Exploring Probation and Parole Records Using Natural Language Processing: A Case Study of Supervisory Condition Notes

Hadeel Elyazori
Teneshia Thurman
Kevin Lybarger
Faye S. Taxman
George Mason University

PROBATION AND PAROLE conditions are generally set by the Judiciary and/or parole board and define obligations that individuals under supervision must address. Officers typically manage compliance with these conditions. Condition management is an important part of client supervision and requires officers to document various degrees of progress towards meeting these conditions. The documentation of conditions is complicated given the high number of conditions (~8-30) per individual on supervision. Further, the documentation technology is cumbersome, with conditions documented through categorical codes, open-ended text, or a combination of both. This combination of categorical data and unstructured text data complicates large-scale analyses to identify patterns or trends. Consequently, an agency is unlikely to use the text information to review benchmarks or assess the performance of the probation or parole system. Agencies often search for ways to use this textual information, especially since officers are asked or required to enter the data into their automated case management system. The following case study illustrates some natural language processing (NLP) methods that can abstract and summarize the text data and demonstrate the utility of this approach.

NLP is a subfield of artificial intelligence (AI) focused on transforming and interpreting human-generated language. Contemporary AI and NLP are based on machine learning techniques, in which algorithms automatically learn patterns from large data sets. Lauriola et al. provide an overview of NLP, including deep learning techniques.[1] Here we explore the use of NLP-based information extraction techniques, which automatically map unstructured text to a structured semantic representation to facilitate large-scale and real-time analyses. Combining extracted information from officer case notes with the available structured data can create a more holistic understanding of clients and provide actionable insights regarding criminal history, behavior patterns, probation compliance, and other outcomes. A review of the published literature suggests a notable gap in the application of machine learning techniques for information extraction specifically within the context of probation and parole case notes. However, information extraction is well established in other contexts, such as legal documents,[2,3] healthcare,[4] and finance.[5]

The goal of this study is to enable data-centric strategies for better understanding probation and parole practices. We explore officer case notes describing conditions of supervision and use information extraction techniques to convert the unstructured case notes to a semantic representation. We developed a fined-grained, hierarchical annotation (coding) schema for 66 *Condition Category* labels associated with supervisory conditions related to substance use, mental health, treatment programs, community service, education, employment, fines, fees, and other conditions. The 66 *Condition Categories* are related to 10 higher level *Parent Categories*. We annotated the records of over 3,000 clients in a state department of parole and probation and used this annotated corpus to develop information extraction models based on traditional machine learning algorithms

Glossary

- *AI - Artificial Intelligence*
- *BERT - Bidirectional Encoder Representations from Transformers*
- *FLAN - Fine-tuned LAnguage Net*
- *LLM - Large Language Model*
- *NLP - Natural Language Processing*
- *PII - Personally Identifiable Information*
- *RF - Random Forest*
- *SVM - Support Vector Machines*
- *T5 - Text-to-Text Transfer Transformer*
- *TF-IDF - Term Frequency-Inverse Document Frequency*

and state-of-the-art Large Language Models (LLMs). Our results demonstrate the feasibility of using information extraction techniques on probation and parole case notes and provide a foundation for enhancing data analytics within criminal justice settings.

Related Work

AI is increasingly explored within criminal justice, including crime detection,[6] prevention,[7] and forecasting[8,9] and decision support.[10] As examples, Shah et al. used computer vision to forecast crime in videos,[7] and Tollenaar et al. developed machine learning models to predict recidivism risk.[11] Advancements in deep learning (neural networks) are expanding the capabilities and performance of AI in criminal justice and other settings. For example, deep learning crime prediction models can successfully leverage diverse data, including videos, images, audio recordings, and text data, and achieve improved performance over traditional machine learning methods.[8] (See Figure 1.)

Information extraction research within the criminal justice domain has been primarily limited to online law enforcement investigations[12] and legal documents, focusing on names, regulations, legal norms, etc.[2,3] Information extraction research is sparse or non-existent within parole and probation settings. Some research explores parole hearing transcripts, focusing on extracting offenses, gang programming, employment, education, and risk scores.[13] Current literature reviews indicate a scarcity of published research exploring the application of information extraction techniques to parole and probation case notes to understand the supervision process. This lack of published research constitutes a missed opportunity to use technology to improve the supervision and management of offenders. While there is an absence of information extraction work focused on parole and probation case notes, there is a robust body of clinical information extraction research focused on clinician-generated notes describing patients within electronic health records.[4] Clinical data is similar to probation and parole data in that both: i) include structured data and narrative text, ii) contain personally identifying information (PII), iii) document individuals through various domain-specific events, and iv) capture information related to socioeconomic status and health. Our experimentation is informed by clinical information extraction methods.

Information extraction has evolved over time, presenting a continuum from rule-based systems to machine learning and deep learning,[2,3,4,5] where the peformance and capabilities of algorithims have increased over time. Rule-based systems consist of manually curated rules to identify predefined linguistic patterns. Frequently employed traditional machine learning models include logistic regression, Random Forest (RF), and Support Vector Machines (SVM).[2,3] RF ensembles multiple decision trees to make predictions (see Figure 2A), and SVM finds the optimal boundary to separate categories[2,3] (see Figure 2B). For traditional methods, a common approach for converting text to input features is Term Frequency-Inverse Document Frequency (TF-IDF), which assigns weights to words based on their frequency[3] (see Figure 2C). TF-IDF word weighting assigns higher values to words that are more frequent in a document and less frequent in the other documents in the corpus. More recently, neural networks, like Convolutional Neural Networks and Recurrent Neural Networks, have achieved prominence over traditional methods due to their capacity for automated feature learning and ability to model complex relationships within text data.[2,3,4,5]

