	70.7 [66.6,74.8]
Possible	2	24	12	9	25.5 [14.4,40.6]	9.2 [5.0,15.8]	66.7 [62.0,71.3]
Conditional	0	1	1	0	0.0 [0.0,80.2]	0.0 [0.0,10.7]	69.1 [61.3,76.9]

2.8. Evaluation strategy and software used

The network assessment was compared with the gold standard (medical expert's global introspection) in terms of sensitivity (recall) and positive predictive values (precision). Also evaluated (using the retrospective validation cohort) was the time to causality assessment (TTA), compared to the manual assessment times recorded in the centre's quality management system. Final validation was done in the prospective validation cohort, along with the specific AUC for each outcome node. Bayesian network structure was defined with *SamIam* [20], while conditional probability tables were learnt from data using R package *bnlearn* [21]. Inference for daily use was done using *SamIam*, while validation was done with the R package *gRain* [22] using Lauritzen-Spiegelhalter algorithm [23] for exact posterior probability inference. ROC curves were computed with R package *pROC* [24], Precision-Recall curves with R package *PRROC* [25], and confidence intervals for proportions were computed with R package *stats* [26].

3. Results

The path for model validation in this work was defined in several steps (derivation, validation, prospective validation and simplified assessment), leading to different levels of results.

3.1. Descriptive analysis

The 2000 to 2012 activity generated 3220 records, from which 593 complete instances were used as derivation cohort. The retrospective validation cohort, collected during 6 months in 2014, included 466 reports. The final prospective validation cohort, collected for the whole year of 2015, included 1041 reports.

Over all 2100 ADR, 85% were described, 29% did not include a drug rechallenge, but 77% were suspended leading to patient status improvement in 72% of the cases, 88% considered concomitant medication, although only 4% actually raised suspicion of interaction. The majority of ADR were reported by physicians (61%) and pharmacists (27%), being mainly related to oral (66%) or injectable (31%) drugs. Table 1 presents the descriptive analysis of the three cohorts, while Table 2 presents a summary of the causality assessment performed by both the medical expert (using global introspection) and the Bayesian network model.

The additional cohort used for the matrix validation included 482 reports: 93% described, 18% without drug rechallenge, 62% suspended leading to patient improvement in 86% of the cases, 44% with concomitant medication.

3.2. Model derivation and prospective validation

The initial analysis of the derivation and validation cohorts gave indications that the network seems better for higher degrees of causality (precision and recall for *Probable* above 87%). However, in the validation cohort, the network actually tends to overrate causality (96.9% of errors on *Possible* cases classified as *Probable*) or give the immediately below level (90.8% of errors on *Definite* cases classified as *Probable*; 69.7% of errors on *Probable* cases classified as *Possible*). The median (Q1:Q3) time to causality assessment was 4 (2:8) days using the network and 8 (5:14) days using global introspection, meaning the network allowed a faster time to assessment.

The prospective validation of the network reinforced the ability to identify higher levels of causality (recall for *Definite* and *Probable* above 80%) while exposing even stronger problems dealing with the lower levels of causality. Table 3 presents the results for that cohort, where the network clearly failed to address *Possible* and *Conditional* levels, although each node, alone, had a specific AUC above 65% (Figs. 4 and 5 present the ROC and Precision–Recall curves for all degrees, which give only insights of the discriminative ability of each node, since the final degree is decided by the team member a posteriori). Nonetheless, classifying *Possible/Probable* cases as *Definite* is the worst error as judge by the experts. Therefore, the fact that only a small fraction of *Probable* cases are classified as *Definite* appears as a good quality of the model.

3.3. Matrix for preliminary assessment

Although the aim of the model is to help the pharmacovigilance teams in the causality assessment process, we envisioned its use as a rough feedback to the notifier, in order to provide an initial expectation of causality for the submitted report. Table 4 presents a matrix using only two variables (available to the notifier) which provides the expected causality assessment for those subgroups of reports. Given the existing conceptual constraints, in case of missing information regarding reappearance after reintroduction and improvement after suspension, categories “Not reintroduced” or “Not suspended” should be assumed, respectively. Additionally, the matrix was validated against a new cohort of reports from the first semester of 2016 (from the same pharmacovigilance centre, which followed similar distributions) using the following rationale:Definiteif probability of *Definite* degree is higher

