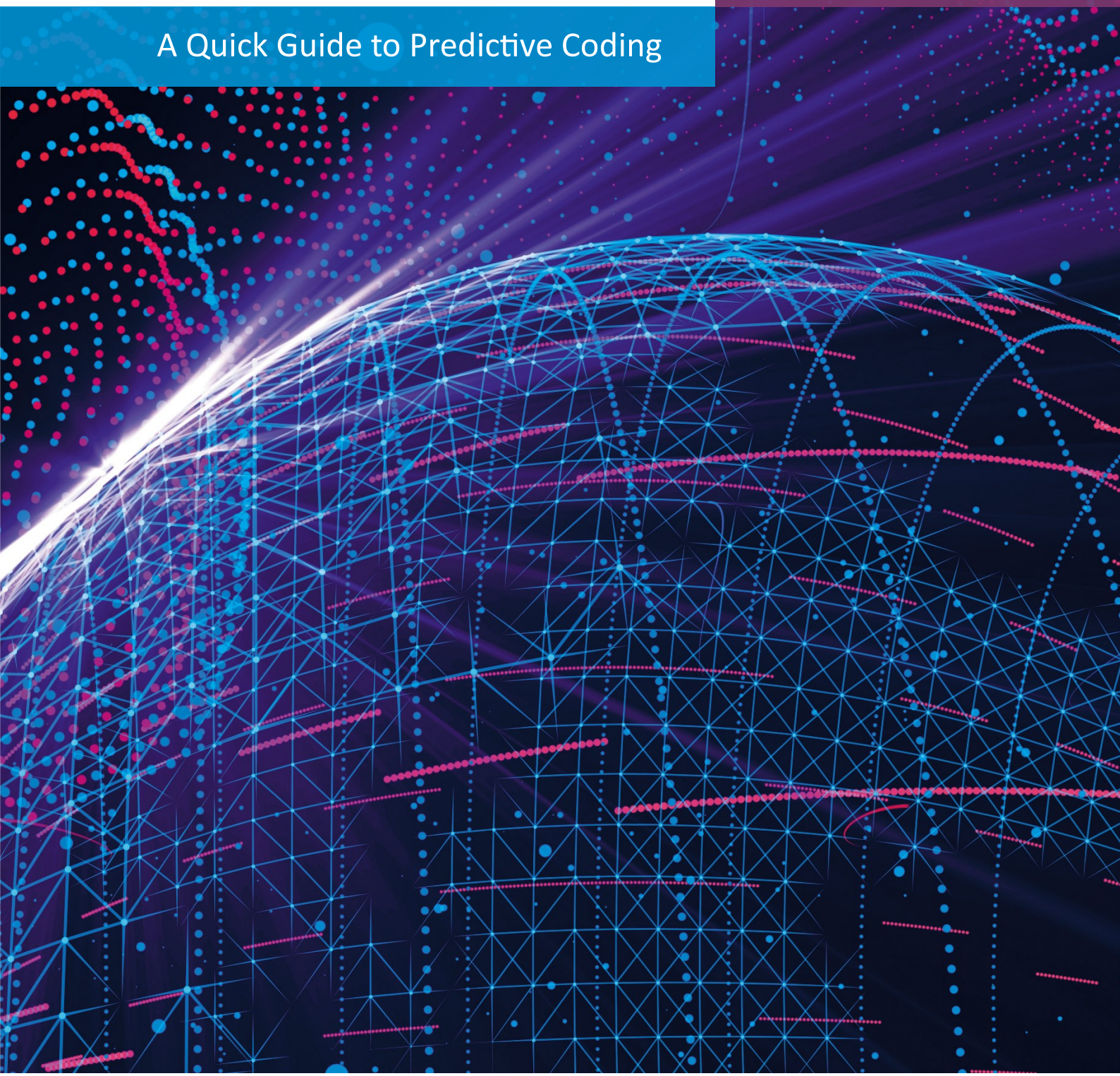
fieldfisher

Technology Assisted Review (TAR)

March 2022

A Quick Guide to Predictive Coding



Predictive Coding (TAR) in eDisclosure

The volume of data (predominantly electronic data) continues to increase exponentially in all business landscapes. This has the knock-on effect of increasing datasets in large-scale and complex litigation. Technology assisted review (TAR) and one of its components, predictive coding, has become an essential legal aid for efficient and cost-effective eDisclosure, particularly when bearing in mind that disclosure is typically the most costly component of an overall case budget.

