Journal of Biomedical Informatics 118 (2021) 103780

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

Original Research

Predicting treatment recommendations in postmenopausal osteoporosis

G. Bonaccorsi [a], M. Giganti [b], M. Nitsenko [c], G. Pagliarini [c], G. Piva [c], G. Sciavicco [c,*]

[a] Dept. of Translat. Med. and for Romagna, Menopause and Osteoporosis Center, University of Ferrara, Italy
[b] Dept. of Translat. Med. and for Romagna, and Exper. Med., Section of Radiology, Univ. of Ferrara, Italy
[c] Dept. of Mathematics and Computer Science, Univ. of Ferrara, Italy

ARTICLE INFO

Keywords:
Keywords: Osteoporosis treatment
Machine learning
Rule extraction
Clinical decision support system

ABSTRACT

We designed, implemented, and tested a clinical decision support system at the Research Center for the Study of Menopause and Osteoporosis within the University of Ferrara (Italy). As an independent module of our system, we implemented an original machine learning system for rule extraction, enriched with a hierarchical extraction methodology and a novel rule evaluation technique. Such a module is used in everyday operation protocol, and it allows physicians to receive suggestions for prevention and treatment of osteoporosis. In this paper, we design and execute an experiment based on two years of data, in order to evaluate and report the reliability of our suggestion system. Our results are encouraging, and in some cases reach expected accuracies of around 90%.

1. Introduction

Osteoporosis is associated with increased fragility of bone and a subsequent increased risk of fracture. The diagnosis of osteoporosis is intimately linked with the imaging and quantification of bone and *bone mass density* (*BMD*). Scanning modalities, such as dual-energy X-ray absorptiometry or quantitative CT, have been developed and honed over the past half century to provide measures of BMD and bone microarchitecture for the purposes of clinical practice and research. *Dual-energy X-ray absorptiometry* (*DXA*), in particular, is a means of measuring BMD using spectral imaging. Two X-ray beams, with different energy levels, are aimed at the patient's bones. When soft tissue absorption is subtracted out, the bone mineral density can be determined from the absorption of each beam by the bone. DXA is the gold standard technique for the diagnosis of osteoporosis as per international guidelines, and it is the most widely used and most thoroughly studied bone density measurement technology. The DXA scan is typically used to diagnose and follow up on osteoporosis, in contrast to the nuclear bone scan, which is sensitive to certain metabolic diseases of bones in which bones are attempting to heal from infections, fractures, or tumors. Patient management, prevention, and treatment of osteoporosis can be eased by associating a *clinical decision support system* (CDSS) to the DXA bone density evaluation, so that each scan can be paired with the patient's clinical history, diagnosis, and treatment recommendations. Automating such an association may facilitate every step of patient management, but above all, it may ease data aggregation, analysis, and knowledge extraction. Many publications convey the belief that CDSS, especially on electronic health record platform, offer large gains in clinical performance, improving its quality, safety, and cost efficiency [6,4].

We designed, implemented, and tested a clinical decision support system at the *Research Center for the Study of Menopause and Osteoporosis* (*CMO*) within the University of Ferrara (Italy). Our CDSS includes an original machine learning system for rule extraction, primarily based on the *Ripper* algorithm [9], enriched with a hierarchical extraction methodology and a novel rule evaluation technique, which offers to physicians suggestions on therapy recommendation. Such a module is used in everyday operation protocol, according to the following high-level algorithm: *(i)* each week, the entire database of patients is consulted, and a new rule extraction is performed during non-active hours; *(ii)* then, once the anamnesis, DXA results, and diagnosis of a new patient are recorded, the system can be queried for a therapy recommendation suggestion, based on the extracted rules; *(iii)* finally, the physician records the actual recommendation, and the new patient is used (starting from the next rule extraction cycle) to improve the suggestions. Our contribution is multi-fold: *(i)* we present an actual, working CDSS with an intelligent component which allows physicians to perform correct and appropriate diagnoses that follows the international standards [1,14,16,23,11], and to identify patients with high and very high fracture risk as per internationally and nationally validated algorithms such as FRAX and DeFRA [15,18,10,17]; *(ii)* we describe its intelligent module, which learns from the records of past patients and provides suggestions for new ones, and *(iii)* we test our system on real data to

* Corresponding author.
E-mail address: guido.sciavicco@unife.it (G. Sciavicco).

https://doi.org/10.1016/j.jbi.2021.103780
Received 17 October 2020; Received in revised form 10 March 2021; Accepted 5 April 2021
Available online 20 April 2021

evaluate its expected accuracy. The motivations and clinical usefulness of our system include the following considerations. First, postmenopausal osteoporosis is a chronic, progressive, and highly undertreated condition, associated with high morbidity, mortality, and economic burden, and recent European data show that only a small proportion of osteoporotic patients are being treated [12,3]; the lack of specific care pathways for osteoporotic patients is certainly to be listed among the causes of such a situation. Our tool eases the design of such pathways, allowing the identification of the patients at risk in the appropriate way and on the basis of a trained evaluation, and promoting the development of a care network modeled on IOF[1] liaison fracture service. Second, it is ever more important to design and implement systems that help physicians during the decision-making phases, not only based on the accepted guidelines, but also based on individual patient risk profiles. Finally, the general practitioners, assistants, and students may not have in-depth knowledge of the specific guidelines for osteoporotic treatment recommendation; our system is trained by specialists, which guarantees the reliability of its rules, and it can be used for decision support for others.

The paper is structured as follows. In Section 2, we review the most important literature on CDSSs for osteoporosis management and treatment. Then, in Section 3 we briefly describe our CDSS and the knowledge extraction algorithm. In Section 4, we design a rule evaluation experiment. To this end, we consider 2052 postmenopausal women over 40 years of age that underwent a DXA exam at CMO from September 1st, 2018 to August 31st, 2020, and we used such patients to train our intelligent system, and test its expected accuracies. Finally, we conclude with some future work directions.

2. Background

2.1. Algorithms for the risk of fracture estimation

The clinical irreversible event of osteoporosis is a fragility fracture. In this clinical setting, an important role is played by the tools for prediction of fracture risk, and among these, in the context of the digitalization of osteoporosis treatment and prevention, a special role is played by the algorithms for absolute fracture risk estimation [21]. FRAX [15,18,17] is an internationally validated and one of the most widely used fracture risk assessment tools. In the past ten years, it has facilitated the assessment of fracture risk on the basis of fracture probability. FRAX integrates the influence of several BMD-independent, well-validated risk factors for fracture with the BMD (when available). The DERIVED FRAX algorithm (DeFRA) was developed by the Italian Society for Osteoporosis, Mineral Metabolism, and Skeletal Diseases (SIOMMMS) with the aim of further increasing the sensitivity of fracture risk assessment, and its international validation is still ongoing [10]. For a comparison between these tools, see also [2,5].

2.2. Clinical decision support systems

A *clinical decision support system* (*CDSS*) is a computerized system intended to improve healthcare delivery by enhancing medical decisions with targeted clinical knowledge, patient clinical risk factors, and other health information. A traditional CDSS is comprised of software designed to be a direct aid to clinical-decision making, in which the characteristics of an individual patient are matched to a computerized clinical knowledge base and patient-specific assessments or recommendations are then presented to the clinician for a decision. CDSSs today are primarily used at the point-of-care, for the clinician to combine their knowledge with information or suggestions provided by the CDSS. Increasingly, however, there are CDSSs being developed with the capability to leverage data and observations otherwise unobtainable or uninterpretable by humans. CDSSs have been classified and subdivided into various categories and types, including intervention timing, and whether they have active or passive delivery. More frequently, CDSS are classified as *knowledge-based* or *non-knowledge based*.

