

ORIGINAL RESEARCH

Open access

Interpretable Random Forest Model for Predicting Diagnostic Test Overuse in Outpatient Clinics Using Visit Type, Physician Ordering History, Patient Complexity, Prior Test Results, and Guideline-Based Appropriateness Indicators

Nguyen Thanh Huy^{1*}, Pham Quang Minh¹, Le Thi Bich²

Abstract

Diagnostic test overuse in outpatient settings contributes to wasteful spending and may expose patients to unnecessary downstream testing, anxiety, and treatment cascades. Common examples include low-value imaging, repetitive laboratory testing, and routine preoperative or annual tests without clear clinical indication. Existing approaches often rely on manual review, retrospective measurement, or rigid rule-based filters. These approaches may miss context-sensitive overuse shaped by visit type, clinician habit, patient complexity, and recent test history. This manuscript proposes an interpretable random forest model that predicts whether a diagnostic test order could represent overuse. The model integrates visit-level, physician-level, patient-level, prior-result, and guideline-based appropriateness features. The proposed model would use historical outpatient orders labelled through appropriateness criteria or expert review. SHAP-based explanations would be used to make each prediction interpretable to clinicians and quality leaders. Conceptually, the model would estimate an overuse probability and identify the relative contribution of physician ordering history, patient complexity, prior results, visit type, and guideline indicators. These explanations would support targeted feedback rather than opaque surveillance. An interpretable random forest model could function as a personalised overuse screening tool within clinical decision support. Its value would depend on transparent explanations, careful guideline alignment, and thoughtful integration into outpatient workflows.

Keywords Explainable AI, SHAP, Random forest, Diagnostic test overuse, Low-value care, Choosing Wisely

*Correspondence:

Nguyen Thanh Huy

huy.nguyen@gmail.com

¹ Department of Health Information Technology, Faculty of Medicine, Vietnam National University, Hanoi, Vietnam

² Department of Clinical Data Analytics and Systems, Faculty of Engineering, Can Tho University, Can Tho, Vietnam

Introduction

Diagnostic test overuse in outpatient care is a persistent quality problem because unnecessary tests can trigger downstream cascades, patient anxiety, false-positive findings, and avoidable spending. Low-value imaging for uncomplicated low back pain, routine preoperative testing, repetitive laboratory testing, and annual screening tests without appropriate indication illustrate how common clinical routines can become sources of waste. Studies of low-value

imaging, preoperative electrocardiography, and broad low-value service use show that overuse is not limited to isolated decisions but can become embedded in care pathways [1-5]. A predictive model for outpatient overuse should therefore focus not only on whether a test was ordered, but also on whether the order fits the clinical context in which it occurred.

The determinants of diagnostic overuse are multifactorial and include clinician habit, perceived patient expectations, defensive medicine, time pressure, and uncertainty in complex patients. Prior work suggests that physician decision patterns can be learned from electronic health record data and that clinician cohorts may exhibit stable ordering tendencies that influence future orders [6-8]. Patient complexity further complicates interpretation because multimorbidity, abnormal findings, and medication burden can make additional testing appear clinically prudent even when guideline support is weak [9, 10]. An explainable model must therefore distinguish between inappropriate repetition and appropriate vigilance.

Guideline-based appropriateness criteria, Choosing Wisely recommendations, and specialty society guidance provide useful reference points for identifying low-value care, but they are often underused at the moment of ordering. The Choosing Wisely campaign has shown value as a framework for reducing unnecessary services, yet awareness and implementation remain inconsistent across clinicians and care settings [11-13]. Electronic health record infrastructure and clinical decision support could operationalise such recommendations by embedding appropriateness indicators directly into predictive workflows [14-16]. Machine learning can support this process by combining guideline signals with patient, visit, and physician context rather than applying rules in isolation.

This article proposes an interpretable random forest model that could estimate the probability that a diagnostic test order represents overuse and explain the factors driving that estimate. Random forest models are well suited to structured outpatient data because they can capture nonlinear relationships and interactions between visit type, physician ordering history, patient complexity, prior results, and guideline flags. Explainability methods such as SHAP can translate model behaviour into global audit patterns and local order-level explanations, strengthening clinical credibility and acceptability. The intended contribution is a conceptual model for behavioural nudges and quality improvement rather than a performance claim or completed experiment.