Since the English High Court approved the use of predictive coding for eDisclosure/eDiscovery in *Pyrrho*[1] in 2016, practitioners in the UK have seen a huge increase in TAR. The Court noted in *Pyrrho* that using predictive coding would result in greater review consistency across disclosure-sets than if the full document corpus were left to traditional linear review.

What is predictive coding?

Predictive coding is a module of TAR (also known as *computer-assisted review* (CAR)). It is used to train a computer algorithm to recognise relevant documents within an eDisclosure dataset, thereby eliminating the requirement to review all documents in scope. Predictive coding uses expert reviewers to train documents (meaning the documents are coded or tagged), thus allowing the algorithm to identify documents that are most likely **relevant**. A human reviewer then reviews these relevant documents. This creates an iterative cycle of prediction and analysis, which is run over other documents within the review corpus to predict accurate coding outcomes. Documents that have been trained by TAR to not respond to relevant criteria can be deemed not relevant, thereby excluding them from review. In practice, a sample of these *not relevant* documents will be evaluated to "*validate*" the algorithm's scoring accuracy.

Predictive coding and exact methodologies used in the process will vary from one eDisclosure provider to the next. Versions of TAR have an associated number: TAR 1.0, TAR 2.0 and, more recently, TAR 3.0.

Each TAR version comes with its merits, therefore a comprehensive understanding of the underlying process and technology is required in selecting the correct TAR model suited to each particular eDisclosure exercise.

[1] *Pyrrho Investments Limited & Anr v MWB Property Limited and Others [2016] EWHC 256 (Ch)*

TAR 1.0

The original TAR (TAR	TAR 1.0 involves a training phase, followed by a review phase. A control set is used to determine the optimal point when the training should progress to the review phase.
Benefits	TAR 1.0 **outperforms** traditional linear review as **responsive documents are prioritised** in the review process. The control set used to train the TAR model allows for an estimation of the number of responsive documents expected to be found. When the full document corpus is scored, the number of documents requiring review can be **practicably assessed**, allowing teams to **efficiently plan workflow**.
Criticisms	TAR 1.0 trains the documents once, which does **not allow the system to adjust where further information becomes available** during the course of the review. Because one-time training relies on early coding of a training-set, the **possibility of bias** in the predictive model could cause potential concerns about a production's sufficiency.

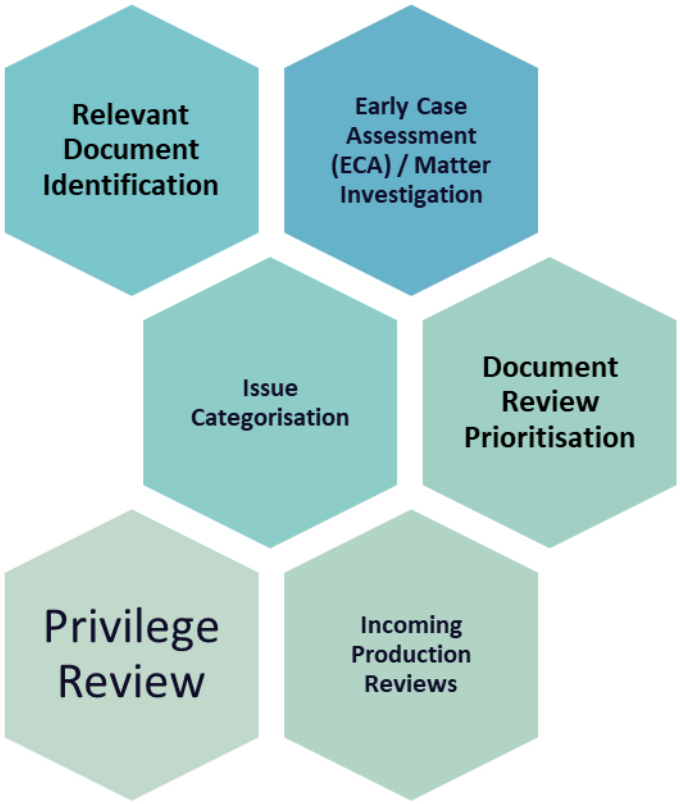
TAR 2.0 (Continuous Active Learning / CAL)