In knowledge-based systems, rules are created, with the system retrieving data to evaluate the rule, and producing an action or output. Rules can be made using literature-based, practice-based, or patient-directed evidence. Two relevant examples of knowledge-based CDSSs for osteoporosis are [13,24]. In [13], the authors present OPAD, a clinical decision support system that utilizes clinical information from international guidelines and experts in the field of osteoporosis. Physicians are provided with a user interface to insert standard patient data, from which OPAD generates instant diagnostic comments, 10-year risk of fragility fracture, treatment options for the given case, and when to offer a follow-up DXA-evaluation. Thus, the medical decision making is standardized according to the best expert knowledge at any given time. OPAD was evaluated with a data set of 308 randomly selected individuals, and its ten-year fracture risk computation is nearly identical to FRAX. In [24], on the other hand, a fuzzy inference system was presented, with the purpose of diagnosing osteoporosis or osteopenia in the same way as an expert physician would do. The degree of severity is calculated via a series of membership functions and inference fuzzy rules, and then confirmed by the physician. The driving idea was to propose an alternative diagnosis system based on X-ray imaging, instead of DEXA, and supported by the fuzzy inference system. Beside pure CDSSs, however, it is worth recalling that there are several studies on rule designing and evaluation; examples include [19,20,27]. Non-knowledge based CDSS still require a data source, but the decision leverages artificial intelligence methods primarily based on pattern recognition, and more in general, machine learning techniques, rather than being programmed to follow expert medical knowledge. Non-knowledge based CDSSs, although a rapidly growing use case for artificial intelligence in medicine, are rife with challenges, including problems in understanding the logic that used to produce recommendations and problems with data availability, and they have yet to reach widespread implementation [25]. Examples include [8], a system based on an artificial neural network trained on seven input variables (sex, age, weight, height, body mass index, postmenopausal status, and coffee consumption), [28], in which an ensemble of data mining techniques was developed for predicting the risk of osteoporosis prevalence in women that consisted of combining decision trees and artificial neural networks, and [22], in which the author proposed a predictive model for osteoporosis based on the random forest algorithm, trained with attributes such as age, gender, early menopause status, heredity, behavioural aspects, lifestyle, and exercise levels, as well as pharmacological data.

2.3. Discussion

A common characteristics of all these systems is that they are designed for diagnosis prediction. The intelligent module for the CDSS designed for CMO, instead, is focused on therapy recommendation prediction. As observed in [7], an effort should be made in order to systematize and standardize the application of artificial intelligence techniques to the osteoporosis problem; this work can certainly be considered a step in this direction, considering that the CDSS is used in the everyday protocol, validated by physicians, and the information contained in it has the potential for offering approaches to several problems, not only therapy recommendation prediction.

3. A CDSS for osteoporosis patient management

3.1. Patient management at CMO

A Hologic Discovery (Horizon W®) DXA system is installed at CMO, and connected to a Windows PC, with proprietary software for single-patient management, archiving, and recall. Each scan produces a PDF

[1] International Osteoporosis Foundation.

Table 1
Attributes: patient sex and menopausal status.

attribute name	domain
menopause state	induced
	premenopause
	spontaneous menopause
	perimenopause
	histerectomy
	histerectomy + unilateral ovariectomy
	histerectomy + bilateral ovariectomy
age at menopause	*numeric*

Table 2
Attributes: current and past therapy osteoprotective status.

attribute name	domain
therapy status	on therapy
	never on therapy
	therapy suspended
hormonal osteoprotective therapy	*binary*
specific osteoprotective therapy	*binary*
vitamin D based treatment	*binary*
other osteoprotective treatment	*binary*

document with a lumbar spine and a hip image, and that contains all relevant values of analysis: bone area, bone mineral content (BMC), bone mineral density (BMD), T-score, and Z-score (refer to the International Society for Clinical Densitometry guidelines for a detailed explanation of these values[2]). Hologic software also allows to collect the relevant clinical risk factors for calculating the FRAX, which is also displayed. Our CDSS is connected with the PC, and automatically recognizes the presence of a new scan; a new scan generates a new record. Such a record is archived in a MySQL database for its successive elaboration. Each record, connected to a specific patient, is associated to several fields (or attributes), filled in by the physician, that integrate and extend those collected by the machine, and may be categorized into: *(i)* personal data, *(ii)* clinical history, *(iii)* previous or current treatment(s), *(iv)* current diagnosis, based on the current scan, and *(v)* therapy and follow-up recommendations.

3.2. Treatment recommendation prediction

Therapy prediction is a dynamic, cyclic process designed in four phases: *(i) data querying and pre-processing, (ii) rules extraction, (iii) rules evaluation, and (iv) prediction*. In order to preserve interpretability, our rule-based system is explicit.

Data querying and pre-processing. Phases (i), (ii), and (iii) are performed at specific, preset times. Once a week, during closing hours, the necessary queries are executed to extract the relevant attributes, each instance corresponding to a patient's record. As a result, all records ever processed by our CDSS are part of the current training, and recurring patients are treated as different individuals. Attributes must be pre-processed before rule extraction, and pre-processing must be automatic. We designed several pre-processing types, depending on the type of attribute, its semantics, and the idiosyncrasies that emerge from daily use by the physician. After pre-processing, attributes can be separated into the following groups:

- General data (Table 1). These include the data on the menopausal status of the patient, that is, *menopausal status* and *age at menopause*.
- Current and past osteoprotective therapy (Table 2). These attributes contain information concerning whether or not the patient has undertaken or is currently on some osteoprotective therapy. We have included the following fields: *therapy status*, *hormonal osteoprotective therapy*, *specific osteoprotective therapy*, *vitamin D treatment*, *other supplementation* (including calcium and other vitamin intake).
- FRAX and DeFRA risk factor data (Table 3). These are *age*, *body mass index*, *vertebral fragility fractures*, *femoral fragility fractures*, and *fragility fracture familiarity*, which take into account parental and

Table 3
Attributes: FRAX and DeFRA data and predictions. *presup**: can be NULL

attribute name	domain
age	*numeric*
body mass index	*numeric*
vertebral fragility fractures	none
	one
	more than one
femoral fragility fractures	none
	one
	more than one
fragility fractures in other sites	*binary*
fragility fracture familiarity	*binary*
smoking habits	no
	$\leqslant$10 cigarettes/day
	$>$ 10 cigarettes/day
alcohol intake	no
	$<$ 3 units/day
	$\geqslant$3 units/day
cortisone (Prednison equivalent)	no
	$\geqslant$2.5 mg/day and $\leqslant$5 mg/day
	$>$ 5 mg/day
rheumatoid arthritis	*binary*
psoriatic arthritis	*binary*
systemic lupus	*binary*
scleroderma	*binary*
other connective tissue diseases	*binary*
diabetes mellitus	*binary*
adult osteogenesis imperfecta	*binary*
untreated chronic hyperthyroidism	*binary*
hypogonadism	*binary*
early menopause	*binary*
chronic malnutrition	*binary*
inflammatory bowel disease	*binary*
chronic liver diseases	*binary*
*FRAX (major fractures)**	*numeric (0 if $<$ 0.1)*
*FRAX (femur)**	*numeric (0 if $<$ 0.1)*
*DeFRA**	*numeric (0 if $<$ 0.1, 50 if $>$ 50)*