Background

Defining and measuring diagnostic test overuse

Diagnostic test overuse can be defined as testing that is unlikely to improve patient outcomes because it is unsupported by clinical indication, recent results, or evidence-based recommendations. Gross overuse may be identifiable through simple rules, such as imaging for uncomplicated low back pain, whereas nuanced overuse depends on timing, symptoms, comorbidity, prior findings, and clinician reasoning. Retrospective measurement using claims or electronic records can identify patterns after care has occurred, but real-time decision support requires features available at the point of ordering [2, 4, 10, 14]. A useful XAI model should therefore treat overuse as a context-dependent probability rather than a binary accusation.

Physician ordering patterns and signature behaviours

Physician ordering behaviour often reflects learned routines, local norms, specialty training, and tolerance for diagnostic uncertainty. Machine learning studies of clinical order patterns show that prior clinician behaviour can be informative for predicting future ordering decisions, suggesting the presence of physician-level “signature” patterns [6-8]. These patterns may persist even after adjustment for patient characteristics, making physician ordering history an important feature for overuse prediction. However, because such features can be sensitive, explanations should be framed as opportunities for reflection and peer learning rather than individual blame.

Patient complexity and risk-adjusted appropriateness

Patient complexity can blur the distinction between low-value testing and clinically appropriate caution. Multimorbidity, polypharmacy, abnormal recent findings, and unclear symptom trajectories may increase diagnostic uncertainty and make repeat testing more defensible. Studies of electronic health record bias and low-value care measurement show that clinical data are shaped by care processes, patient selection, and prior utilisation, so risk adjustment is essential when assessing appropriateness [9, 10, 17]. A model that ignores patient complexity could incorrectly penalise clinicians caring for more complex populations.

Visit type and the rhythm of overuse

Visit type structures the rhythm of outpatient testing because annual physicals, urgent visits, preoperative evaluations, follow-up visits, and problem-focused encounters create different expectations for diagnostic activity. Preoperative visits and routine checkups may be especially prone to habitual testing when historical templates, perceived completeness, or institutional routines encourage orders without new indications. Evidence on preoperative electrocardiography and low-value imaging demonstrates how visit context can initiate cascades of downstream care [3, 5, 18]. Encoding visit type allows the model to distinguish clinically plausible testing from orders that appear driven primarily by visit routines.

Interpretable machine learning for quality-of-care audit

Interpretable machine learning is increasingly important in healthcare because clinical users need to understand why a model identifies a decision as risky, inappropriate, or potentially low value. Explainability methods have been used to make clinical predictions more transparent, including SHAP-based approaches that attribute predictions to individual features [19–22]. In quality-of-care audit, interpretability is especially important because opaque models may be viewed as punitive, biased, or misaligned with clinical nuance. A random forest with SHAP explanations could support both individual order review and clinic-level audit while preserving a clear link between predictions and actionable features.

Model Overview and Design Principles

High-level predictive framework

At the time a diagnostic test is ordered, or prospectively before the order is finalised, the proposed model would ingest structured information from the visit, patient record, ordering clinician profile, prior test history, and guideline indicator library. It would output an estimated probability that the order could represent overuse, using the probability as a decision-support signal rather than an automated denial mechanism. This framework is consistent with prior work showing that electronic health record data can support clinical order prediction and that such predictions can be embedded into clinical decision support workflows [6, 8, 16]. The emphasis is on surfacing context and uncertainty at the point of care.

Figure 1 illustrates the proposed interpretable random forest workflow linking outpatient ordering context, guideline-based appropriateness indicators, SHAP explanations, and governed clinician-facing decision support.