FIGURE 1

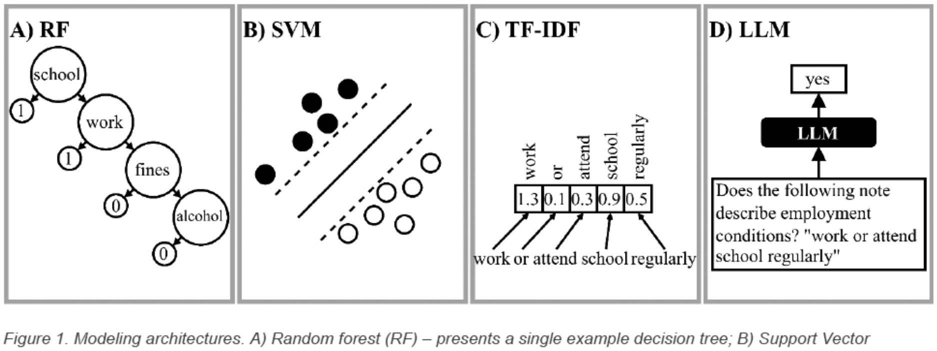

A) RF
school
1
work
1
fines
0
alcohol
0
B) SVM
C) TF-IDF
work
or
attend
school
regularly
1.3
0.1
0.3
0.9
0.5
work or attend school regularly
D) LLM
yes
LLM
Does the following note describe employment conditions? "work or attend school regularly"

Figure 1. Modeling architectures. A) Random forest (RF) – presents a single example decision tree; B) Support Vector Machines – illustrates defining a decision boundary that optimally separates samples; C) Term frequency-inverse document frequency (TF-IDF) – presents example mapping of input text to a feature vector; and D) Large language model (LLM) – presents an example where the model input and output are text.

LLMs, like ChatGPT,[14] currently dominate the NLP landscape and achieve state-of-the-art performance in myriad tasks, including information extraction. LLMs are built on transformer architectures and include millions to trillions of trainable parameters. The typical training approach involves *pre-training* on extensive unlabeled text corpora to acquire a generalized understanding of language, followed by *fine-tuning* (supervised learning) on labeled data to learn a specific task. This transfer learning paradigm is particularly advantageous in domains where annotated data is limited, a condition relevant to corrections and community supervision settings. To address privacy concerns related to PII, we focus on two publicly available architectures: Bidirectional Encoder Representations from Transformers (BERT)[15] and Text-to-Text Transfer Transformer (T5)[16] (see Figure 2D). BERT encodes text by transforming input word sequences into vectors that can be used for classification. BERT has achieved state-of-the-art performance in many information extraction tasks across domains.[2,3,4,5] T5 is a generative model that transforms input text to output text and can be used for many tasks, including classification. T5 has achieved state-of-the-art performance in many tasks, including the extraction of social determinants of health in clinical notes.[17]

Methods and Materials

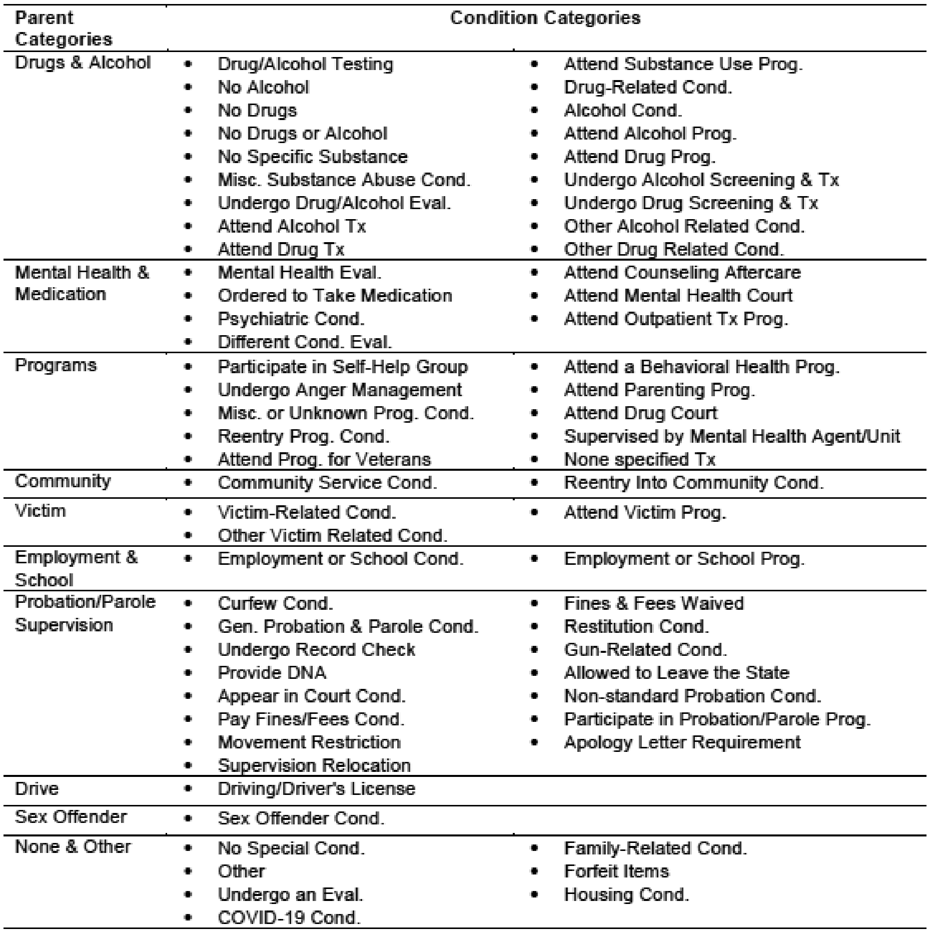
Data

In this study, we used client case plan data from a parole and probation agency located in a mid-Atlantic state. The data includes over 3,000 unique clients, which covers cases opened from 2017-2021. Client case plans describe the requirements and conditions an individual must follow during supervision. Probation/parole officers use an agency's database to document conditions and design goals to achieve them. The goals can refer to activities such as random urinalysis, taking prescribed medication, obtaining mental health evaluation, participating in mental health treatment, and other requirements. In our study, the agency-provided data included 120 *Condition Codes*, each with a corresponding *Condition Description* that is constant across all records. For example, agency-provided *Condition Codes* 9532 and 16028 have the *Condition Descriptions* "Other" and "Additional Drug condition," respectively. Within the dataset, there were 34 *Condition Codes* that also included an officer-generated *Condition Note* documenting case plan details through unstructured narrative text. For example, the *Condition Code* 9532 with *Condition Description* “Other” serves as one of several catchall codes for conditions that do not easily fit more specific codes. The officer-generated *Condition Notes* for *Condition Code* 9532 document a wide range of conditions, such as “Defendant not to drive,” “Seek employment/school,” or “Gun registry.” The 34 *Condition Codes* with associated *Condition Notes* include: 1) *Other* – 12 codes were described as “other” and serve as a catchall for undefined conditions; 2) *Programs* – 4 codes require officers to specify particular programs, for example behavioral health, domestic violence, veteran, family counseling, and vocational programs; 3) *Substance Use* – 5 codes pertain to drug or alcohol conditions; 4) *Victim* – 2 codes were victim-focused conditions; 5) *Sex Offender* – 2 codes were related to sex offenders’ special conditions; and 6) *Additional Conditions* – 9 codes addressed specific requirements or restrictions, which involved completion of assigned tasks or community service, financial obligations such as court costs and restitution, geographical limitations, and specified durations of home confinement or other monitoring requirements. Officers can amend their case plans through supervision. Each client may have multiple parole or probation cases, and each case can include multiple conditions. We treat each condition record (*Condition Code*, *Condition Description*, and *Condition Note*) as a sample or record. In addition to conditions, the data set includes: 1) *case type* – parole vs. probation and 2) *case level* – low, low-moderate, moderate, maximum, special cases, or violent.