Table 4
Attributes: Previous and current illnesses, previous DXA. *: can be null.

attribute name	domain
endometrial pathologies	*binary*
breast cancer	*binary*
vasomotor symptoms	*binary*
distrofic symptoms	*binary*
dyslipidemia	*binary*
hypertension	*binary*
venous thromboembolism risk factors	*binary*
cardiac pathologies	*binary*
vascular pathologies	*binary*
kidney failure	*binary*
respiratory pathologies	*binary*
oral pathologies	*binary*
esophageal pathologies	*binary*
gastroduodenitis	*binary*
gastrectomy	*binary*
bowel resection	*binary*
other diseases	*binary*
*serum vitamin level**	*numeric*
*previous total spine T-score**	*numeric*
*previous total spine Z-score**	*numeric*
*previous neck T-score**	*numeric*
*previous neck Z-score**	*numeric*

[2] *https://www.iscd.org/official-positions/*.

Table 5
Attributes: Current scan and diagnosis.

attribute name	domain
severe osteoporosis	*binary*
femur status	normal
	osteopenia
	osteoporosis
spine status	normal
	osteopenia
	osteoporosis
total spine T-score	*numeric*
total spine Z-score	*numeric*
neck T-score	*numeric*
neck Z-score	*numeric*

Table 6
Therapy and supplementation types. MHT: menopause hormonal therapy. TSEC: tissue selective estrogen complex.

name	type	domain	abbr.
	(first level)	(second level)	
T_{horm}	hormonal therapy	MHT (tibolone)	*tib*
		MHT (oral)	*oral*
		MHT transdermal	*tsec*
		MHT (TSEC)	*trans*
T_{osteop}	osteoprotective therapy	alendronate	*ale*
		alendronate + vit D	*ale* + *vD*
		risedronate	*ris*
		ibandronate	*iba*
		clodronate	*clo*
		raloxifene	*ral*
		bazedoxifene	*baz*
		denosumab	*den*
		teriparatide	*ter*
		zoledronate	*zon*
T_{vitDth}	vitamin D therapy	colecalciferol	*colec*
		calcifediol	*calci*
$S_{vitDsup}$	vitamin D supplementation	colecalciferol	*colec*
		calcifediol	*calci*
S_{calsup}	calcium supplementation	carbonated calcium	*carb*
		citrated calcium	*citr*

personal (after 50 years of age) history of hip or vertebral fragility fractures, *fragility fractures in other sites* to account for prior low-trauma fractures at other sites (after 50 years of age), *smoking habits* to take into account smoking habits in terms of number of cigarettes/day, *alcohol intake* to account for alcohol use in terms of units/day, *cortisone*, that is, long-term use and dosage of glucocorticoids, *rheumatoid arthritis* and other connective tissue diseases (*psoriatic arthritis*, *systemic lupus*, *scleroderma*, and *other connective tissue diseases*), and causes of secondary osteoporosis (*diabetes mellitus*, *adult osteogenesis imperfecta*, *untreated chronic hyperthyroidism*, *hypogonadism*, *early menopause*, *chronic malnutrition* or malabsorption, *inflammatory bowel disease*, and *chronic liver diseases*). Moreover, we collect here the results of FRAX and DeFRA algorithms, that is: *FRAX (major fractures)*, *FRAX (femur)*, and *DeFRA*. With respect to the raw values of these indicators, the following pre-processing has been applied: if FRAX (resp., DeFRA) is not applicable, the value is NULL; if it is applicable and the result is less than 0.1, its value becomes 0, and if it is greater than 50 (for DeFRA only), the value is 50.

- Previous and current illnesses (those not already taken into account in the previous group), and previous DXA values (Table 4). Here, we collect useful information that concerns the presence or absence of previous and current illnesses that may influence therapy recommendation. All attributes here are binary. They are: *endometrial pathologies*, *breast cancer*, *vasomotor symptoms*, *distrofic symptoms*, *dyslipidemia*, *hypertension*, *venous thromboembolism risk factors*, *cardiac pathologies*, *vascular pathologies*, *kidney failure*, *respiratory pathologies*, *oral pathologies*, *esophageal pathologies*, *gastroduodenitis*, *gastrectomy*, *bowel resection*, *other diseases*, *serum vitamin level*. If a previous DXA is present, then we record *previous DXA total spine T-score*, *previous total spine Z-score*, *previous neck T-score*, and *previous neck Z-score*.
- Scan values and diagnosis (Table 5). The last group of attributes deals with the current scan, the values that it contains, and the diagnosis that corresponds to these values. In particular we consider the attributes *severe osteoporosis* (that takes into account previous fragility fractures - this is a computed value), *femur diagnosis*, *spine diagnosis*, *neck T-score*, *neck Z-score*, *total spine T-score*, and *total spine Z-score*.

All attributes in Table 1 to Table 5 are used in the predictions.

Rule extraction. A *rule*, in general, is a formula of the type:

$$\rho : p_1 \wedge \ldots \wedge p_n \rightarrow \mathscr{C},$$

where $p_1, \ldots, p_n$ are logical propositions over the attributes of the problem, and $\mathscr{C}$ is a class. In our context, examples of p_is include *ethnic group is white* and *body mass index is less than 20*. More in general, p_i is called a *decision*, and its format depends on the particular type of attribute on which it is taken. CMO's CDSS allows for five types of therapy and supplementation recommendations, as detailed in Table 6 (first and second columns). These are considered separate problems, and denoted by $T_{problem}$ (therapy), where $problem \in \{horm, osteop, vitDth\}$ or $S_{problem}$ (supplementation), where $problem \in \{vitDsup, calsup\}$; so for example, T_{horm} is the problem of establishing if a patient needs a hormonal therapy; these, in our nomenclature, are called *upper level* problems. As we have designed them, each problem is binary, and in the previous example, establishing if a patient needs hormonal therapy is answered either with yes or No.To each problem is associated a *classifier* (that is, a structured set of rules), extracted by executing (our implementation of) the so-called Ripper algorithm. Suppose, now, that we are able to solve a particular upper level problem, say T_{horm}. In turn, this raises the *lower level* problem of establishing which drug or supplement to recommend to a particular patient for whom a certain therapy or supplementation has been recommended by the system. Lower level problems are also listed in Table 6 (third column), and are denoted with superscripts. So, continuing our initial example, T^{tib}_{horm} would be the problem of establishing if a patient, to whom a hormonal therapy has been recommended, should be recommended to take tibolone or not. As upper level ones, lower level problems are binary.