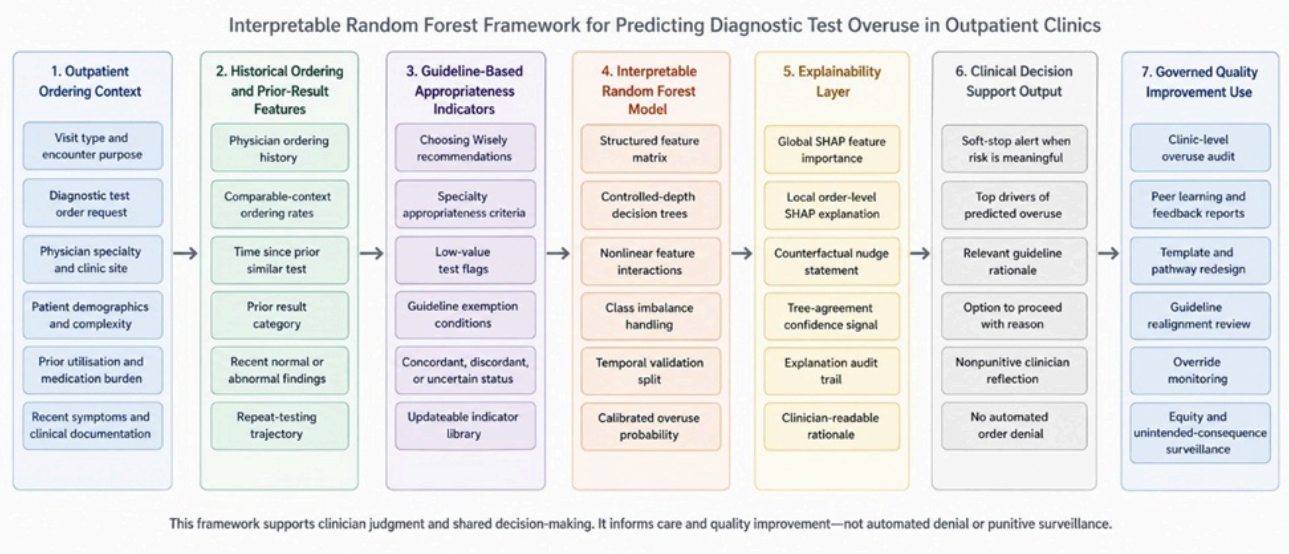


Figure 1. Interpretable random forest architecture for context-sensitive prediction and explanation of outpatient diagnostic test overuse.

Core input features

The core input features would include visit type, ordering physician history, clinician specialty, patient complexity, recent relevant test results, and guideline-based appropriateness indicators. Visit type would capture whether the encounter is new, follow-up, urgent, preventive, annual, or preoperative, while physician ordering history would summarize prior testing tendencies in comparable contexts. Patient complexity features would capture comorbidity burden, medication burden, and recent abnormal or normal results, and guideline flags would indicate whether a test conflicts with Choosing Wisely or other appropriateness criteria [11-13, 15]. These features are intended to let the model learn when a superficially low-value test may be justified by context.

Design principles

The model should be interpretable, guideline-aligned, clinically contestable, and adaptable as recommendations evolve. A tree-based ensemble offers a practical balance because it can capture feature interactions while remaining compatible with global and local explanation methods [19, 22, 23]. The system should provide physician-friendly explanations that identify the most influential features without overwhelming clinicians or implying that guidelines replace judgment. Its design should also support periodic updating as evidence changes and as local practice patterns shift.

Data Sources and Feature Engineering

Extracting visit, order, and physician data

Structured electronic health record data would provide the foundation for visit type, order characteristics, ordering physician identifier, specialty, clinic location, and patient demographics. Historical ordering rates could be calculated within comparable visit categories so that physician behaviour is interpreted relative to clinical context rather than as a crude total volume measure. Prior studies of order prediction and electronic health record-based quality measurement support the feasibility of using structured clinical data for modelling ordering behaviour, while also cautioning that EHR data reflect workflow and documentation practices [6, 7, 9, 14]. Feature definitions should therefore be transparent and reviewed with clinicians before deployment.

Table 1 presents the proposed feature architecture for distinguishing potentially low-value diagnostic testing from clinically justified testing in context-sensitive outpatient workflows.

Table 1. Conceptual Feature Architecture for Context-Sensitive Diagnostic Test Overuse Prediction