Annotation

The primary objective of the annotation was to identify and categorize the 34 *Condition Codes* that included an officer-generated *Condition Note*; however, we developed a comprehensive set of *Condition Category* labels that summarized the meaning of all 120 *Condition Codes*. Officers manually type the *Condition Notes*, requiring a comprehensive review and categorization process. Based on our review of the data, we developed a set of 66 *Condition Category* labels to map the condition records, including unstructured *Condition Note* information, to a fixed set of classes. Annotation involved assigning one or more of the researcher-defined 66 *Condition Category* labels to the agency-provided records. Table 1 summarizes

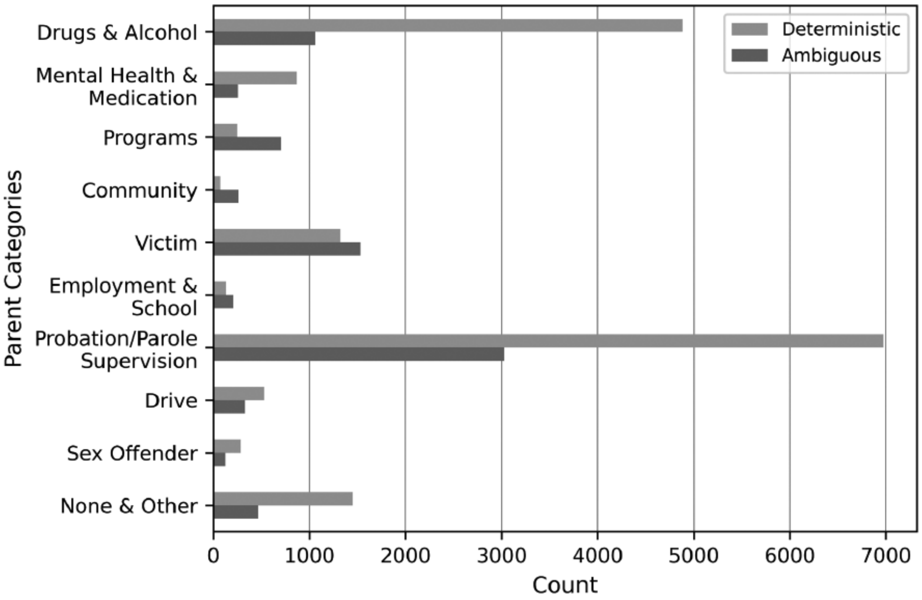
TABLE 1
Condition Category Hierarchy



Parent Categories	Condition Categories	
Drugs & Alcohol	• Drug/Alcohol Testing • No Alcohol • No Drugs • No Drugs or Alcohol • No Specific Substance • Misc. Substance Abuse Cond. • Undergo Drug/Alcohol Eval. • Attend Alcohol Tx • Attend Drug Tx	• Attend Substance Use Prog. • Drug-Related Cond. • Alcohol Cond. • Attend Alcohol Prog. • Attend Drug Prog. • Undergo Alcohol Screening & Tx • Undergo Drug Screening & Tx • Other Alcohol Related Cond. • Other Drug Related Cond.
Mental Health & Medication	• Mental Health Eval. • Ordered to Take Medication • Psychiatric Cond. • Different Cond. Eval.	• Attend Counseling Aftercare • Attend Mental Health Court • Attend Outpatient Tx Prog.
Programs	• Participate in Self-Help Group • Undergo Anger Management • Misc. or Unknown Prog. Cond. • Reentry Prog. Cond. • Attend Prog. for Veterans	• Attend a Behavioral Health Prog. • Attend Parenting Prog. • Attend Drug Court • Supervised by Mental Health Agent/Unit • None specified Tx
Community	• Community Service Cond.	• Reentry Into Community Cond.
Victim	• Victim-Related Cond. • Other Victim Related Cond.	• Attend Victim Prog.
Employment & School	• Employment or School Cond.	• Employment or School Prog.
Probation/Parole Supervision	• Curfew Cond. • Gen. Probation & Parole Cond. • Undergo Record Check • Provide DNA • Appear in Court Cond. • Pay Fines/Fees Cond. • Movement Restriction • Supervision Relocation	• Fines & Fees Waived • Restitution Cond. • Gun-Related Cond. • Allowed to Leave the State • Non-standard Probation Cond. • Participate in Probation/Parole Prog. • Apology Letter Requirement
Drive	• Driving/Driver's License	
Sex Offender	• Sex Offender Cond.	
None & Other	• No Special Cond. • Other • Undergo an Eval. • COVID-19 Cond.	• Family-Related Cond. • Forfeit Items • Housing Cond.

Abbreviations: condition (cond.), evaluation (eval.), general (gen.), miscellaneous (misc.), program (prog.), and treatment (Tx).