The *Ripper* (*Repeated incremental pruning to produce error reduction*) algorithm [9] improves upon earlier deterministic rule-based classification algorithms by generating systems that in most cases have been shown to be more reliable than decision trees. It can be seen as a three-step process; such steps are usually called *grow*, *prune*, and *optimize*. The first step uses a 'separate and conquer' method to add conditions to a rule until it perfectly classifies a subset of data. Just like decision trees, the *information gain* criterion is used to identify the next splitting attribute. When increasing a rule's specificity no longer reduces entropy, the rule is immediately pruned. Until reaching stopping criterion, step one and two are repeated; then the whole set of rules is optimized using a variety of heuristics. Pruning is performed by extracting a subset of the training set called *pruning* set, and using it to confirm, or modify, the extracted rules. For a single problem, the resulting classification system (denoted, in general, with Γ) has the following aspect:

$$\Gamma = \begin{cases} p_1^1 \wedge \ldots \wedge p_{n_1}^1 \rightarrow \mathscr{C}_1, & \text{else} \\ p_1^2 \wedge \ldots \wedge p_{n_2}^2 \rightarrow \mathscr{C}_2, & \text{else} \\ \quad \ldots & \\ p_1^m \wedge \ldots \wedge p_{n_m}^m \rightarrow \mathscr{C}_m, & \text{else} \\ & \mathscr{C}_{m+1} \end{cases} \quad (1)$$

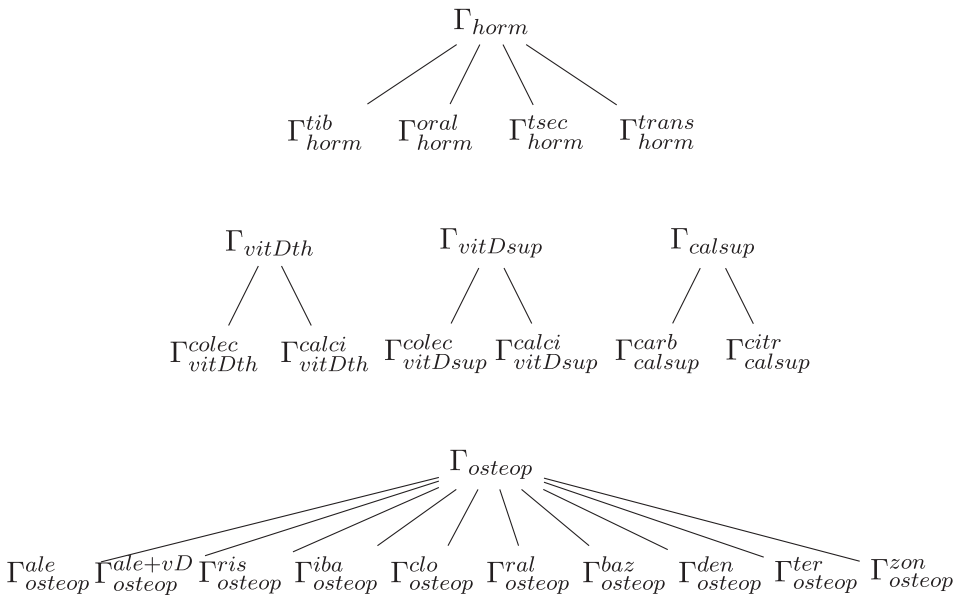

Fig. 1. A schematic representation of the hierarchical system.

As we have explained, our system is hierarchical. For each of the five problems in Table 6, a different classifier Γ is extracted that goes along with it and is denoted in the same way; therefore, to the problem $T_{problem}$ (resp., $S_{problem}, T^{drug}_{problem}, S^{drug}_{problem}$), we associate the classifier $\Gamma_{problem}$ (resp., $\Gamma^{drug}_{problem}$). Because each problem is different, it is extracted from a different data set; however, each of such data sets is a snapshot of the original one. So, for example, in order to extract the classifier Γ_{horm} that solves the problem T_{horm} we consider the original data set D, keep all columns explained from Table 1 to Table 5 plus the binary column concerning the hormonal therapy, and produce in this way the data set D_{horm}, which is used by Ripper for the extraction of the classifier. For all other problems, at both the upper and the lower level, we proceed in the same way. In general, therefore, to each problem we associate both a classifier $\Gamma^{drug}_{problem}$ and a data set $D^{drug}_{problem}$ from which the classifier is extracted. Once the data set is created, however, a simple *balancing* condition is checked; if indeed, a data set is too unbalanced, the problem is simply ignored for lack of information. For example, suppose that, in the set of all patients, almost all have been treated with hormonal therapy: in this case, there is not enough information for us to extract significant rules to decide if a patient should be recommended an hormonal therapy or not. To this end, we used a pre-defined parameter p, and filtered out all problems in which the less numerous class has a cardinality less than p% of the most numerous one. It is important at this point to stress that our system is designed to work in a continuous way; rules are continuously extracted, evaluated, and applied, and as new patients enter our database, rules may be replaced by new ones. So, our design includes a *self-adapting* extraction technique, and each problem must be re-evaluated at each step. In this way, for example, the same problem may present a balanced data set at a certain moment, which becomes unbalanced after a month, and then balanced again. This justifies the introduction of parameters such as p, above. (Fig. 1)

Rule evaluation. Let us focus now on a specific classifier Γ and its corresponding data set D. Phases (*ii*) and (*iii*), that is, (automatic) classifier extraction and rule evaluation, start by separating D into two data sets, D_{train} and D_{test}, so that the latter contains the k% *most recent* instances, where k is a system parameter. The former is the one actually used by the Ripper algorithm to extract Γ, while the latter is used for evaluation. So, the first time an instance enters the system is when a classifier is applied to it and the physician receives a suggestion; later, the same instance starts playing the role of a test instance, for the rules to be evaluated before a new suggestion is made; finally, the instance will eventually enter the training set, where it will stay. In this way, the learning system is somehow aware of the time component. If a new policy in therapy recommendation starts being applied, at the beginning it will not be taken into account for suggestion, but at a certain moment it will become part of the training, and eventually rules will be extracted for that policy as well. Separation between training and test is performed in a *stratified* fashion: for each class, the k% of instances that belong to that class are selected, and, at the end, all selected subsets are joined together. There are two levels of evaluation: we study how Γ behaves *as a whole* (*global evaluation*), and how each of its specific rule does (*local evaluation*). Global evaluation is used for internal assessment only. whereas local evaluation is used also in everyday protocol, to add useful information to the physician:

- Global evaluation. Being based on a classical machine learning algorithm, our system is subject to classical statistical evaluation. For binary problems (such as, for example, deciding if a patient needs hormonal therapy), evaluation on a test set starts by computing the four classic indicators: *true positive* cases (TP, the cases in which the prediction and the real value are both 'yes'), *true negative* cases (TN, the cases in which the prediction and the real value are both 'no'), *false positive* cases (FP, the cases in which the prediction is 'yes' but and the real value is 'no'), and *false negative* cases (FN, the cases in which the prediction is 'no' but and the real value is 'yes'). On the basis of these four values, the standard performance indicators can be computed, e.g., accuracy, sensitivity, and so on.
- Local evaluation. Although rules are applied, as explained, in a specific order, in the context of making a prediction, each rule can be evaluated separately using the classic rule evaluation metrics. These are: *support* (s, the ratio of cases in which both the antecedent and the consequence of the rule are verified over the total number of cases), and *confidence* (c, the ratio of the of cases in which only the antecedent is verified over the cases in which both the antecedent and the consequent are verified). Intuitively, the support of a rule indicates its relevance, while the confidence indicates its reliability. In the context of association rule extraction (which is not our case), these values are used in the extraction algorithm. In our context, instead, we can use them to evaluate each single rule at test time; then, each time a particular rule is used for a prediction we can offer the physician not only a therapy suggestion, but also an estimation of the

Table 7
Single rule evaluation with support and confidence.