Feature domain	Example variables	Why it matters for overuse prediction	Interpretation risk	Governance requirement
Visit context	Preventive visit, urgent visit, follow-up, preoperative evaluation, annual examination	Distinguishes routine-driven ordering from clinically prompted testing	May label routine care as low value without symptom nuance	Define visit categories transparently and review with clinicians
Physician ordering history	Prior ordering frequency, comparable-context test rate, specialty-adjusted ordering pattern	Captures stable ordering tendencies and possible habit-driven overuse	Could be perceived as physician surveillance or blame	Use nonpunitive framing and risk-adjusted peer comparison
Patient complexity	Comorbidity burden, medication burden, recent utilisation, multimorbidity indicators	Prevents inappropriate flagging in clinically complex patients	Under-adjustment may penalise clinicians caring for sicker populations	Audit calibration across complexity strata
Prior test results	Time since last similar test, normal result, abnormal result, unresolved finding	Differentiates unnecessary repetition from clinically justified follow-up	EHR data may not capture full clinical reasoning	Encode result trajectory rather than simple repeat-test status
Guideline indicators	Choosing Wisely flag, specialty appropriateness rule, exemption condition	Anchors predictions in evidence-based low-value care logic	Guidelines may lag evidence or fail in complex cases	Maintain an updateable, clinically reviewed indicator library
Ordering environment	Clinic site, specialty, local pathway, order set or template	Identifies organisational routines that may drive overuse	Site effects may reflect documentation or workflow differences	Use aggregate audit for pathway redesign, not individual punishment
Model confidence	Tree agreement, prediction variability, calibration band	Communicates uncertainty and discourages deterministic interpretation	False confidence may encourage inappropriate deferral	Pair probability with uncertainty and override option

Constructing patient complexity and prior-result features

Patient complexity features would include comorbidity burden, active medication count, recent utilisation, and relevant prior abnormal findings. Prior-result features would represent the time since the last comparable test, the result category, and whether the result was normal, borderline, abnormal, or unresolved. This structure is important because a repeat test after a recent normal result may carry a different overuse implication than a repeat test after an abnormal or worsening result [9, 10, 17]. The model should therefore encode clinical trajectory rather than treating all repeat testing as low value.

Mapping guideline-based appropriateness indicators

Guideline-based appropriateness indicators would translate Choosing Wisely recommendations, specialty criteria, and evidence-based statements into structured binary or ordinal features. For example, a feature could indicate that imaging for uncomplicated low back pain is generally low value unless red flags are present, or that routine preoperative testing is discouraged in low-risk contexts. Prior research on Choosing Wisely implementation and low-value care measurement shows that guidelines can identify actionable opportunities, but their real-world use requires careful operationalisation [2, 4, 11–13]. These indicators should function as model inputs and interpretive anchors rather than rigid substitutes for clinical reasoning.

Interpretable Random Forest Architecture

Model structure and training regimen

The proposed model would use a random forest classifier composed of decision trees with controlled depth and prespecified feature governance. Historical outpatient orders could be labelled through guideline concordance, expert review, or hybrid adjudication, while recognising that labels may reflect uncertainty in appropriateness judgments. Random forests are suitable for structured clinical data because they can capture nonlinear relationships and interactions among physician history, patient complexity, prior results, and visit type [7, 8, 19, 22]. Training would be conceptualised as a quality-improvement modelling process rather than a claim of completed experimental performance.

Handling class imbalance and temporal trends

The frequency of labelled overuse may vary by test type, specialty, clinic, and guideline definition, so class imbalance should be handled through transparent weighting or sampling choices. Temporal splitting would be important to avoid future data leakage and to evaluate whether patterns learned from earlier ordering behaviour remain relevant as guidelines and practice norms change. Research on EHR-derived prediction and low-value care measurement highlights the need to account for changing clinical workflows, documentation practices, and secular trends [9, 14, 16, 17]. Model governance should therefore include periodic review of feature stability and label validity.

Output: overuse probability and confidence

For each diagnostic order, the model would return a calibrated overuse probability and an accompanying indication of prediction confidence based on agreement or variability across trees. This probability should be interpreted as a decision-support signal that invites reflection, not as a definitive judgment that the test is inappropriate. Prior work on clinical decision support and explainable AI suggests that predictive outputs are more likely to be accepted when paired with understandable explanations and when clinicians retain the ability to override recommendations [16, 18, 20, 21]. The model's output should therefore be embedded in a workflow that supports review, documentation, and learning.

Incorporating Guideline-Based Appropriateness Indicators

Indicator selection and translation

A guideline indicator library would be curated from Choosing Wisely recommendations, specialty society appropriateness criteria, and evidence-based recommendations for common outpatient tests. Each recommendation would be translated into structured features that identify whether the order appears concordant, potentially discordant, or clinically exempt based on available EHR context. This approach aligns with prior work showing that Choosing Wisely can define actionable low-value care targets, but that practical implementation requires translation into measurable clinical logic [11–13, 15]. The indicator library would require clinical governance so that each feature remains transparent, contestable, and updatable.