FIGURE 2
Distribution of Condition Categories



Deterministic indicates the Condition Category label can be assigned based solely on the agency-provided Condition Code. Ambiguous indicates the Condition Category label assignment requires interpretation of the office-generated Condition Note text.

the assigned labels, which are hierarchically arranged with 66 *Condition Category* labels assigned to 10 *Parent Categories*. Among the 120 *Condition Codes*, 86 *Condition Codes* do not include *Condition Notes* and always correspond with the same *Condition Category*, so they can be deterministically assigned a *Condition Category* label; and 34 *Condition Codes* include *Condition Notes* that must be interpreted to resolve ambiguity regarding the relevant *Condition Category* label. Before manual coding of the case plan requirements, *Condition Categories* were automatically assigned to the 86 deterministic *Condition Codes* that do not include associated *Condition Notes*, and manual annotation focused on resolving the ambiguity associated with the 34 *Condition Codes* that included narrative text through *Condition Notes*. During the annotation process, new *Condition Category* labels were added to the label set if the condition did not align with existing categories. Samples were annotated by three individuals with domain expertise, including backgrounds in criminology and policy. Extensive annotation training ensured data quality and annotation consistency.

Figure 2 summarizes the distribution of the *Parent Category* labels broken down by: 1) *deterministic* – the record does not include officer-generated text (*Condition Note*), and the *Condition Category* label can be assigned to the record based solely on the *Condition Code* and 2) *ambiguous* – officer-generated *Condition Note* text must be interpreted to determine the appropriate category label. In total, 48 percent of records require interpretation of the officer-generated *Condition Note*, indicating the text's importance in understanding the assigned condition.

Condition Category Dependence

To better understand the relationship between the *Condition Category* labels and the client case type and level, we performed Chi-squared test of independence between each *Condition Category* label and the case type and level. The case type is binary (probation vs. parole). The case level is multiclass, and we converted the case level labels to a binary one-versus-rest representation before performing the statistical test.

Information Extraction

We explored the *Condition Category* prediction task for the records with a *Condition Note* (ambiguous records in Figure 2), using traditional machine learning models and LLMs. For all experiments, the model input is the client record (*Condition Code*, *Condition Description*, and *Condition Note*). In our annotation scheme, each record can be assigned multiple *Condition Category* labels, so we treat this task as a multi-label binary prediction task, where each record is assigned a set of 66 binary labels (1 indicates category relevant, and 0 indicates category irrelevant). Figure 3 presents an overview of the modeling approaches, including examples of how the record is represented in the input. The *Condition Code* and *Condition Description* are included with the *Condition Note* in the model input to provide important context for interpretation.

Traditional Machine Learning

We explored two traditional machine learning models: 1) RF and 2) SVM. The input to these models includes the *Condition Code* and TF-IDF representation of the *Condition Description* and *Condition Note*. The RF/SVM models learn feature weights for the features to predict the *Condition Category* labels. Separate RF and SVM models were developed for each *Condition Category*, and predictions from the category-specific models were combined to form a set of predictions for each record. Figure 3A presents an example of a single RF/SVM classifier, where the output is a binary prediction for a single *Condition Category*.

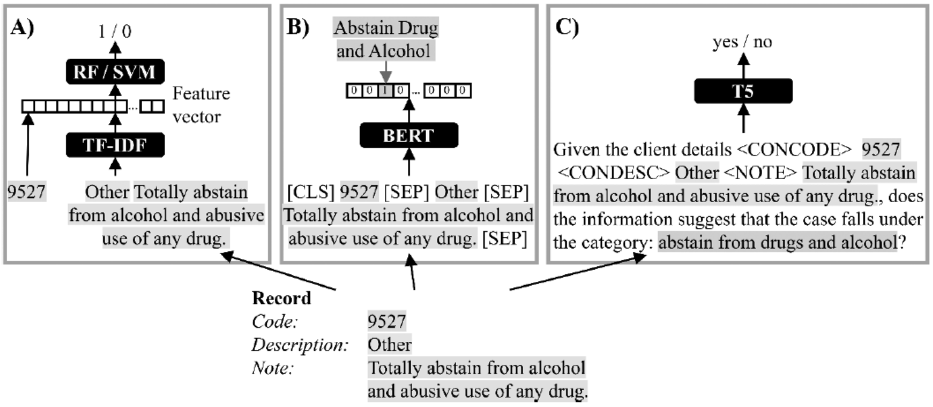
LLMs

We explored two LLMs: BERT and T5. BERT is pretrained on a large body of text to learn a general representation of language. In this pretraining, special tokens are included to define the input format, including: CLS – specifies the start of the input and SEP – serves as a separator for different inputs. As shown in Figure 3B, the BERT input consists of the *Condition Code*, *Condition Description*, and *Condition Note* separated by the SEP token. BERT maps this input text to an output vector, and separates linear functions for each *Condition Category* to generate binary predictions. In this configuration, a single BERT model can generate all 66 multi-label predictions. As is common practice, we started with a pretrained BERT model, then trained the BERT model and output linear functions on the labeled data. As presented in Figure 3C, we used T5 to assign *Condition Category* labels using a question-answering (QA) setting. In this QA setting, a separate yes/no question is formulated for each *Condition Category*, and the set of yes/no questions spanning all *Condition Category* labels is applied to each record. The input to T5 includes a *Condition Category*-specific question and the *Condition Code*, *Condition Description*, and *Condition Note* separated by special tokens (e.g., <Code> or <Description>) to differentiate input information. The T5 output is a "yes" / "no" answer to the *Condition Category*-specific question.

Experimental Paradigm

Modeling was implemented using the Python packages Scikit-learn[18] and Transformers.[19] Records were divided into three subsets at the client level: 70 percent training, 10 percent validation, and 20 percent testing. The optimal configuration (hyperparameters) for each model was determined by training models on the training set and evaluating performance on the validation set. We report the performance on the withheld test set using the optimal configurations. Detailed model configurations are presented in the Appendix.

FIGURE 3
Information Extraction Architectures


A)
1 / 0
RF / SVM
Feature vector
TF-IDF
9527
Other Totally abstain from alcohol and abusive use of any drug.
B)
Abstain Drug and Alcohol
0 0 1 0 ... 0 0 0
BERT
[CLS] 9527 [SEP] Other [SEP] Totally abstain from alcohol and abusive use of any drug. [SEP]
C)
yes / no
T5
Given the client details <CONCODE> 9527 <CONDESC> Other <NOTE> Totally abstain from alcohol and abusive use of any drug., does the information suggest that the case falls under the category: abstain from drugs and alcohol?
Record
Code: 9527
Description: Other
Note: Totally abstain from alcohol and abusive use of any drug.