Continuous Active Learning – this is more precise (and therefore, more economical) review process.	TAR 2.0/CAL does not require a control set. It relies on human reviewers to make coding decisions – there is no separation between training and review as the system is continuously learning. The algorithm scores the coding manually applied by reviewers and is continuously learning so that it can **constantly prioritise unreviewed potentially relevant documents** in the review queue. TAR 2.0/CAL tends to be very efficient even when prevalence is low.
Benefits	A document review exercise can **begin immediately**. Linear review is improved upon as a **high proportion of responsive documents are presented to reviewers early on**.
Criticisms	Due to the continuous evolution of the model, predicting the overall volume of document population that will require review is difficult to gauge. The constant promotion of high-ranked documents introduces the **risk of showing reviewers similar documents repetitively, instead of promoting diverse potentially responsive content early in the review**. A risk is that surprises may occur later in the review if unexpected responsive content is lower-ranked.

TAR 3.0 (CAL + Clustering)

TAR 3.0 uses CAL on cluster sets. A clustering algorithm can be used to initially group documents, over which CAL is run. 1.0)	TAR 3.0 does not require a control set. TAR 3.0 uses a high-quality conceptual clustering algorithm, which forms targeted fixed-size cluster-groups in concept space. TAR 3.0 applies TAR 2.0/CAL methodology to the focused clusters in isolation - this allows for the review of a diverse set of potentially relevant documents early in the review. When relevant clusters have been reviewed (and no more clusters can be found), the reviewed clusters train the documents to make predictions over the entire document corpus.
Benefits	A document review exercise can begin immediately. The system is **continuously updated to reflect ongoing reviewer-based coding decisions**. The early focus on cluster-groups **creates a diverse set of reviewable documents, which minimises surprises later on** and which allows for **responsive documents to be identified at an early stage.** When the prevalence of responsive content is low, completion of the review can be determined.
Criticisms	Case managers and senior stakeholders need to consider whether documents marked as relevant, that have not been subject to review, should be produced or whether these documents should have secondary privilege/high-risk key words run over them to assess their content.