Feature interaction with physician and patient context

Guideline flags should not be interpreted in isolation because patient complexity, recent abnormal results, and visit context can modify appropriateness. A random forest could learn that a test flagged as low value in a routine visit may be more defensible in a patient with unresolved symptoms, multimorbidity, or concerning prior results. This is important because overuse measurement can be biased when claims or EHR features fail to capture clinical nuance [9, 10, 17]. The model should therefore combine guideline indicators with patient and physician context rather than simply reproducing rules.

Override logic and clinical nuance

The proposed model would not enforce guidelines or block orders automatically. Instead, it would estimate overuse risk and allow clinicians to override the alert when undocumented nuance, patient preference, or evolving clinical concern justifies testing. Clinical decision support research suggests that acceptability improves when tools are explainable, nonpunitive, and integrated into workflow rather than imposed as rigid restrictions [16, 18, 20, 21]. Override reasons could also become feedback signals for refining guideline logic and improving future explanations.

Periodic guideline re-alignment

Guideline-based features would need scheduled review because recommendations, evidence thresholds, and specialty norms may change over time. Periodic realignment would involve updating the indicator library, reviewing feature definitions, and checking whether model explanations still reflect current clinical guidance. This is consistent with concerns that low-value care definitions and EHR-based measurements can drift as practice patterns and documentation change [9, 12, 14, 17]. Recalibration should therefore be treated as part of ongoing clinical governance rather than a one-time technical task.

Explainability Methods for Overuse Predictions

Global feature importance for clinic-level audit

Global SHAP analysis could identify which features most strongly influence overuse predictions across clinics, specialties, or test categories. For example, clinic-level summaries might reveal that prior imaging frequency, visit type, recent normal results, or physician ordering history consistently drive predicted overuse risk. SHAP-based interpretation has been used to make healthcare predictions more transparent and can support quality audit when explanations are aggregated carefully [19–22]. These summaries should be de-identified or appropriately governed to avoid turning explainability into punitive surveillance.

Local explanation for the ordering clinician

Local explanations would show why a specific order was assigned a higher overuse probability at the moment of ordering. For example, an explanation might indicate that a recent normal result, a low-risk visit type, and a guideline discordance flag were the most influential features. Explainable clinical decision support is more likely to be trusted when clinicians can see the concrete factors behind a recommendation rather than only a score or alert label [20, 21, 23]. The explanation should be concise enough for point-of-care use while still allowing the clinician to inspect relevant evidence.

Counterfactual explanation for behavioural nudge

Counterfactual explanations could frame the alert as a reflective behavioural nudge rather than a judgment. A clinician might see that the overuse probability would be lower if the patient had no recent normal test, if the visit were problem-focused rather than routine, or if guideline exemption criteria were present. This approach is consistent with overuse reduction strategies that emphasise clinician engagement, feedback, and practical workflow design [15, 16, 18, 24]. The goal would be to encourage reconsideration while preserving clinical autonomy.

Transparent audit trail for administration

Administrative dashboards could use aggregated explanations to identify system-level drivers of overuse, such as routine preoperative testing pathways, annual physical templates, or specialty-specific ordering norms. These audit trails would focus on modifiable processes rather than individual blame, helping leaders redesign defaults, education, and feedback mechanisms. Prior studies of low-value care cascades and health-system-level variation show that overuse often reflects organisational routines as much as individual clinician choices [1, 5, 17]. Transparent audit trails would therefore support improvement planning while maintaining accountability.

Table 2 defines how model explanations should be translated into safe, nonpunitive, and governable decision-support functions across clinician, administrative, and quality-improvement users.