Performance

Performance is evaluated using precision, recall, and F1, as defined in Equation 1. Given the high number of *Condition Category* labels, we report the micro-averaged performance at the *Parent Category* level and include individual *Condition Category* performance in the Appendix. The statistical significance of the results was evaluated using a pairwise non-parametric test (bootstrap test, p-value<0.05).[20]

Results

Condition Category Dependence

Our study first focused on comparing our *Condition Category* labels with the agency-provided case types and case levels through Chi-squared tests of independence summarized in Table 2. For space, Table 2 only presents 32 of the 66 *Condition Category* labels that are dependent on case type or level. The triangles (▲ or ▼) indicate that the *Condition Category* label and case level or type (the *variables*) are dependent. An upward-facing triangle (▲) indicates the variables co-occur more frequently and a downward-facing triangle (▼) indicates the variables co-occur less frequently than expected, if the variables were independent. The diversity in the conditions across different case levels and types illustrates the complexity of decision-making and the tailored strategies employed to address the varying needs and risks associated with each case; however, several themes emerged from this analysis. Probation tends to have higher rates than parole for conditions related to drugs and alcohol, self-help, anger management, community service, victims, waiving fees, guns, driver's licenses, and forfeiture of items. Conversely, parole has higher rates than probation for conditions related to employment, curfew, paying fees, and sex offender conditions. Lower level offenders (low and low/moderate) tend to have higher prevalence than higher level offenders (moderate, maximum, special case, and violent) for conditions related to drugs and alcohol, self-help, community service, attending victim programs, and driver's license. Conversely, higher level offenders have higher rates than lower level offenders for conditions related to anger management, victim conditions (other than victim programs), curfew, and sex offender conditions.

Classification Performance

Table 3 presents the prediction performance on the withheld test set for the *Condition Category* labels micro-averaged for each *Parent Category*. In information extraction research, performance varies by task and data set, and there are not predefined thresholds for good/acceptable performance; however, we consider performance ≥ 0.90 F1 to be very high. The LLMs (BERT and T5) outperformed the traditional machine learning models (RF and SVM) in the overall performance, as well as the performance in 5 of the 10 *Parent Categories*, with significance, demonstrating the natural language understanding capabilities of the LLMs. Among all models, T5 achieved the highest overall performance and *Mental Health & Medication* performance with significance. These results demonstrate the feasibility of developing high-performing information extraction models for probationary notes and highlight the value of using LLMs. Table 4 in the Appendix presents the performance for the individual *Condition Category* labels.

Error Analysis

Each *Parent Category* includes a set of topically relevant *Condition Categories*. The performance for the *Parent Categories* tends to be higher when there are fewer associated *Condition Categories*, as the classification models need to disambiguate fewer topics. For example, the T5 performance is ≥ 0.97 F1 for the *Parent Categories* – *Community*, *Victim*, and *Drive* – which have 2, 3, and 1 child labels respectively. Additionally, the highest performing *Parent Categories* include *Condition*

EQUATION 1
Precision, Recall, and F1 Formulas

$$Precision = \frac{TP}{TP + FP}; \quad Recall = \frac{TP}{TP + FN}; \quad F1 = 2\frac{Precision \times Recall}{Precision + Recall}$$

Abbreviations: true positive (TP), false positive (FP), and false negative (FN)

TABLE 2
Condition Category Dependency

Parent Categories	Level - Low	Level - Low/moderate	Level - Moderate	Level - Maximum	Level - Special Case	Level - Violent	Type - Probation	Type - Parole	Condition Categories	Level - Low	Level - Low/moderate	Level - Moderate	Level - Maximum	Level - Special Case	Level - Violent	Type - Probation	Type - Parole	Condition Categories
Drugs & Alcohol		▲				▼			Drug/Alcohol Testing	▲				▼	▼			Attend Substance Use Prog.
		▲				▲			Misc. Substance Abuse Cond.	▲	▲	▼	▼		▼	▲	▼	Attend Alcohol Prog.
		▲				▼	▲	▼	Attend Alcohol Tx					▼				Attend Drug Tx
					▼		▲	▼	Undergo Drug/Alcohol Eval.									
Mental Health & Medication				▲	▲				Psychiatric Cond.					▲				Attend Mental Health Court
Programs	▲					▼	▲	▼	Participate in Self-Help Group				▲					Attend a Behavioral Health Prog.
	▼	▼	▲	▲			▲	▼	Undergo Anger Management					▲				Attend Parenting Prog.
						▲			Misc. or Unknown Prog. Cond.				▲					Attend Drug Court
			▲						Reentry Prog. Cond.									
Community	▲				▼	▼	▲	▼	Community Service Cond.									
Victim	▼	▼	▲	▲	▲		▲	▼	Victim-Related Cond.	▲	▲	▼	▼		▼	▲	▼	Attend Victim Prog.
Employment & School	▼		▲				▼	▲	Employment or School Cond.						▲			Employment or School Prog.
Probation/ Parole Supervision				▲	▲	▲	▼	▲	Curfew Cond.				▲			▲	▼	Fines & Fees Waived
				▼			▼	▲	Gen. Probation & Parole Cond.			▲			▲			Restitution Cond.
	▲	▼		▼			▲	▼	Undergo Record Check	▼			▲		▲	▲	▼	Gun-Related Cond.
	▲			▼	▼	▲	▼	▲	Pay Fines/Fees Cond.									
Drive	▲		▲	▼	▼	▼	▲	▼	Driving/Driver's License									
Sex Offender	▼	▼	▲	▲	▲	▼	▼	▲	Sex Offender Cond.									
None & Other		▼		▲		▲	▲	▼	Forfeit Items							▼	▲	COVID-19 Cond.

An upward or downward facing triangle (▲ or ▼) indicates the Condition Category label and case level or type are dependent (p<0.05, null hypothesis of independence rejected). An upward facing triangle (▲) indicates the variables co-occur more frequently than expected if independent, and a downward facing triangle (▼) indicates the variables co-occur less frequently than expected if independent. Abbreviations: condition (cond.), evaluation (eval.), general (gen.), miscellaneous (misc.), program (prog.), and treatment (Tx).

Categories with very consistent linguistic cues (keywords). For example: i) *Community Service Condition* – "community service" or time commitment, like "10 hours per week"; ii) *Victim-Related Condition* – "no contact with" or "do not enter"; iii) *Attend Victim Program* – "victim impact panel" or "VIP"; and iv) *Driving/Driver's License* – "drive," "interlock," or "license."