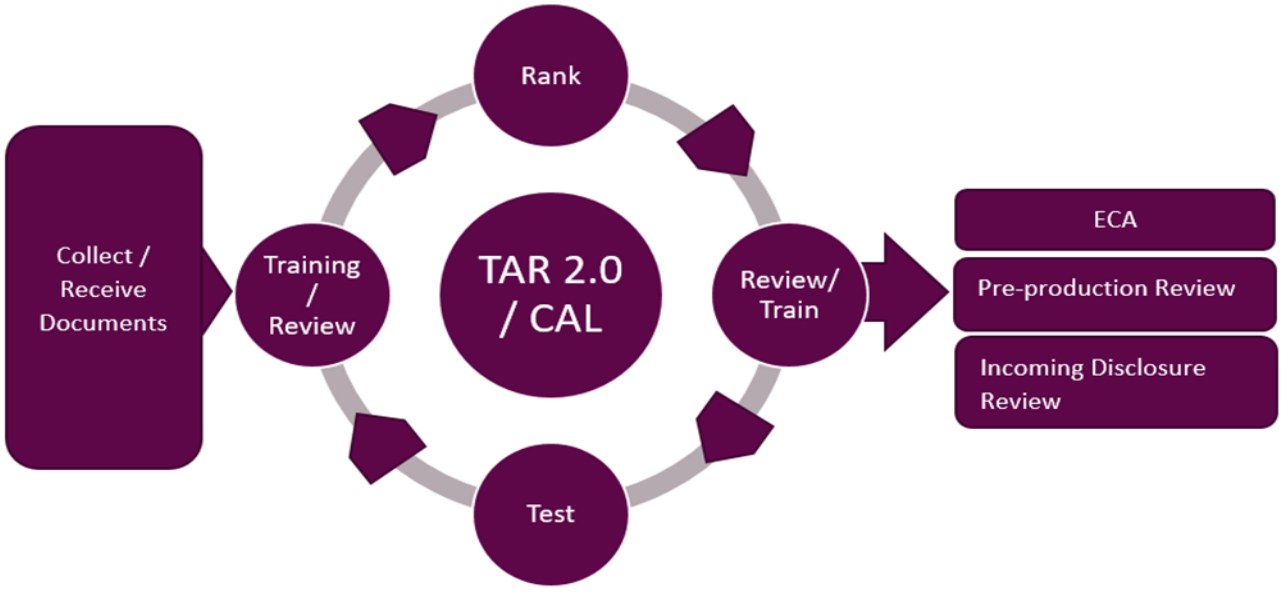
Practicable applicability of TAR



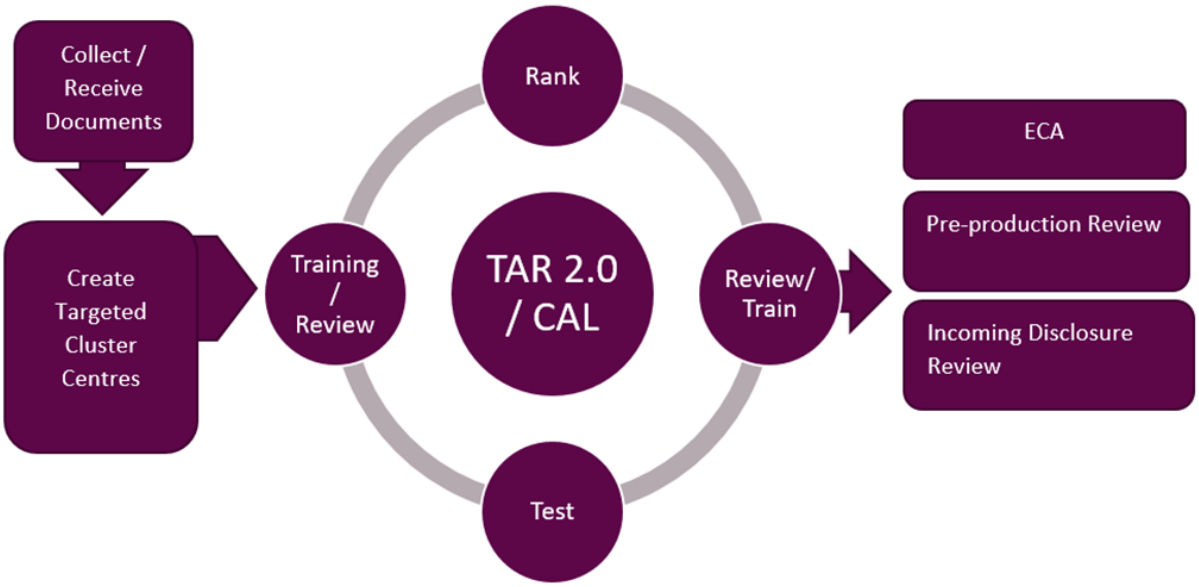
A Practicable Application of TAR Models

TAR 1.0		
Step 1	**Seed Sets**	A representative cross-section of documents are taken from the full review pool.
Step 2	**Coding for Responsiveness – Relevant / Not Relevant**	Expert Reviewers apply relevant/not relevant coding/tagging to each document in the seed set. As the technology needs to be trained properly, for the seed set review, it is best practice to engage senior lawyers who possess expert subject-matter knowledge. (SMEs) In *Pyrrho* the Court said, "*best practice would be for a single, senior lawyer who has mastered the issues in the case to consider the whole* [seed/training] *sample*". The results of this exercise are fed into the predictive coding software.
Step 3	**The Predictive Coding formula is generated**	The predictive coding software runs an analysis over the seed set to create the appropriate algorithm for predicting the relevance status of future documents by ranking the documents.
Step 4	**Perform more training rounds to stabilise the predictive coding model**	Expert reviewers trial the results of the algorithm on further document samples taken from the general review pool. The results are used to refine the algorithm by continually coding and inputting sample documents until the desired results are achieved. Decisions on the level of recall (i.e., the percentage of the total predicted relevant documents that will be returned by the selected criteria) and cut-off scores (i.e., when to stop the document review) required.
Step 5	**Apply the final prediction score to review prioritised documents.**	After all training samples are complete and the predictive coding model stabilised, the final prediction scores are applied to the remaining review pool – this allows the prioritisation of relevant material.
Step 6	**Quality control and validation**	Documents filtered out are randomly sampled and reviewed to validate their low –level of relevance.
Step 7	**Conclude the review**	