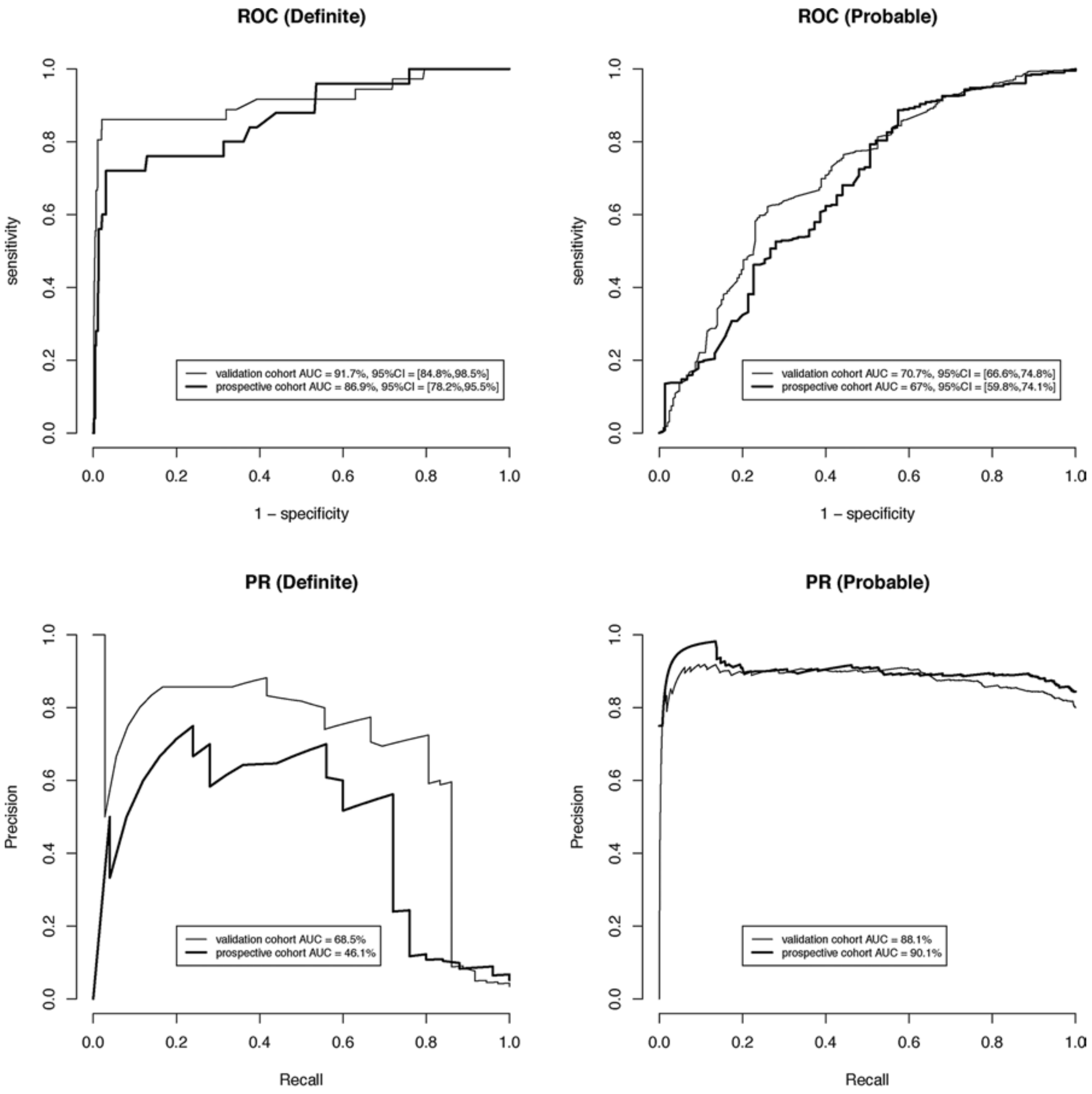


Fig. 4. ROC and Precision–Recall curves for the top causality degrees, reflecting the application of the model to both the validation and prospective cohorts.

than that of *Probable*;Possibleif probability of *Possible* degree is ≥ 40%;Probableotherwise. Results of applying the matrix to the prospective cohort are presented in Table 5 (top half), which are not significantly different from the prospective validation of the model itself (Table 3). Looking towards the application to the new 2016 cohort (Table 5, bottom half), although most validity measures are not significantly different, we should note a decrease in sensitivity in the *Definite* degree. Nonetheless, this decrease represents a shift towards a more conservative approach (classifying *Definite* cases as *Probable*) which is recommended assuming its future direct use by the notifier.