Table 2. Explainability and Governance Matrix for Clinician-Facing Overuse Decision Support

Explanation layer	Primary user	Output format	Intended function	Safety concern	Required safeguard
Global SHAP summary	Quality leaders and clinical governance teams	Ranked feature contributions by clinic, specialty, or test type	Identifies system-level drivers of predicted overuse	May become punitive performance monitoring	Aggregate reporting with explicit nonpunitive use policy
Local SHAP explanation	Ordering clinician	Top contributing factors for one order	Supports real-time reconsideration at the point of ordering	May oversimplify nuanced clinical judgment	Allow override with documented reason
Guideline rationale	Ordering clinician and reviewer	Relevant recommendation or appropriateness indicator	Connects model output to evidence-based guidance	Guideline may not fit patient-specific nuance	Display exemption logic and uncertainty category
Counterfactual nudge	Ordering clinician	“Risk would be lower if...” statement	Encourages reflection without blocking the order	Could be perceived as coercive	Phrase as decision support, not denial
Override audit	Governance team	Aggregated override reasons and alert outcomes	Detects alert fatigue, valid disagreement, or weak rules	Could incentivise superficial documentation	Review overrides qualitatively and refine indicators
Equity monitoring	Quality and safety committees	Calibration and flagging rates by patient subgroup	Detects inequitable overuse predictions or under-testing risk	Model may amplify EHR or access biases	Require subgroup performance review before and after deployment
Guideline realignment	Clinical governance group	Scheduled update log for indicator library	Keeps appropriateness logic current	Outdated rules may misclassify appropriate care	Establish periodic review and version control

Clinical Integration and Feedback Mechanisms

Seamless EHR decision support integration

The model would run within the EHR at the moment a diagnostic order is placed and return a soft-stop alert only when the estimated overuse risk is clinically meaningful. The alert would include a short explanation, the relevant guideline indicator, and an option to continue with a documented reason. Prior work on clinical decision support and order recommendation shows that timing, workflow fit, and interpretability are central to whether clinicians accept or ignore model-driven tools [8, 16, 18]. The system should therefore minimise interruption while making the reasoning clear.

Post-Visit Feedback and Peer Comparison

Post-visit feedback reports could summarise model-derived overuse patterns by test type, visit type, and guideline category. Peer comparison should be framed carefully, using department-level benchmarks and contextual adjustment so clinicians caring for complex patients are not unfairly judged. Evidence from Choosing Wisely and low-value care research suggests that awareness alone is often insufficient, but feedback linked to specific behaviours may support change [11–13]. The feedback mechanism should promote learning rather than competition or blame.

Evaluation Strategy

Predictive performance and calibration

A future evaluation should examine whether predicted overuse probabilities are well calibrated across physicians, clinics, test categories, and patient complexity strata. Performance assessment should include discrimination, calibration, and error analysis, but these results should be interpreted alongside clinical review because appropriateness labels can be uncertain. Prior EHR prediction and low-value care studies show that measurement validity depends on how labels, features, and clinical context are defined [6, 9, 10, 14]. Evaluation should therefore focus on whether the model supports safer, more consistent decision-making rather than only technical accuracy.

Explanation quality and clinician trust

Explanation quality should be evaluated through clinician assessment of fairness, understandability, usefulness, and actionability. Clinicians should be asked whether local explanations identify clinically relevant reasons, whether guideline indicators are credible, and whether alerts support reconsideration without undermining autonomy. Explainable AI literature emphasises that transparency is not only a technical property but also a relational feature affecting trust, responsibility, and adoption [20–23]. Override patterns should be reviewed to distinguish valid clinical disagreement from alert fatigue.

Impact on overuse rates

The clinical impact of the model should be evaluated through pragmatic implementation designs that compare ordering appropriateness before and after explainable decision support. Outcomes could include changes in low-value ordering, downstream cascades, clinician acceptance, and unintended consequences such as under-testing or documentation burden. Prior research on low-value imaging, preoperative testing, and overuse cascades shows that interventions must be assessed beyond the initial order because downstream effects are central to patient harm and waste [2–5]. Any impact evaluation should preserve clinician discretion and monitor equity across patient groups.

Limitations

Guideline limitations and expert disagreement

A central limitation is that appropriateness is sometimes subjective, and guidelines may lag behind evidence, conflict across societies, or fail to address complex clinical presentations. Labels based on guideline concordance can therefore inherit uncertainty and may not fully capture patient preference, evolving symptoms, or undocumented clinical reasoning.

Prior work on overuse measurement and EHR-derived quality assessment shows that low-value care definitions can be incomplete when applied mechanically [9, 10, 12, 14]. The model should therefore communicate uncertainty and allow clinically justified disagreement.