	$c > 0.7$	$c \leqslant 0.7$
$s > 0.2$	Relevant, reliable (I)	Relevant, unreliable (II)
$s \leqslant 0.2$	Irrelevant, reliable (III)	Irrelevant, unreliable (IV)

trustworthiness of the rule that has been used for that suggestion, according to Table 7. In short, such an estimation works as follows: if a rule has a sufficiently high confidence (more than or equal to 0.7), we say that it is *reliable*, and if it has a sufficiently high support (more than or equal to 0.2), we say that it is *relevant*. By combining these two indicators, we classify each rule. Along with the application of a rule for a suggestion, we display its classification; moreover, in the particular case of only having unreliable and irrelevant rules, we simply do not offer any suggestion.

Prediction. A classifier Γ is applied as follows: when a new record comes with all its relevant attributes filled up, rules are tried in the same order in which they are listed. A certain rule cannot be fired if at least one of its antecedents is not fulfilled. Thus, when a specific rule is in fact fired, we can list all the conditions that prevented the previous ones to be fired along with those that allowed that one to be. For example, consider Γ as in Eq. (1) and assume that the first rule cannot be fired because p_1^1 does not hold, but the second can; we can say that it holds:

$$\neq gp_1^1 \wedge p_1^2 \ldots \wedge p_{n_2}^2$$

and therefore C_2 can be predicted. A schematic representation of the entire system is shown in Fig. 2, and it can be summarized as follows. Patients already in the database constitute a data set from which rules are extracted. For new patients, the physician receives a suggestion (if possible) on therapy recommendation. These new patients are then recommended a therapy (based on the accepted protocols and gold standards), and inserted in the database; they will be part of the data used for the next rule extraction in the future.

4. Experiment

In this section we consider the recommendations that have been given from Sept. the 1st, 2018 to Aug. the 31th, 2020. Focusing on the specific attributes that we have identified in the previous section, we first give some basic descriptive statistics of our data, and, then, we execute a complete experiment and give the results. Finally, we discuss them.

4.1. Data, basic descriptive statistics, and experiment parameters

From Sept. the 1st, 2018 to Aug. the 31st, 2020, 2052 post-menopausal women over 40 years of age underwent a DXA examination at CMO. Of these, 18 patients returned more than once; our original data set, then, is composed of 2070 reports. Out of these reports, 16 presented a null BMI, 4 of them presented a recommendation of some type without drug or supplement specification, and 2 of them presented the same recommendation more than once (with the same drug or supplementation) because of human error during data insertion; these have been filtered out, leaving us with 2048 instances. Observe that at the first level, it holds that each instance is a record; at the second level, however, after selecting only those reports in which a particular recommendation type has been given, more than one recommendation for the same patient may have been selected, and therefore it holds that each instance is a recommendation. In Table 8, we give some basic statistic values of important attributes in our data set (neck BMD, total spine BMD, FRAX and DeFRA statistics have been computed limited to the applicable ones). The minimum recorded FRAX (major fractures) was 1.80; on the other hand, there are several instances in which FRAX

Table 8
Basic statistic values. *: values inferior than 0.1 have been replaced by 0 (25 cases); **: values inferior than 0.1 have been replaced by 0 (1 case); ***: values superior than 50 have been replaced by 50 (114 cases).

attribute name	min	max	mean	st. dev.
age	40	93	63.57	9.33
body mass index	15.62	56.36	25.41	4.54
spine BMD	0.45	2.62	0.83	0.17
neck BMD	0.33	1.31	0.66	0.10
FRAX (major fractures)	1.80	73	9.70	7.35
FRAX (femur)	0*	67	2.77	4.36
DeFRA	0**	50***	12.94	13.39
total spine T-score	−6.8	7.6	−1.77	1.28
total spine Z-score	−5.2	9.9	−0.14	1.32
neck T-score	−4.7	2.2	−1.71	0.90
neck Z-score	−2.6	4.1	−0.26	0.88

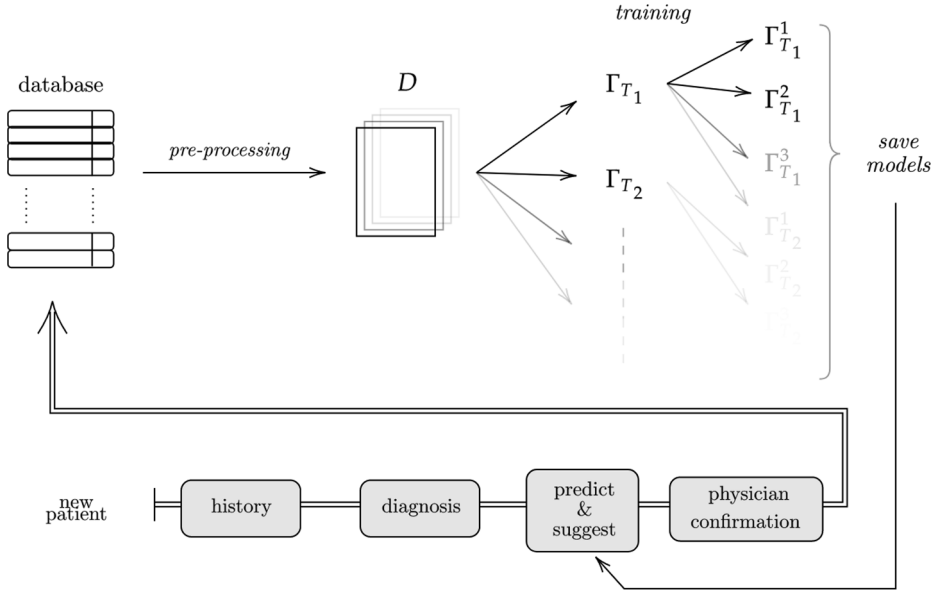

Fig. 2. A diagram showing the high-level flow of the system.

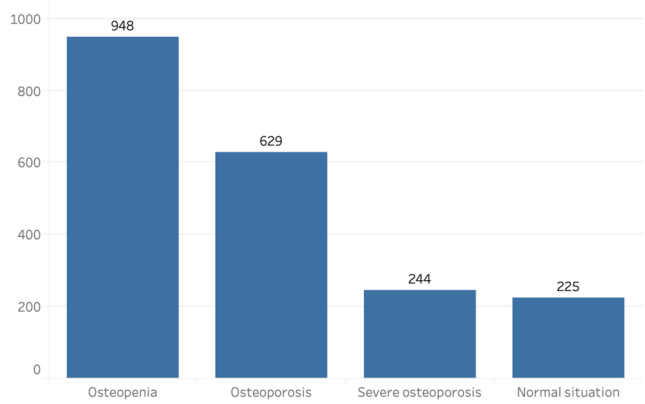

Fig. 3. Distribution of diagnoses.