4. Discussion

The Bayesian network model allowed a faster time to causality assessment, which has a procedural deadline of 30 days, improving daily activities in the pharmacovigilance centre. Moreover, the model was accurate on most cases, showing satisfactory results to the higher degrees of causality.

4.1. On results

Although the overall problem is quite imbalanced, with *Probable* cases being the vast majority which could allow a majority default model resulting in 100% recall and 80% precision for the *Probable* degree (against 95% recall and 84% precision of our strategy), we believe the beneficial trade-off rises from being able to address the most important *Definite* degree (83% recall and 64% precision). Likewise, ROC and PR curves seem to indicate a somewhat low discriminative ability. However, since we have modelled each level as a separate node, the ROC and PR curves are drawn from the probability output of each node which does not translate directly into a global probability for that level. Therefore, ROC and PR curves are only indicative of each node's quality, not of the overall system.

On the other hand, it had a non-adequate behaviour with the two lowest degrees of causality. We believe the Bayesian network failed to learn the degree *Possible* because this degree is much related with the existence of concomitant diseases or conditions that could explain the ADR [1]. However, this kind of information is not collected in the ADR form as a structured field. The notifier may provide this information in a free text field (comments) or by phone. For this reason, the network does not consider any node with this question. On the other hand, the medical expert, while performing global introspection, is aware of this information (if any) and can fully assess the case. For example, there were several reports of headache and fatigue involving new drugs used