Physician gaming and alert fatigue

Clinicians may learn to bypass alerts, document superficial justifications, or modify ordering workflows in ways that reduce the tool's value. Alert fatigue is especially concerning if the system interrupts too frequently, provides vague explanations, or appears misaligned with clinical reality. Research on clinical decision support and explainable AI suggests that tools must balance transparency, workflow fit, and cognitive burden to remain useful [16, 18, 20, 21]. Continuous monitoring of override reasons and clinician feedback would be needed to prevent the model from becoming another ignored EHR prompt.

Conclusion

An interpretable random forest model could provide a practical framework for predicting diagnostic test overuse in outpatient clinics. By combining visit type, physician ordering history, patient complexity, prior test results, and guideline-based appropriateness indicators, the model would move beyond simple rule-based detection. Its purpose would be to support reflective ordering decisions rather than replace clinical judgment.

The model's key strength is its integration of evidence-based appropriateness logic with real-world clinical context. SHAP explanations would allow both local order-level interpretation and global quality audit. This makes the approach suitable for point-of-care nudges, peer feedback, and health-system learning.

Important challenges remain. Guideline definitions must be kept current, clinician pushback must be avoided, and explanations must be clear enough to support trust without increasing cognitive burden. The model must also be evaluated for unintended consequences, including inequitable flagging or under-testing in complex patients.

Future work should prioritise pragmatic clinical trials and shared benchmarks for measuring outpatient diagnostic overuse. Health systems should develop transparent governance processes for guideline translation, model review, and clinician feedback. With careful implementation, explainable AI could help reduce low-value testing while preserving the nuance of outpatient care.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 22 Dec 2021

Revised: 29 Jan 2022

Accepted: 29 Mar 2022

Published online: 20 July 2022

Rights and permissions

Open Access The author(s) retain copyright. This article is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License. It may be shared and adapted for non-commercial purposes with appropriate attribution, an indication of changes, and distribution of adaptations under the same license. Third-party material may be subject to separate terms identified in its credit line. View the license at <https://creativecommons.org/licenses/by-nc-sa/4.0/>.

References

Radomski TR, Zhao X, Lovelace EZ, Sileanu FE, Rose L, Schwartz AL, et al. Use and cost of low-value health services delivered or paid for by the Veterans Health Administration. *JAMA Intern Med.* 2022;182(8):832-9.
<https://doi.org/10.1001/jamainternmed.2022.2285>.

Belavy DL, Tagliaferri SD, Buntine P, Saueressig T, Samanna C, McGuckian T, et al. Reducing low-value imaging for low back pain: systematic review with meta-analysis. *J Orthop Sports Phys Ther.* 2022;52(4):175-91.
<https://doi.org/10.2519/jospt.2022.10669>.

Al-Hihi E, Gibson C, Lee J, Mount RR, Irani N, McGowan C. Improving appropriate imaging for non-specific low back pain. *BMJ Open Qual.* 2022;11(1):e001731.
<https://doi.org/10.1136/bmj-2021-001731>.

Hall AM, Aubrey-Bassler K, Thorne B, Maher CG. Do not routinely offer imaging for uncomplicated low back pain. *BMJ.* 2021;372:n291.
<https://doi.org/10.1136/bmj.n291>.

Ganguli I, Lupo C, Mainor AJ, Raymond S, Wang Q, Orav EJ, et al. Prevalence and cost of care cascades after low-value preoperative electrocardiogram for cataract surgery in fee-for-service Medicare beneficiaries. *JAMA Intern Med.* 2019;179(9):1211-9.
<https://doi.org/10.1001/jamainternmed.2019.1739>.

Hunter-Zinck HS, Peck JS, Strout TD, Gaehde SA. Predicting emergency department orders with multilabel machine learning techniques and simulating effects on length of stay. *J Am Med Inform Assoc.* 2019;26(12):1427-36.

Wang JK, Hom J, Balasubramanian S, Schuler A, Shah NH, Goldstein MK, et al. An evaluation of clinical order patterns machine-learned from clinician cohorts stratified by patient mortality outcomes. *J Biomed Inform.* 2018;86:109-19.
<https://doi.org/10.1016/j.jbi.2018.09.004>.