Table 9
Number of instances per type of therapy and supplementation.

type	# of inst.	class value	# of inst.
(first level)		(second level)	
hormonal th.	56	MHT (tibolone)	7
		MHT (oral)	11
		MHT (transdermal)	9
		MHT (TSEC)	29
osteop. th.	447	alendronate	148
		alendronate + vit D	0
		risedronate	50
		ibandronate	2
		clodronate	39
		raloxifene	1
		bazedoxifene	39
		denosumab	140
		teriparatide	23
		zoledronate	5
vitamin D th.	53	colecalciferol	23
		calcifediol	30
vitamin D supp.	1125	colecalciferol	986
		calcifediol	141
calcium supp.	580	carbonated calcium	322
		citrated calcium	258

(femur) and/or DeFRA was less than 0.1 or more than 50 (for DeFRA only): these values have been replaced by 0 and 50, respectively (and discarded when computing mean and standard deviation). Diagnoses are distributed as in Fig. 3. In terms of the classes in whose prediction we are interested, the situation is as shown in Table 9. In Fig. 4 we have graphically represented the situation at the first level. An important observation at this point is that the single instance does not always end in a therapy recommendation, but it may also end in a recommendation of further exams and investigations; in this experiment, such cases have been computed as negative cases for therapy recommendation, causing some imbalance between osteoporosis/severe osteoporosis cases and effectively recommended therapies. In this experiment we fixed $p = 10\%$; thus, only three data set have been considered for classifier extraction at the first level, namely $D_{osteop}, D_{vitDsup}$, and D_{calsup}. Under the same parameter, at the second level, whose situation is depicted in Fig. 5, the following data set have been considered: $D^{ale}_{osteop}, D^{den}_{osteop}, D^{ris}_{osteop}$, $D^{colec}_{vitDsup}, D^{calci}_{vitDsup}, D^{carb}_{calsup}$, and D^{citr}_{calsup}.

4.2. Results

After running the entire system, having fixed $k = 20\%$ (so that, as explained, the 20% stratified most recent records of each data set are used for testing purposes) the results are as in Table 10. For each problem and its corresponding test, we reported the following values: *accuracy*, that is, the rate of corrected classification, *sensitivity*, that is, the rate of true positives, *specificity*, that is, the rate of true negatives, *positive predicted value*, that is, the inverse of the false discovery rate, the *negative predicted value*, that is, the inverse of the false omission rate, and the F_1 *score*, that is, the harmonic mean of sensitivity and positive predicted value. Let us focus, first, on the behaviour of Γ_{osteop}. Out of 2048 instances, in 447 a osteoprotective therapy has been prescribed; our system is able to correctly predict if that is the case for a new instance in 86% of the cases. If, in fact, a osteoprotective therapy should be recommended, the system returns the correct suggestion in the 55% of the cases, while if the therapy should not be recommended, the system gives a correct prediction in 95% of the cases. In the case of predicting if a patient needs calcium supplementation (580 positive instances), our system gives a correct prediction in 83% of the cases, which drops to 47% in the positive cases and raises to 97% in the negative ones. Finally, in the case of supplementation of vitamin D (1125 positive cases), the rate of overall correct prediction is 75%, which is 74% and 76% in the positive and the negative cases, respectively. As it turns out, predicting the correct drug or supplement is quite more difficult at least in some cases, because the data set are still unbalanced even after our initial screening. Local evaluation is shown in Table 11, where we have displayed, for each type, the number of rules of each classifier, and their distribution among the four types; types I and III (which include only reliable rules) are the most common ones, indicating that our approach

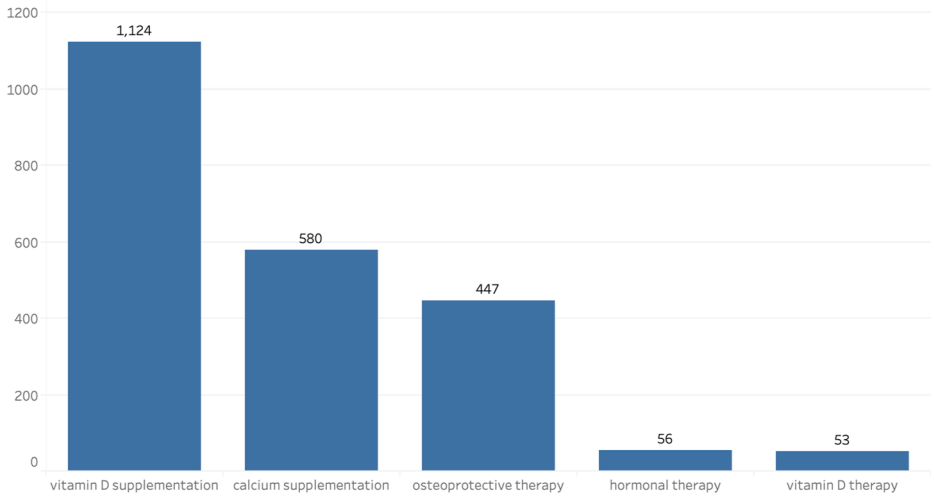

Fig. 4. Distribution at the first level.

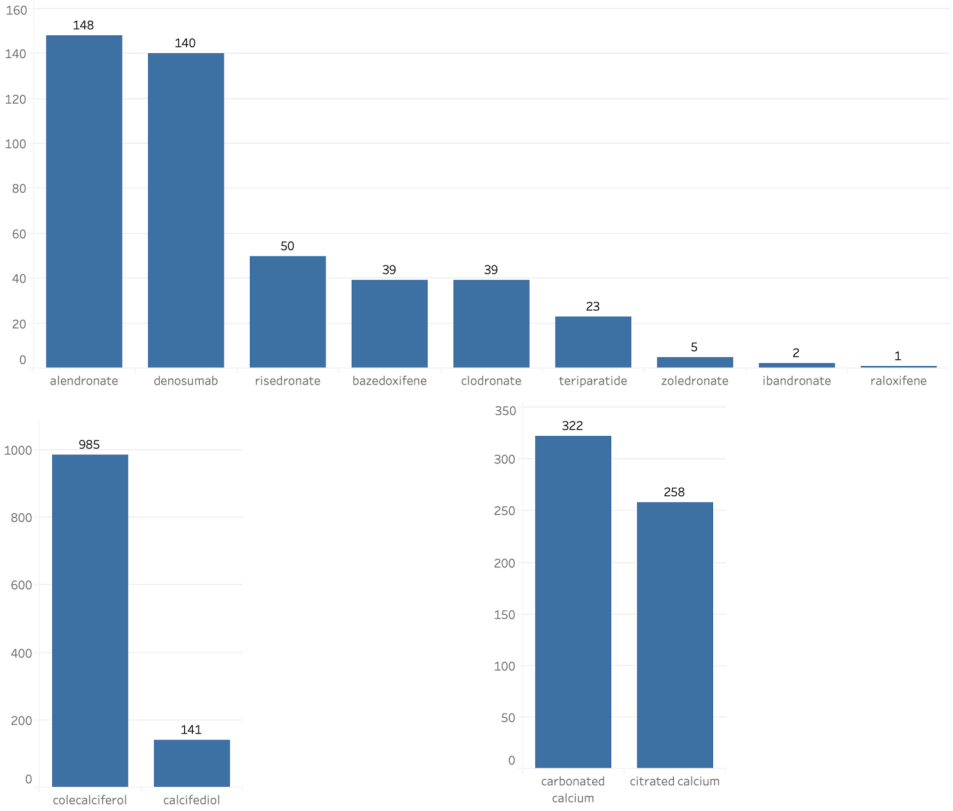

Fig. 5. Distribution at the second level.