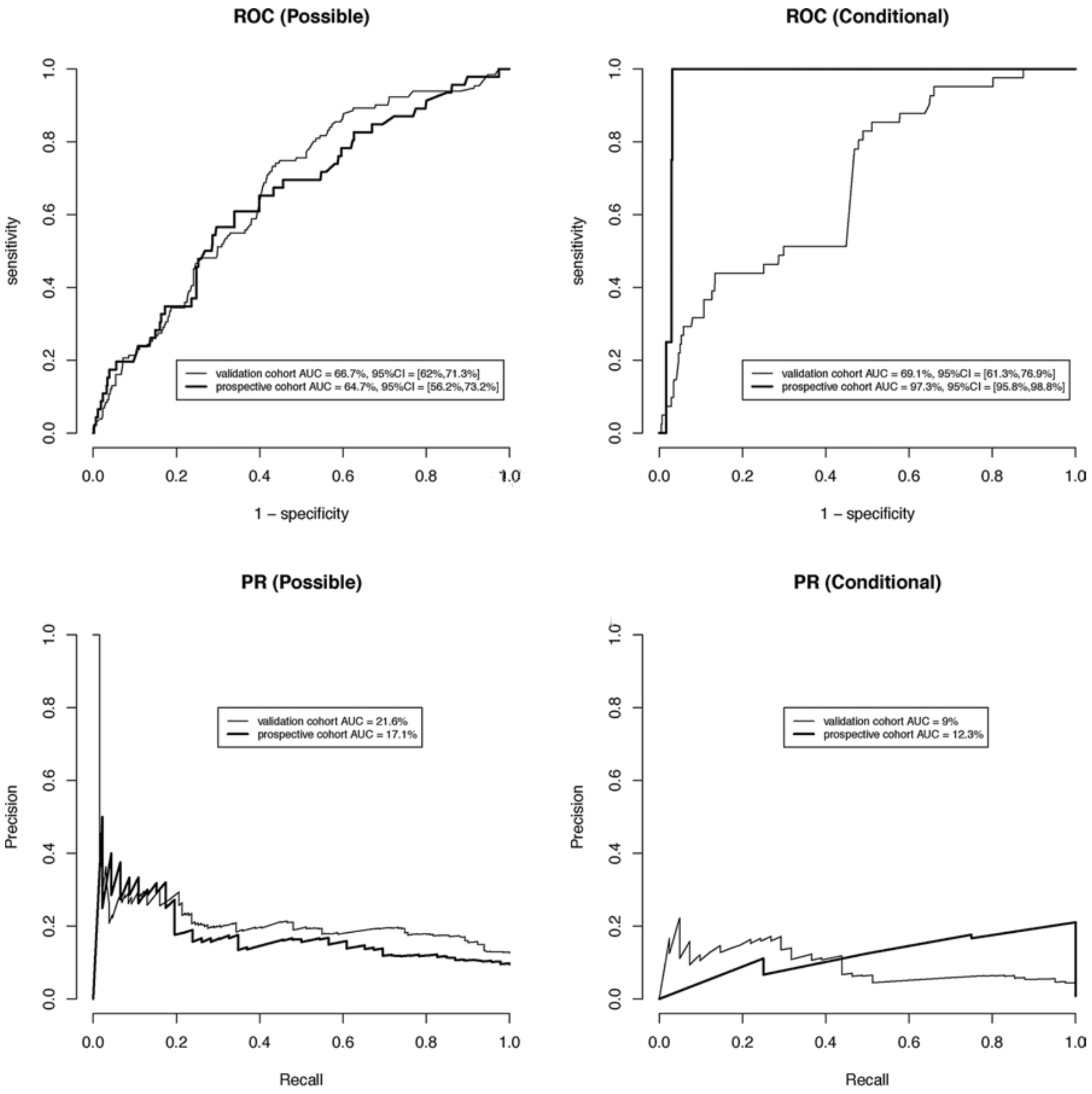


Fig. 5. ROC and Precision–Recall curves for the bottom causality degrees, reflecting the application of the model to both the validation and prospective cohorts.

in Hepatitis C. These reports were mainly assessed by the expert as *Possible*, because the ADR reported could also be explained by the disease (Hepatitis), contrary to the Bayesian network which assessed these cases as *Probable*, since these ADR are described in the summary of product characteristics. We believe that this issue can be solved with the inclusion of a new question on the ADR form on (a) the existence of any other eventual cause to the ADR other than the suspect drug, or (b) if the drug is considered to be new in the population; consequently, new nodes in the Bayesian network are needed.

The network also failed, and with greater magnitude, to learn the degree *Conditional*. This is a temporary degree, attributed to those cases with insufficient information and also to those cases which is expected to obtain more data. For this reason, it is a degree difficult to fit in a model and was hence not considered for the preliminary matrix assessment. After February 2015, this interpretation has changed (which prevented a joint analysis of both cohorts together). Thus, a particular analysis will be performed to node *Innefectiveness*.