TAR 2.0 (CAL)		
Step 1	**Document Corpus Assessment**	Documents that have not been OCRed/ do not contain extracted text, or are otherwise unsuitable for predictive coding are removed from the CAL population (for which a textual analysis is required).
Step 2	**Coding for Responsiveness—Relevant / Not Relevant**	Expert reviewers apply "*relevant/not relevant*" tags to a random selection of documents to train the initial statistical model. Rankings are then applied across the documents, teaching the system to locate more relevant documents.
Step 3	**Continuous application of the predictive coding formula / Iterative Review**	As the expert reviewers assess and code relevant documents, the system continuously updates the statistical model and promotes potentially relevant documents for priority review.
Step 4	**Unreviewed Document Sampling**	This approach can incorporate measurable exposure to the breadth of documents, resulting in the application of this snapshot model onto the unreviewed population. This will define the outer boundary of a review population to meet case-specific needs.
Step 5	**Cut-off**	As the percentage of documents ranked as relevant will begin to tail off, considerations should be made for an appropriate point in time at which to conclude the review.
Step 6	**Quality control and validation**	Documents filtered out are randomly sampled and reviewed to validate their low level of relevance



TAR 3.0		
Step 1	**Targeted Cluster Centre Samples Created**	Documents that have not been OCRed/ do not contain extracted text, or are otherwise unsuitable for predictive coding are removed from the CAL population (for which a textual analysis is required). Cluster centres are created, with documents drawn from targeted samples.
Step 2	**Coding for Responsiveness – Relevant / Not Relevant**	Expert reviewers apply "*relevant/not relevant*" tags to a random selection of documents to train the initial statistical model. Rankings are then applied across the documents - teaching the system to locate more relevant documents
Step 3	**Continuous application of the predictive coding formula / Iterative Review**	As the expert reviewers assess and code relevant documents from cluster centres, the system continuously updates the statistical model and promotes potentially relevant documents
Step 4	**Unreviewed Document Sampling**	This approach can incorporate measurable exposure to the breadth of documents, resulting in the application of this snapshot model onto the unreviewed population. This will define the outer boundary of a review population to meet case-specific needs.
Step 5	**Cut-off**	As the percentage of documents ranked as relevant will begin to tail off, considerations should be made for an appropriate point in time at which to conclude the review.
Step 6	**Quality control and validation**	Documents filtered out are randomly sampled and reviewed to validate their low level of relevance.



Consideration when using TAR

Review Readiness	Confirm whether the document collection is complete or whether TAR will accommodate the rolling ingestion of new data. The richness or prevalence of responsive documents can influence the performance of TAR and workflows.
Cost	Costs of resource, including staff and contract document reviewers, as well as the technology and the eDisclosure provider costs. Some approaches will allow an earlier estimation of the numbers that will help with effective work flow and cost planning, i.e. the number of documents requiring review and the number of responsive documents expected to be found.
Time	Time considerations, including the time it will take to achieve key milestones – starting review, understanding the contents of a document collection and, ultimately, production – as well as the time it will take subject matter experts to train a system, when that is required.
Subject Matter Knowledge	Subject matter knowledge is essential to correctly train the algorithm. SMEs' / Expert Reviewers' availability and experience, as well as their in-depth knowledge of the matter, will need due consideration to ensure effective output from the TAR model.
Quality Standards	Optimal precision and recall standards need to be drawn in line with eDisclosure obligations.

fieldfisher

Predictive Coding Terminology

Prevalence / Richness	The percentage of relevant documents within a disclosure-set. By way of example, if, in a case of 100 documents, 10 are relevant, prevalence would be 10%. Prevalence varies from matter to matter. Lower prevalence in any TAR exercise will make it difficult to locate relevant material, which will have a knock-on effect on the cost, time and effort required to locate relevant documents.
Precision	The percentage of relevant documents identified by the TAR model. Precision is a model of accuracy. High precision is desirable as it means that the algorithm has correctly identified relevant documents. Therefore the review team are not wasting time and effort reviewing not relevant documents that the model may have incorrectly identified as relevant. By way of example, if, the TAR model identifies 100 documents as relevant, but only 90 are truly relevant and 10 not relevant, the precision of the model is 90%.
Recall	Recall is a measure of completeness – it assesses the percentage of the number of relevant documents retrieved by the system. High recall is desirable for defensibility purposes, and is usually the metric that legal teams focus on, because it estimates whether a reasonable number of relevant documents were found. By industry standards, recall of 75% - 80% is considered reasonable. The prevalence of relevant documents will influence recall.
F1 / F score / F measure	F1 is the weighted average of precision and recall. It scores both false positives and false negatives to determine effectiveness and strike a balance between precision and recall. In order to achieve a high F1 score, both high recall and high precision are required.

Contact

Fiona Campbell
Senior Associate, Dispute Resolution
+44 (0)330 460 6620
+44 (0)7741 905675
fiona.campbell@fieldfisher.com