The performance for *Parent Categories* tends to be lower when there are more associated *Condition Categories*, as the models must distinguish between more closely related topics. For example, the T5 performance is ≤ 0.82 F1 for the *Parent Categories* – *Drug & Alcohol*, *Programs*, and *None & Other* – which have 18, 10, and 8 child labels respectively. Within these *Parent Categories*, the individual *Condition Category* performance varies, and performance decreases as linguistic diversity increases. For example, the *Condition Category Miscellaneous or Unknown Program Condition* is a catchall for requirements related to a range of programs, and the notes contain diverse language, references to specific treatment facilities, and ambiguous statements like "successfully complete treatment." As another example, the *Condition Category* – *Attend Substance Use Program* – includes notes describing several different specific treatment programs and facilities and includes less common shorthand, like "ALC PGM" for "Alcohol Program."

Discussion

The overarching goal of this study is to enable probation and parole agencies to use the information captured in officer-generated notes in large-scale and real-time analyses. This goal is highly significant, due to the prevalence of open-ended text fields in management information systems, importance of the textual information, and challenges associated with converting this textual information into quantifiable data. Through NLP information extraction techniques, the unstructured text can be converted to a structured representation to examine patterns and assess performance at all levels, including the program, officer, and individual under supervision. Agencies currently grapple with the complexity of summarizing these text data, but the strategies presented in this case study demonstrate how NLP can generate usable metrics that can easily be combined with existing categorial data. While these strategies require specific technical expertise, this work illustrates the value of AI methods.

In our study, the LLMs (BERT and T5) outperformed traditional machine learning models (RF and SVM). For the traditional models, all model learning originates from annotated training data. However, the LLMs use transfer learning, where the models first pretrain on large corpora of unlabeled text to learn language understanding and then fine-tune (train) on the annotated training data to learn the target task. The improved performance of the LLM can be attributed to the success of this learning transfer, which provides a general understanding of language. The improved performance of the T5 model relative to BERT can be attributed to the larger model size (higher number of parameters) and larger pretraining corpus.

We are unaware of any prior information extraction work exploring officers' documentation of parole and probation conditions. Our results demonstrate the feasibility of using information extraction techniques in this setting by achieving high performance across most of the *Parent Categories*. The use of NLP with correctional system data, including parole and probation notes, has the potential to improve management and supervision by enabling the automatic analysis of vast amounts of information-dense text data. It can provide a richer, data-driven understanding of offender behavior and risks and could lead to more tailored intervention strategies and more informed decision-making processes, ultimately contributing to improved rehabilitation and public safety.

TABLE 3
Classification Results on Withheld Test Set*

Parent Category	# Human-annotated Labels	F1 Micro Average			
		RF	SVM	BERT	T5
Drugs & Alcohol	219	0.61	0.71	0.79*	**0.82***
Mental Health & Medication	46	0.68	0.74	0.78*	**0.86*†**
Programs	122	0.66	0.70	**0.76**	0.75
Community	20	0.80	0.86	0.97*	**0.98***
Victim	194	0.94	0.94	0.96	**0.97**
Employment & School	41	0.85	0.89	0.92	**0.93**
Probation/Parole Supervision	347	0.74	0.78	0.86*	**0.90***
Drive	66	0.98	0.98	0.97	**0.99**
Sex Offender	25	0.89	0.92	0.94*	**0.99***
None & Other	71	0.70	0.71	**0.73**	**0.73**
OVERALL MICRO AVG.	1151	0.77	0.81	0.85*	**0.89*†**

** Indicates LLM significantly outperforms traditional model (RF and SVM). † Indicates T5 significantly outperforms BERT.*

This research has key limitations related to data heterogeneity. First, we explored a moderately sized client population from a single agency, and the populations in the analyzed data set may not be representative of other agencies. The conditions and documentation practices, including the authoring of notes by officers, may vary by agency, and additional work is needed to understand the variability of the conditions and notes across institutions. Second, we explored officer descriptions of conditions, which represent only one of many types of free-text records within correctional data. Additional analyses with more comprehensive text record types are needed to understand the feasibility and challenges associated with applying information extraction techniques more broadly within correctional system data.

Conclusions

We explored a corpus of officer-generated notes documenting the parole and probation conditions of clients under supervision and investigate the use of state-of-the-art information extraction techniques. We annotated the records of over 3,000 clients with a fine-grained annotation schema of 66 *Condition Categories* and developed information extraction models based on traditional machine learning methods and LLMs. The LLMs outperformed the traditional machine learning methods, with the generative T5 model achieving the best overall performance at 0.89 F1. This high performance demonstrates the feasibility of using NLP in this parole and probation setting and provides a foundation for future exploration of correctional system data.

Ethics

We had the necessary approvals from our institution's Institutional Review Board (IRB) to obtain, store, and analyze the probation and parole data set. All researchers and annotators received the necessary human subjects training to interact with the client data, including the PII.

Acknowledgments

We would like to thank the state agency for offering this data and giving us the opportunity to understand how to use various technologies to explore strategies to abstract information. We acknowledge that Dr. Faye S Taxman is the PI of this project and supported this study.