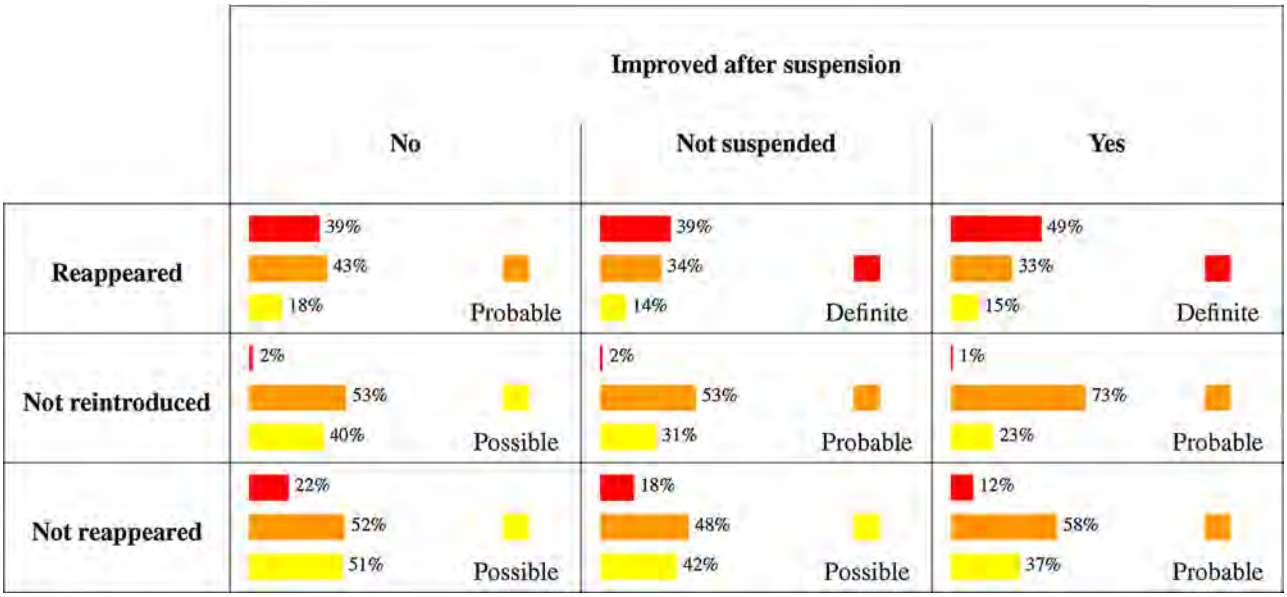
4.2. *On the final decision*

The use of a maximum a posteriori probability rule to define the final degree (or a cut-off rule as used in the preliminary assessment matrix) is clearly debatable. To better assess the quality of our strategy, we have compared it with a 2-layer solution, where the a posteriori probabilities of each degree are fed to a naive Bayes classifier, with factors being (quintile-based) discretized versions of each degree probability from the original network, and conditional probability tables learnt using the validation cohort (2015). This classifier was then applied to the prospective cohort (2016), and results are present in Table 6, where we can stress that the strategy is mostly equivalent, being only slightly more flexible towards *Conditional* degree.

4.3. *On model validation*

As noted in the results section, there are changes in the data priors (which were overviewed in the detailed descriptive analysis table) which might affect the validation of the model. As is, given the changes in data distribution, the validation can be considered a worst-case scenario. This has led us to the prospect of learning the models online, updating them with new data as available. However, to be validated to be used in real practice, an online model would need to be evaluated not only in its validity, but also in its adaptability, which would require a more complex study design. Although we are looking towards adaptive models in the future, we need to have a stable model validated for

Table 4
Preliminary assessment matrix showing the probability [%] of causality degrees, according to variables *Reappeared after reintroduction* and *Improved after suspension*, and the likely degree according to defined thresholds: *Definite*, if probability of degree is higher than that of *Probable*; *Possible* if probability of degree is ≥40%; *Probable* otherwise. In case of missing information about reappearance or about improvement after suspension, categories "Not reintroduced" and "Not suspended" should be assumed, respectively.



	Improved after suspension		
	No	Not suspended	Yes
Reappeared	🟥 39% 🟧 43% 🟨 18% 🟧 Probable	🟥 39% 🟧 34% 🟨 14% 🟥 Definite	🟥 49% 🟧 33% 🟨 15% 🟥 Definite
Not reintroduced	🟥 2% 🟧 53% 🟨 40% 🟨 Possible	🟥 2% 🟧 53% 🟨 31% 🟧 Probable	🟥 1% 🟧 73% 🟨 23% 🟧 Probable
Not reappeared	🟥 22% 🟧 52% 🟨 51% 🟨 Possible	🟥 18% 🟧 48% 🟨 42% 🟨 Possible	🟥 12% 🟧 58% 🟨 37% 🟧 Probable

Table 5
Matrix validity assessment for the 2015 and 2016 cohorts of adverse drug reaction (ADR) reports. Columns represent the assessment done by the expert using global introspection, while lines represent the causality degree according to the preliminary assessment matrix.

	Definite	Probable	Possible	Conditional	Precision % [95%CI]	Recall % [95%CI]
2015						
Definite	33	16	4	3	58.9 [45.0,71.6]	91.7 [76.4,97.8]
Probable	3	793	121	36	83.2 [80.7,85.5]	95.2 [93.5,96.5]
Possible	0	24	6	2	18.8 [7.9,37.0]	4.6 [1.9,10.1]
2016						
Definite	15	12	2	0	51.7 [32.9,70.1]	60.0 [38.9,78.2]
Probable	10	380	40	4	87.6 [84.0,90.4]	93.4 [90.4,95.5]
Possible	0	15	4	0	21.1 [7.0,46.1]	8.7 [2.8,21.7]

Table 6
Validity assessment of the 2-layer classifier strategy, for the 2016 cohorts of adverse drug reaction (ADR) reports. Columns represent the assessment done by the expert using global introspection, while lines represent the causality degree according to the naive Bayes classifier applied to the a posteriori probabilities computed by the original network.

	Definite	Probable	Possible	Conditional	Precision % [95%CI]	Recall % [95%CI]
2016						
Definite	14	8	1	0	60.9 [38.8,79.5]	56.0 [35.3,75.0]
Probable	10	367	41	0	87.8 [84.2,90.7]	90.2 [86.8,92.8]
Possible	0	5	0	0	0.0 [0.0,53.7]	0.0 [0.0,9.6]
Conditional	1	27	4	4	11.1 [3.6,27.0]	100.0 [39.6,100.0]

real-world use.