Table 10
Results of the experiment: global evaluation. Top side: first level classifiers. Bottom side: second level classifiers.

	classif.	acc.	F_1	sens.	spec.	ppv.	npv.
1^{st} level	Γ_{osteop}	0.86	0.63	0.55	0.95	0.74	0.88
	$\Gamma_{vitDsup}$	0.75	0.76	0.74	0.76	0.79	0.71
	Γ_{calsup}	0.83	0.61	0.47	0.97	0.85	0.82
2^{nd} level	Γ^{ale}_{osteop}	0.72	0.39	0.27	0.95	0.73	0.72
	Γ^{den}_{osteop}	0.71	0.35	0.25	0.92	0.58	0.73
	Γ^{ris}_{osteop}	0.83	0.12	0.10	0.92	0.14	0.89
	$\Gamma^{calci}_{vitDsup}$	0.93	0.65	0.54	0.98	0.83	0.94
	$\Gamma^{colec}_{vitDsup}$	0.92	0.96	0.98	0.54	0.94	0.79
	Γ^{citr}_{calsup}	0.62	0.52	0.46	0.75	0.60	0.63
	Γ^{carb}_{calsup}	0.62	0.68	0.75	0.46	0.63	0.60

Table 11
Results of the experiment: local evaluation. Top side: first level classifiers. Bottom side: second level classifiers.

	classif.	#	# I	# II	# III	# IV
1^{st} level	Γ_{osteop}	7	1	0	5	1
	$\Gamma_{vitDsup}$	7	2	2	3	0
	Γ_{calsup}	4	1	0	2	1
2^{nd} level	Γ^{ale}_{osteop}	2	0	1	1	0
	Γ^{den}_{osteop}	4	0	1	1	2
	Γ^{ris}_{osteop}	2	1	0	0	1
	$\Gamma^{calci}_{vitDsup}$	5	1	0	2	2
	$\Gamma^{colec}_{vitDsup}$	5	1	0	2	2
	Γ^{citr}_{calsup}	3	0	2	1	0
	Γ^{carb}_{calsup}	3	0	2	1	0

is relatively stable.

Further considerations should be made with respect to the cases in which our approach showed a lower accuracy. First of all, the worst result in the entire panel of tests concerns the prediction of *risedronate*, for which the number of cases is only 50; ignoring the positive predicted value of 14% in this particular situation, the next worst performance (predicting *denosumab*) is 58%. In our experience, *risedronate* is usually recommended for a monthly dosing, and it includes the evaluation of possible gastric diseases; at the moment of the test, such cases were probably under-represented in our data base, causing the low performances of the system. However, as we have explained, our system is constantly improving itself as more patients are treated and recorded, so it could be the case that in a future evaluation better rules will be learned, and this problem will be solved. Now, consider the entire situation at the first level, and in particular, consider the subset of patients that are common to all test sets at that level; by combining all predictions for each patients, we can obtain a *confusion matrix* that allows us to further evaluate the performances of the predictions. Three problems have been considered at the first level: predicting *osteoprotective therapy*, *vitamin D therapy*, and *calcium supplementation*. This means that, per patient, there are eight possible recommendations. Table 12 shows the corresponding confusion matrix, on the left–right diagonal of which the correct predictions are shown. By observing the non-diagonal values, we can, in a sense, evaluate the seriousness of the mistake during the

Table 12
Confusion matrix for therapy prediction recommendation.

actual \ prediction	*osteop vitDsup calsup*	*osteop vitDsup*	*osteop calsup*	*vitDsup calsup*	*osteop*	*vitDsup*	*calsup*	∅
osteop vitDsup calsup	31	7	2	3	1	16	1	2
osteop vitDsup	1	3	0	1	0	10	0	0
osteop calsup	0	0	0	0	0	0	0	0
vitDsup calsup	4	0	0	4	0	12	0	3
osteop	0	0	0	0	0	0	0	0
vitDsup	1	0	0	1	1	61	0	49
calsup	0	0	0	0	0	0	0	1
∅	6	3	0	0	1	32	0	124

prediction. As it turns out, out of 373 patients, 223 have had a correct prediction as a whole. Moreover, the two single major incorrect ones were *vitamin D supplementation* versus *no treatment*; summing both directions, the total cases were 81. In other words, 54% of patients with some mistake in the prediction at the first level needed supplementation of vitamin D but the system suggested no supplementation, or the other way around. Another source of error is the suggestion of calcium supplementation in addition to vitamin D supplementation. In 13 cases, patients who have been recommended a supplementation of both calcium and vitamin D have received a prediction with vitamin D supplementation only, or the other way around. In conclusion, at the first level, we can say that 62% of wrongly predicted patients have received a prediction which was *not too far* from (our) ground truth. Unfortunately, at the second level, the number of patients that are common to all our test sets is not enough for such an analysis to be statistical significant.

5. Conclusions

Postmenopausal osteoporosis is a chronic, progressive condition characterized by reduced bone mass and impairment of bone architecture, leading to an increased risk of fragility fractures. Osteoporotic-related fractures decrease quality of life and are associated with high morbidity, mortality, and economic burden [3]. Therefore, it is mandatory upon the clinical practitioners to improve the tools that allow to identify patients at high risk of fracture and therapy based on the estimated risk. In this paper we have described the intelligent module of the CDSS that is currently in use at the Research Center for the Study of Menopause and Osteoporosis at the University of Ferrara (Italy). Our module is based on machine learning techniques, and in particular on the Ripper algorithm for deterministic rule-based classifier extraction, and it allows a physician to receive suggestions for treatment recommendation for osteoporotic patients. We adapted the algorithm to our purpose, integrated it into our proprietary CDSS, and used it in a self-adapting system that is able to update the rules on which recommendations are made as new patients populate the database. Then, based on two years of data, we have run a complete experiment to assess and report the reliability of the predictions, with encouraging results in most cases. Taking into account the amount of data that we collect for every patient, we believe that other problems can be approached and solved in similar ways; among them, we mention: *(i)* the problem of predicting the expected bone mass density for a patient based on his/her anamnesis; *(ii)* the problem of suggesting the optimal date for the next visit, and *(iii)* the problem of monitoring, even in an intelligent fashion, the T-score improvement in treated patients, and the occurrence of new fractures, as suggested in [26].

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] S. Adami, E. Romagnoli, V. Carnevale, A. Scillitani, A. Giusti, M. Rossini, D. Gatti, R. Nuti, S. Minisola, Guidelines on prevention and treatment of vitamin d deficiency, Reumatismo 63 (3) (2011) 129–147.

[2] G. Bonaccorsi, E. Fila, C. Cervellati, A. Romani, M. Giganti, M. Rossini, P. Greco, L. Massari, Assessment of fracture risk in a population of postmenopausal italian women: A comparison of two different tools, Calcif. Tissue Int. 97 (1) (2015) 50–57.

[3] F. Borgström, L. Karlsson, G. Ortsäter, N. Norton, P. Halbout, C. Cooper, M. Lorentzon, E.V. McCloskey, N.C. Harvey, M.K. Javaid, J.A. Kanis, Fragility fractures in europe: burden, management and opportunities, Arch. Osteoporos. 15 (59) (2020) 1–21.

[4] D.W. Bates, G.J. Kuperman, S. Wang, T. Gandhi, A. Kittler, L. Volk, Ten commandments for effective clinical decision support: making the practice of evidence-based medicine a reality, J. Am. Med. Inform. Assoc. 10 (6) (2003) 523–530.