References

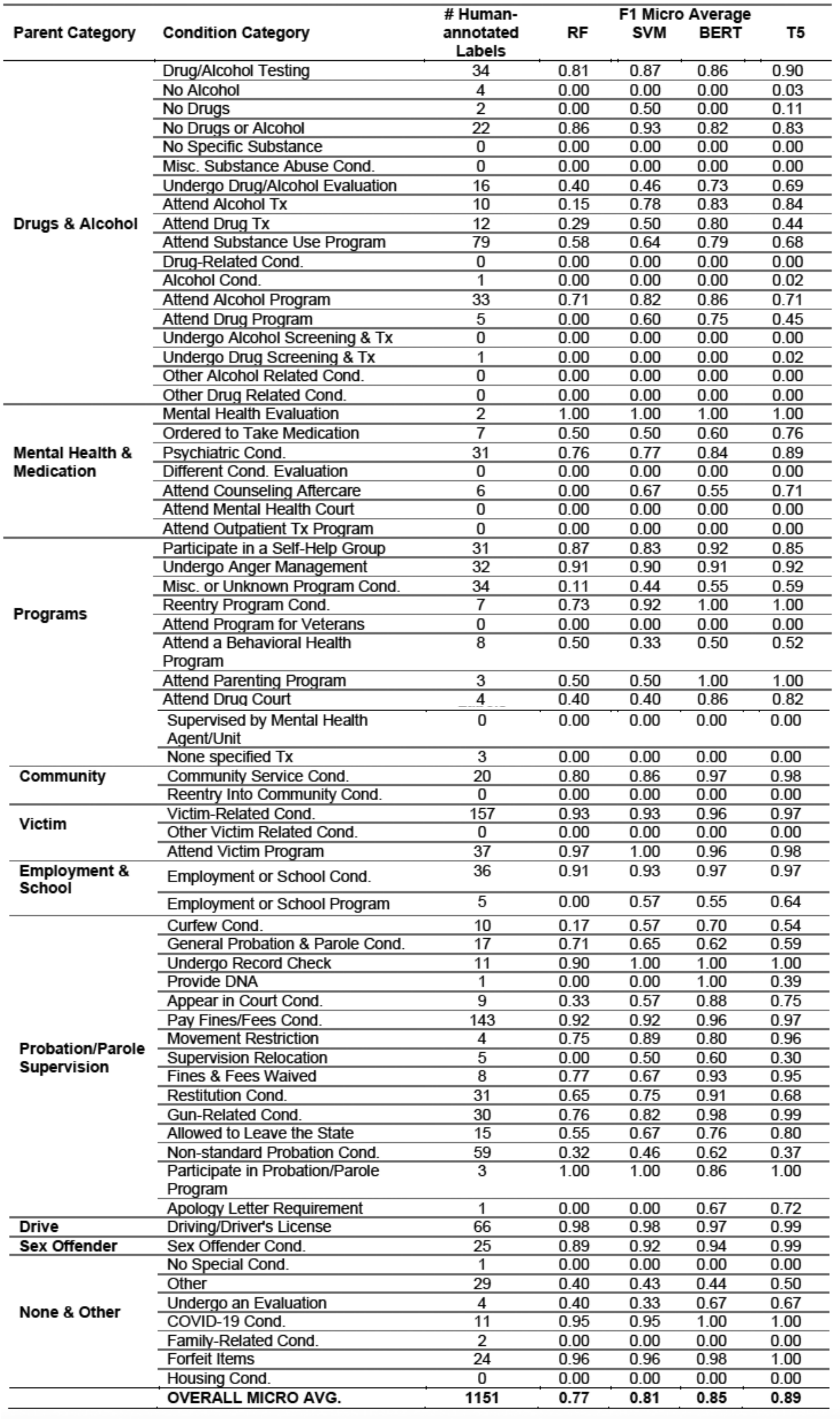
1. I. Lauriola, A. Lavelli and F. Aiolli, "An introduction to deep learning in natural language processing: Models, techniques, and tools," *Neurocomputing,* vol. 470, pp. 443-456, 1 2022.
2. F. Solihin, I. Budi, R. F. Aji and E. Makarim, "Advancement of information extraction use in legal documents," *International Review of Law, Computers and Technology,* vol. 35, no. 3, pp. 322-351, 2021.
3. C. Sansone and G. Sperlí, "Legal information retrieval systems: State-of-the-art and open issues," *Information Systems,* vol. 106, 5 2022.
4. Y. Wang, L. Wang, M. Rastegar-Mojarad, S. Moon, F. Shen, N. Afzal, S. Liu, Y. Zeng, S. Mehrabi, S. Sohn and H. Liu, "Clinical information extraction applications: A literature review," *Journal of Biomedical Informatics,* vol. 77, pp. 34-49, 1 2018.
5. M. H. A. Abdullah, N. Aziz, S. J. Abdulkadir, H. S. A. Alhussian and N. Talpur, "Systematic literature review of information extraction from textual data: Recent methods, applications, trends, and challenges," *IEEE Access,* vol. 11, pp. 10535-10562, 2023.
6. C. Rigano, "Using artificial intelligence to address criminal justice needs," *National Institute of Justice Journal,* vol. 280, pp. 1-10, 2019.
7. N. Shah, N. Bhagat and M. Shah, "Crime forecasting: A machine learning and computer vision approach to crime prediction and prevention," *Visual Computing for Industry, Biomedicine, and Art,* vol. 4, no. 1, p. 9, 2021.
8. V. Mandalapu, L. Elluri, P. Vyas and N. Roy, "Crime prediction using machine learning and deep learning: A systematic review and future directions," *IEEE Access,* vol. 11, pp. 60153-60170, 2023.
9. O. Kounadi, A. Ristea, A. Araujo Jr and M. Leitner, "A systematic review on spatial crime forecasting," *Crime Science,* vol. 9, no. 1, p. 7, 5 2020.
10. J. Mitchell, S. Mitchell and C. Mitchell, "Machine learning for determining accurate outcomes in criminal trials," *Law, Probability and Risk,* vol. 19, no. 1, p. 43–65, 2020.
11. N. Tollenaar and P. G.M. van der Heijden, "Which method predicts recidivism best?: A comparison of statistical, machine learning and data mining predictive models," *Journal of the Royal Statistical Society Series A: Statistics in Society,* vol. 176, no. 2, p. 565–584, 2013.
12. M. Edwards, A. Rashid and P. Rayson, "A systematic survey of online data mining technology intended for law enforcement," *ACM Computing Surveys,* vol. 48, no. 1, 9 2015.
13. J. Hong, C. Voss and C. D. Manning, "Challenges for information extraction from dialogue in criminal law," in *Workshop on NLP for Positive Impact*, 2021.
14. OpenAI, "GPT-4 technical report," 3 2023.
15. J. Devlin, M.-W. Chang, K. Lee and . K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Minneapolis, MN, 2019.
16. C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *Journal of Machine Learning Research,* vol. 21, pp. 1-67, 2020.
17. B. Romanowski, A. Ben Abacha and Y. Fan, "Extracting social determinants of health from clinical note text with classification and sequence-to-sequence approaches," *Journal of the American Medical Informatics Association : JAMIA,* vol. 30, no. 8, pp. 1448-1455, 7 2023.
18. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. P. Passos, D. Cournapeau, M. Brucher, M. Perrot and É. Duchesnay, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research,* vol. 12, no. 85, pp. 2825-2830, 2011.
19. T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. Von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest and A. M. Rush, "Transformers: State-of-the-Art Natural Language Processing," in *Association for Computational Linguistics*, 2020.
20. T. Berg-Kirkpatrick, D. Burkett and D. Klein, "An Empirical Investigation of Statistical Significance in NLP," in *Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, Jeju Island, Korea, 2012.