4.4. On model building

The Bayesian network has two main objectives: (a) predict the most probable causality degree the expert will assign to the report, and (b) be interpretable by the pharmacist team members to inspect which variables are causing the shifts in causality classification. The first objective would clearly take advantage of a classifier approach (e.g. naive Bayes or Tree-Augmented Naive Bayes). However, the interpretation of such models is challenging for a health professional. Therefore, the definition of the structure was, indeed, ad-hoc, trying to give meaning to the dependences, while not losing in terms of predictive accuracy (we have compared results with classifier approaches and the differences were negligible, hence not shown). The rationale was to define as root nodes those factors which describe drug-event causality in a broader sense (is the event described in literature, was it reintroduced previously and did it reappeared) and as leaf nodes those who represent observations of the actual event (who reported, how was the drug administrated, if it was suspended and condition improved, if there are suspicion of interactions). There are, however, two variables which are general and should, therefore, be modelled as root nodes and were not, for different reasons. First, pharmaceutical group includes dozens of sparse categories; modelling it as an ascendant node would block the possible inference we would like to do with it, as conditional probabilities of degrees given each category of this variable are several times not possible to compute. Then, drug ineffectiveness relates to the known possibility of a drug simply failing in its main purpose; this factor was known to be used by the medical expert directly as a trigger for the *Conditional* degree, so we applied the classifier approach to it, instead.

The choice to have four different nodes for causality degrees instead of one single node with four categories is not without debate. Although, in theory, the four degrees are mutually exclusive (despite *Conditional* being not so clearly exclusive), the fact is that the uncertainty surrounding the classification is so high that we prefered to model the dependences of the factors to each causality degree separately, to allow for higher degrees of freedom in the inference process. We tested models with single node causality degree and the predictive results were similar. Furthermore, the pharmacists team expressed also benefits in having such modelling (instead of usual one outcome classifier approach) to enable a richer interactive analysis of the reports.

Also debatable is the use of separate nodes for *Suspended / Improved*, *Reintroduced / Reappeared* and *Concomitant / SuspectedInteraction*, when there seems to exist a straightforward link between each pair which could allow to use a single node for each pair, with three states (as used in the matrix). In reality, this theoretical definition is not entirely true. The team usually works with incomplete forms (e.g. one could know that the drug was suspended but not knowing if it improved or not the condition, *Reintroduced* actually means it might be a rechallenge of a drug which was taken a long time ago and not directly related to this case, so we do not know the exact response for *Reappeared*). Furthermore, *Suspected Interaction* might be defined even if there were no concomitant medications reported (the team might induce this from a narrative section of the report). Therefore, although in theory these could be modelled as single nodes, we preferred to have separate nodes where the team members could exactly work with the uncertainty of each of them (leaving them random if no information comes from the report).

Causality assessment by the expert has also some limitations [5,7]. During this activity, personal expectations and beliefs can influence the assessment. As so, there is always some randomness at the time of evaluation. This subjectivity is hard to be replicated in a model as ours. Although our model learned data from the expert assessment, it tends to follow the causality assessment guidelines, which is not in line with this kind of subjectivity. For example, in ADR reports made by physicians the signs and symptoms are usually better described than in ADR reports made by other health professionals or consumers. As a consequence, the medical expert has more information about the ADR, as it is detailed by a peer, with the same language and structure. To try to solve this issue, we have included in the network the node *Notifier* which is intended to be a proxy to the manner the ADR is explained. Future developments could include learning a model with latent variables to try to capture these phenomena.

5. Concluding remarks

The derived Bayesian network model has been used in the Northern Pharmacovigilance Centre, in Portugal, for more than three years now, for causality assessment of ADR reports. Upon reception of an ADR report, at the pharmacovigilance centre, whilst the expert is still and always consulted for final assessment, the centre pharmacists can, in parallel, use the network to inform the notifier about the preliminary assessment, speeding up the process of the centre.

One important aspect of the creation of our model was to endow the pharmacovigilance team with an interactive model which could provide, along with the best prediction of the causality degree assigned by the expert, a visual interpretation of the interactions and dependences of different factors in the causality assessment process, more than the applied in other assessment algorithms (e.g. decision trees) or other less-interpretable models (e.g. neural nets). The additional goal (which is out of scope of this paper) was to improve the parallel assessment done by the team, preparing the final report for the national and European institutions of pharmacovigilance.