[5] G. Bonaccorsi, C. Messina, C. Cervellati, E. Maietti, M. Medini, M. Rossini, L. Massari, P. Greco, Fracture risk assessment in postmenopausal women with diabetes: comparison between defra and frax tools, Gynec. End. 34 (5) (2017) 404–408.

[6] D.W. Bates, M. Cohen, L.L. Leape, J.M. Overhage, M.M. Shabot, T. Sheridan, Reducing the frequency of errors in medicine using information technology, J. Am. Med. Inform. Assoc. 8 (2001) 299–308.

[7] A.S. Cruz, H.C. Lins, R.V.A. Medeiros, J.M.F. Filho, S.G. da Silva SG, Artificial intelligence on the identification of risk groups for osteoporosis, a general review, Biomed. Eng. Online 17 (1) (2018) 1–12.

[8] J.S. Chiu, Y.C. Li, F.C. Yu, Y.F. Wang, Applying an artificial neural network to predict osteoporosis in the elderly, Stud. Health Technol. Inform. 124 (2006) 609–614.

[9] W.W. Cohen, Fast effective rule induction, in: In Proc. of the 20th International Conference on Machine Learning, 1995, pp. 115–123.

[10] C. Cipriani, J. Pepe, F. Bertoldo, G. Bianchi, F.P. Cantatore, A. Corrado, M. Di Stefano, B. Frediani, D. Gatti, A. Giustina, T. Porcelli, G. Isaia, M. Rossini, L. Nieddu, S. Minisola, G. Girasole, M. Pedrazzoni, The epidemiology of osteoporosis in italian postmenopausal women according to the National Bone Health Alliance (NBHA) diagnostic criteria: a multicenter cohort study, J. Endocrinol. Invest. 41 (2018) 431–438.

[11] P.M. Camacho, S.M. Petak, N. Binkley, D.L. Diab, L.S. Eldeiry, A. Farooki, S. T. Harris, D.L. Hurley, J. Kelly, E.M. Lewiecki, R. Pessah-Pollack, M. McClung, S. J. Wimalawansa, N.B. Watts, American association of clinical endocrinologists/ american college of endocrinology practice guideline for the diagnosis and treatment of postmenopausal osteoporosis - 2020 update, Endocr. Pract. 26 (5) (2020) 564–570.

[12] L. Degli Esposti, A. Girardi, S. Saragoni, S. Sella, M. Andretta, M. Rossini, S. Giannini, Use of antiosteoporotic drugs and calcium/vitamin D in patients with fragility fractures: impact on re-fracture and mortality risk, Endocr. 64 (2) (2018) 367–377.

[13] B. Halldorsson, A. Björnsson, H. Gudmundsson, E. Birgisson, B. Ludviksson, B. Guobjörnsson, A clinical decision support system for the diagnosis, fracture risks and treatment of osteoporosis, Comput. Math. Methods Med. 1–7 (2015) 2015.

[14] K. Hind, G. Slater, B. Oldroyd, M. Lees, S. Thurlow, M. Barlow, J. Shepherd, Interpretation of dual-energy x-ray absorptiometry-derived body composition change in athletes: A review and recommendations for best practice, J. Clin. Densitom. 21 (3) (2018) 429–443.
[15] J.A. Kanis, F. Borgstrom, C. De Laet, H. Johansson, O. Johnell, B. Jonsson, A. Oden, N. Zethraeus, B. Pfleger, N. Khaltaev, Assessment of fracture risk, Osteop. Int. 16 (6) (2005) 581–589.
[16] J.A. Kanis, C. Cooper, R. Rizzoli, J.Y. Reginster, European guidance for the diagnosis and management of osteoporosis in postmenopausal women, Osteoporos. Int. 30 (1) (2019) 3–44.
[17] J.A. Kanis, N.C. Harvey, E. McCloskey, O. Bruyere, N. Veronese, M. Lorentzon, C. Cooper, R. Rizzoli, G. Adib, N. Al-Daghri, C. Campusano, M. Chandran, B. Dawson-Hughes, K. Javaid, F. Jiwa, Johansson, J.K. Lee, E. Liu, D. Messina, O. Mkinsi, D. Pinto, D. Prieto-Alhambra, K Saag, W. Xia, L. Zakraoui, and J. Reginster. Algorithm for the management of patients at low, high and very high risk of osteoporotic fractures. Osteoporos. Int., 31(1), 2020.
[18] J.A. Kanis, A. Oden, H. Johansson, F. Borgstrom, F. Borgström, O. Stroöm, E. McCloskey, FRAX and its applications to clinical practice, Bone 44 (5) (2009) 734–743.
[19] J. LaFleur, C. McAdam-Marx, C.V. Asche, S. Alder, X. Sheng, J. Nebeker, D. I. Brixner, S. Silverman, Clinical risk factors for fracture among postmenopausal patients at risk for fracture: A historical cohort study using electronic medical record data, J. Bone Min. Metab. 29 (2) (2011) 193–200.
[20] J. LaFleur, R.E. Nelson, Y. Yao, R. Adler, J. Nebeker, Validated risk rule using computerized data to identify males at high risk for fracture, Osteop. Int. 23 (3) (2012) 1017–1027.
[21] A. Marques, R.J. Ferreira, E. Santos, E. Loza, L. Carmona, J.A. da Silva, The accuracy of osteoporotic fracture risk prediction tools: a systematic review and meta-analysis, Ann. Rheum. Dis. 74 (11) (2015) 1958–1967.
[22] W. Moudani, A. Shahin, F. Chakik, D. Rajab, Intelligent decision support system for osteoporosis prediction, Int. J. Intell. Inf. Technol. 8 (1) (2012) 26–45.
[23] R. Nuti, M.L. Brandi, G. Checchia, O. Di Munno, L. Dominguez, P. Falaschi, C. E. Fiore, G. Iolascon, S. Maggi, R. Michieli, S. Migliaccio, S. Minisola, M. Rossini, G. Sessa, U. Tarantino, A. Toselli, G.C. Isaia, Guidelines for the management of osteoporosis and fragility fractures, Intern. Emerg. Med. 14 (1) (2019) 85–102.
[24] C. Reshmalakshmi and M. Sasikumar. Fuzzy inference system for osteoporosis detection. In Proc. of the 5th IEEE Global Humanitarian Technology Conference, pages 675–681. IEEE, 2016.
[25] R. Sutton, D. Pincock, D. Baumgart, D. Sadowski, R. Fedorak, K. Kroeker, An overview of clinical decision support systems: benefits, risks, and strategies for success, Digital Medicine (2020) 3–17.
[26] T. Thomas, E. Casado, P. Geusens, W.F. Lems, J. Timoshanko, D. Taylor, L. C. Hofbauer, Is a treat-to-target strategy in osteoporosis applicable in clinical practice? consensus among a panel of European experts, Osteoporos. Int. 31 (1) (2020) 2303–2311.
[27] S. Unni, K. Gunning K, J. Curtis, Y. Yao, and J. LaFleur. An evaluation of clinical risk factors for estimating fracture risk in postmenopausal osteoporosis using an electronic medical record database. Osteop. Int., 26(2):581–587, 2015.
[28] W. Wang and S. Rea. Intelligent ensemble system aids osteoporosis early detection. In Proc. of the 6th World Scientific and Engineering Academy and Society International Conference on Evolutionary Computing, pages 123–128. WSEAS, 2005.