Appendix

Model Configuration

Each architecture includes some model-specific configuration. For the RF, the optimum hyperparameters include class weight = balanced subsample, maximum depth = 50, and number of estimators = 200. For the SVM, the optimum hyperparameters include C = 100. For BERT, we started with the pretrained model *bert-base-uncased* and trained the model for 29 epochs. For T5, we started with the pretrained *model flan-t5-large* and trained the model for 20 epochs.

TABLE 4
Detailed Performance for the Individual Condition Category Labels



Parent Category	Condition Category	# Human-annotated Labels	F1 Micro Average			
			RF	SVM	BERT	T5
Drugs & Alcohol	Drug/Alcohol Testing	34	0.81	0.87	0.86	0.90
	No Alcohol	4	0.00	0.00	0.00	0.03
	No Drugs	2	0.00	0.50	0.00	0.11
	No Drugs or Alcohol	22	0.86	0.93	0.82	0.83
	No Specific Substance	0	0.00	0.00	0.00	0.00
	Misc. Substance Abuse Cond.	0	0.00	0.00	0.00	0.00
	Undergo Drug/Alcohol Evaluation	16	0.40	0.46	0.73	0.69
	Attend Alcohol Tx	10	0.15	0.78	0.83	0.84
	Attend Drug Tx	12	0.29	0.50	0.80	0.44
	Attend Substance Use Program	79	0.58	0.64	0.79	0.68
	Drug-Related Cond.	0	0.00	0.00	0.00	0.00
	Alcohol Cond.	1	0.00	0.00	0.00	0.02
	Attend Alcohol Program	33	0.71	0.82	0.86	0.71
	Attend Drug Program	5	0.00	0.60	0.75	0.45
	Undergo Alcohol Screening & Tx	0	0.00	0.00	0.00	0.00
	Undergo Drug Screening & Tx	1	0.00	0.00	0.00	0.02
	Other Alcohol Related Cond.	0	0.00	0.00	0.00	0.00
	Other Drug Related Cond.	0	0.00	0.00	0.00	0.00
Mental Health & Medication	Mental Health Evaluation	2	1.00	1.00	1.00	1.00
	Ordered to Take Medication	7	0.50	0.50	0.60	0.76
	Psychiatric Cond.	31	0.76	0.77	0.84	0.89
	Different Cond. Evaluation	0	0.00	0.00	0.00	0.00
	Attend Counseling Aftercare	6	0.00	0.67	0.55	0.71
	Attend Mental Health Court	0	0.00	0.00	0.00	0.00
	Attend Outpatient Tx Program	0	0.00	0.00	0.00	0.00
Programs	Participate in a Self-Help Group	31	0.87	0.83	0.92	0.85
	Undergo Anger Management	32	0.91	0.90	0.91	0.92
	Misc. or Unknown Program Cond.	34	0.11	0.44	0.55	0.59
	Reentry Program Cond.	7	0.73	0.92	1.00	1.00
	Attend Program for Veterans	0	0.00	0.00	0.00	0.00
	Attend a Behavioral Health Program	8	0.50	0.33	0.50	0.52
	Attend Parenting Program	3	0.50	0.50	1.00	1.00
	Attend Drug Court	4	0.40	0.40	0.86	0.82
	Supervised by Mental Health Agent/Unit	0	0.00	0.00	0.00	0.00
	None specified Tx	3	0.00	0.00	0.00	0.00
Community	Community Service Cond.	20	0.80	0.86	0.97	0.98
	Reentry Into Community Cond.	0	0.00	0.00	0.00	0.00
Victim	Victim-Related Cond.	157	0.93	0.93	0.96	0.97
	Other Victim Related Cond.	0	0.00	0.00	0.00	0.00
	Attend Victim Program	37	0.97	1.00	0.96	0.98
Employment & School	Employment or School Cond.	36	0.91	0.93	0.97	0.97
	Employment or School Program	5	0.00	0.57	0.55	0.64
Probation/Parole Supervision	Curfew Cond.	10	0.17	0.57	0.70	0.54
	General Probation & Parole Cond.	17	0.71	0.65	0.62	0.59
	Undergo Record Check	11	0.90	1.00	1.00	1.00
	Provide DNA	1	0.00	0.00	1.00	0.39
	Appear in Court Cond.	9	0.33	0.57	0.88	0.75
	Pay Fines/Fees Cond.	143	0.92	0.92	0.96	0.97
	Movement Restriction	4	0.75	0.89	0.80	0.96
	Supervision Relocation	5	0.00	0.50	0.60	0.30
	Fines & Fees Waived	8	0.77	0.67	0.93	0.95
	Restitution Cond.	31	0.65	0.75	0.91	0.68
	Gun-Related Cond.	30	0.76	0.82	0.98	0.99
	Allowed to Leave the State	15	0.55	0.67	0.76	0.80
	Non-standard Probation Cond.	59	0.32	0.46	0.62	0.37
	Participate in Probation/Parole Program	3	1.00	1.00	0.86	1.00
	Apology Letter Requirement	1	0.00	0.00	0.67	0.72
Drive	Driving/Driver's License	66	0.98	0.98	0.97	0.99
Sex Offender	Sex Offender Cond.	25	0.89	0.92	0.94	0.99
None & Other	No Special Cond.	1	0.00	0.00	0.00	0.00
	Other	29	0.40	0.43	0.44	0.50
	Undergo an Evaluation	4	0.40	0.33	0.67	0.67
	COVID-19 Cond.	11	0.95	0.95	1.00	1.00
	Family-Related Cond.	2	0.00	0.00	0.00	0.00
	Forfeit Items	24	0.96	0.96	0.98	1.00
	Housing Cond.	0	0.00	0.00	0.00	0.00
	OVERALL MICRO AVG.	**1151**	**0.77**	**0.81**	**0.85**	**0.89**