Furthermore, the simplified matrix will allow the notifier to have an immediate feedback on their submission, improving engagement in the entire voluntary process. As suggested by other authors, to provide technical information during the act of ADR reporting seems to stimulate health professionals for this activity [27–29]. Thus, information on causality assessment presented in real time can improve the experience of ADR reporting, turning it into an engaging activity. Therefore, we believe that this tool will increase the motivation of health professionals to report their suspicions of ADR, with a consequent improvement of drug safety profile knowledge, resulting in a better public health protection.

We believe that this network can be very useful to other pharmacovigilance centres, mainly to those that do not have access to a full-time expert to evaluate ADR reports. As every method for ADR causality assessment [4], the presented Bayesian network has some advantages but also some limitations. Nonetheless, the network allows to shorten the time to causality assessment, which is a main issue in pharmacovigilance activities, and is accurate for most of the cases. Therefore, this method does not replace the expert evaluation, but can be used to complement it. Furthermore, future work will focus on refining the model, learning a new classifier from the (now more complete) data recorded from 2015 onwards, validate it against other alternatives while studying the interpretability and usability of different approaches. Hopefully, this will be included in a future global comparison on the cost-effectiveness of the system, measuring both the validity and the effort needed to achieve a classification by each method.

Acknowledgements

This work has been developed under the scope of project NanoSTIMA [NORTE-01-0145-FEDER-000016], which was financed by the North Portugal Regional Operational Programme [NORTE 2020], under the PORTUGAL 2020 Partnership Agreement, and through the European Regional Development Fund [ERDF]. UFN is an entity funded by INFARMED, the Portuguese National Authority of Medicines and Health Products, IP.

References

[1] World Health Organization. Uppsala Monitoring Centre. 2017http://www.who-umc.org/.

[2] Lindquist M, Staahl M, Bate A, Edwards IR, Meyboom RH. A retrospective evaluation of a data mining approach to aid finding new adverse drug reaction signals in the WHO international database. Drug Saf 2000;23(6):533–42. https://doi.org/10.2165/00002018-200023060-00004.

[3] INFARMED. Farmacovigilância em Portugal. Tech. rep. 2004.

[4] Agbabiaka TB, Savović J, Ernst E. Methods for causality assessment of adverse drug reactions: a systematic review. Drug Saf 2008;31(1):21–37http://www.ncbi.nlm.nih.gov/pubmed/18095744.

[5] Arimone Y, Bégaud B, Miremont-Salamé G, Fourrier-Réglat A, Moore N, Molimard M, Haramburu F. Agreement of expert judgment in causality assessment of adverse drug reactions. Eur J Clin Pharmacol 2005;61:169–73. https://doi.org/10.1007/s00228-004-0869-2.

[6] Arimone Y, Miremont-Salamé G, Haramburu F, Molimard M, Moore N, Fourrier-Réglat A, Bégaud B. Inter-expert agreement of seven criteria in causality assessment of adverse drug reactions. Br J Clin Pharmacol 2007;64(4):482–8. https://doi.org/10.1111/j.1365-2125.2007.02937.x.

[7] Miremont G, Haramburu F, Bégaud B, Péré JC, Dangoumau J. Adverse drug reactions: physicians' opinions versus a causality assessment method. Eur J Clin Pharmacol 1994;46:285–9. https://doi.org/10.1007/BF00194392.

[8] Naranjo CA, Busto U, Sellers EM, Sandor P, Ruiz I, Roberts EA, Janecek E, Domecq C, Greenblatt DJ. A method for estimating the probability of adverse drug reactions.

Clin Pharmacol Ther 1981;30:239–45. https://doi.org/10.1038/clpt.1981.154.

[9] Jones J. Adverse drug reactions in the community health setting: approaches to recognizing, counseling, and reporting. Fam Community Health 1982;5(2):58–67http://www.ncbi.nlm.nih.gov/pubmed/10278126.

[10] Karch FE, Lasagna L. Toward the operational identification of adverse drug reactions. Clin Pharmacol Ther 1977;21(3):247–54. https://doi.org/10.1002/cpt1977213247.