Kumar A, Aikens RC, Hom J, Shieh L, Chiang J, Morales D, et al. OrderRex clinical user testing: a randomized trial of recommender system decision support on simulated cases. *J Am Med Inform Assoc.* 2020;27(12):1850-9.

Agniel D, Kohane IS, Weber GM. Biases in electronic health record data due to processes within the healthcare system: retrospective observational study. *BMJ.* 2018;361:k1479.
<https://doi.org/10.1136/bmj.k1479>.

Oakes AH, Chang HY, Segal JB. Systemic overuse of health care in a commercially insured US population, 2010-2015. *BMC Health Serv Res.* 2019;19(1):280.
<https://doi.org/10.1186/s12913-019-4085-0>.

Colla CH, Mainor AJ. Choosing Wisely campaign: valuable for providers who knew about it, but awareness remained constant, 2014-17. *Health Aff (Millwood)*. 2017;36(11):2005-11.
<https://doi.org/10.1377/hlthaff.2017.0845>.

Mafi JN, Parchman M. Low-value care: an intractable global problem with no quick fix. *BMJ Qual Saf*. 2018;27(5):333-6.
<https://doi.org/10.1136/bmjqs-2017-007477>.

Levinson W, Born K, Wolfson D. Choosing wisely campaigns: a work in progress. *JAMA*. 2018;319(19):1975-6.
<https://doi.org/10.1001/jama.2018.2202>.

Rumball-Smith J, Shekelle PG, Bates DW. Using the electronic health record to understand and minimize overuse. *JAMA*. 2017;317(3):257-8.
<https://doi.org/10.1001/jama.2016.18693>.

Mehta N, Born K, Fine B. How artificial intelligence can help us 'Choose Wisely'. *Bioelectron Med*. 2021;7(1):5.
<https://doi.org/10.1186/s42234-021-00068-3>.

Poly TN, Islam MM, Muhtar MS, Yang HC, Nguyen PA, Li YC. Machine learning approach to reduce alert fatigue using a disease medication-related clinical decision support system: model development and validation. *JMIR Med Inform*. 2020;8(11):e19489.
<https://doi.org/10.2196/19489>.

Ganguli I, Morden NE, Yang CW, Crawford M, Colla CH. Low-value care at the actionable level of individual health systems. *JAMA Intern Med*. 2021;181(11):1490-500.
<https://doi.org/10.1001/jamainternmed.2021.5531>.

Min A, Chan VW, Aristizabal R, Peramaki ER, Agulnik DB, Strydom N, et al. Clinical decision support decreases volume of imaging for low back pain in an urban emergency department. *J Am Coll Radiol*. 2017;14(7):889-97.
<https://doi.org/10.1016/j.jacr.2017.01.046>.

Lundberg SM, Nair B, Vavilala MS, Horibe M, Eisses MJ, Adams T, et al. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nat Biomed Eng*. 2018;2(10):749-60.
<https://doi.org/10.1038/s41551-018-0304-0>.

Amann J, Blasimme A, Vayena E, Frey D, Madai VI, Precise4Q Consortium. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. *BMC Med Inform Decis Mak*. 2020;20(1):310.
<https://doi.org/10.1186/s12911-020-01332-6>.

Amann J, Vetter D, Blomberg SN, Christensen HC, Coffee M, Gerke S, et al. To explain or not to explain?—Artificial intelligence explainability in clinical decision support systems. *PLOS Digit Health*. 2022;1(2):e0000016.
<https://doi.org/10.1371/journal.pdig.0000016>.

Loh HW, Ooi CP, Seoni S, Barua PD, Molinari F, Acharya UR. Application of explainable artificial intelligence for healthcare: a systematic review of the last decade (2011-2022). *Comput Methods Programs Biomed*. 2022;226:107161.
<https://doi.org/10.1016/j.cmpb.2022.107161>.

Ahmad MA, Eckert C, Teredesai A. Interpretable machine learning in healthcare. In: *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*; 2018; Washington, DC, USA. New York: ACM; 2018. p. 559-60.
<https://doi.org/10.1145/3233547.3233667>.

van Dulmen SA, Naaktgeboren CA, Heus P, Verkerk EW, Weenink J, Kool RB, et al. Barriers and facilitators to reduce low-value care: a qualitative evidence synthesis. *BMJ Open*. 2020;10(10):e040025.
<https://doi.org/10.1136/bmjopen-2020-040025>.