[11] Latoszek-Berendsen A, Tange H, van den Herik HJ, Hasman A. From clinical practice guidelines to computer-interpretable guidelines. A literature overview. Methods Inf Med 2010;49(6):550–70. https://doi.org/10.3414/ME10-01-0056.

[12] Lucas P. Bayesian analysis, pattern analysis, and data mining in health care. Curr Opin Crit Care 2004;10(5):399–403http://www.ncbi.nlm.nih.gov/pubmed/15385759.

[13] Mitchell TM. Machine Learning. McGraw-Hill; 1997.

[14] Schurink CAM, Lucas PJF, Hoepelman IM, Bonten MJM. Computer-assisted decision support for the diagnosis and treatment of infectious diseases in intensive care units. Lancet 2005;5(5):305–12. https://doi.org/10.1016/S1473-3099(05)70115-8.

[15] Lucas PJF, van der Gaag LC, Abu-Hanna A. Bayesian networks in biomedicine and health-care. Artif Intell Med 2004;30(3):201–14. https://doi.org/10.1016/j.artmed.2003.11.001.

[16] Sackett DL, Rosenberg WM, Gray JA, Haynes RB, Richardson WS. Evidence based medicine: what it is and what it isn't. BMJ 1996;312(7023):71–2.

[17] Darwiche A. Bayesian networks. Commun ACM 2010;53(12):80–90. https://doi.org/10.1145/1859204.1859227.

[18] Velikova M, van Scheltinga JT, Lucas PJ, Spaanderman M. Exploiting causal functional relationships in Bayesian network modelling for personalised healthcare. Int J Approx Reason 2014;55(1):59–73. https://doi.org/10.1016/j.ijar.2013.03.016.

[19] Dias CC, Rodrigues PP, Coelho R, Santos PM, Fernandes S, Lago P, Caetano C, Rodrigues Â, Portela F, Oliveira A, Ministro P, Cancela E, Vieira AI, Barosa R, Cotter J, Carvalho P, Cremers I, Trabulo D, Caldeira P, Antunes A, Rosa I, Moleiro J, Peixe P, Herculano R, Gonçalves R, Gonçalves B, Sousa HT, Contente L, Morna H, Lopes S, Magro F. Development and validation of risk matrices for Crohn's disease outcomes in patients who underwent early therapeutic interventions. J Crohn's Colitis 2017;11(4):445–53. https://doi.org/10.1093/ecco-jcc/jjw171.

[20] Darwiche A. Modeling and Reasoning with Bayesian Networks. Cambridge University Press; 2009http://www.amazon.com/Modeling-Reasoning-Bayesian-Networks-Darwiche/dp/0521884381.

[21] Scutari M. Learning Bayesian networks with the bnlearn R Package. J Stat Softw 2010;35:22http://arxiv.org/abs/0908.3817.

[22] Højsgaard S. Graphical independence networks with the gRain package for R. J Stat Softw 2012;46(10)http://www.jstatsoft.org/v46/i10/paper.

[23] Lauritzen DJ, Spiegelhalter SL. Local computations with probabilities on graphical structures and their application to expert systems. J R Stat Soc Ser B 1988;50(2):157–224. https://doi.org/10.2307/2345762http://wrap.warwick.ac.uk/24233/.

[24] Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez J-C, Müller M. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinf 2011;12:77.

[25] Grau J, Grosse I, Keilwagen J. PRROC: computing and visualizing precision–recall and receiver operating characteristic curves in R. Bioinformatics 2015;31(15):2595–7. https://doi.org/10.1093/bioinformatics/btv153.

[26] R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2015https://www.r-project.org/.

[27] Jean-Pastor M, Affaton M, Prost N, Rodor F. Pharmacovigilance information via electronic mail. Fundam Clin Pharmacol 2001;15(Suppl. 1):12.

[28] Johansson ML, Brunlöf G, Edward C, Wallerstedt SM. Effects of e-mails containing ADR information and a current case report on ADR reporting rate and quality of reports. Eur J Clin Pharmacol 2009;65(5):511–4. https://doi.org/10.1007/s00228-008-0603-6.

[29] Lapphra K, Dobson S, Bettinger JA. Acceptability of Internet adverse event self-reporting for pandemic and seasonal influenza immunization among health care workers. Vaccine 2010;28(38):6199–202. https://doi.org/10.1016/j.vaccine.2010.07.019.