Federal Register Notice: 89 FR 51554, Federal Register :: Networking and Information Technology Research and Development Request for Information on Digital Twins Research and Development, June 18, 2024.

Request for Information on the National Digital Twins R&D Strategic Plan

Charlotte Alexander

Shany Azaria

Opher Baron

Dmitry Krass

Response to RFI on the Digital Twins Research and Development[1]: Digital Twin to Improve Access to Services - the Case of the U.S. Federal Court System

Shany Azaria
Postdoctoral Fellow
Operations Management and Statistics
Rotman School of Management, University of Toronto

Charlotte S. Alexander
Professor of Law and Ethics and Director, Law, Data, and Design Lab
Scheller College of Business, Georgia Institute of Technology

Opher Baron
Professor of Operations Management and Statistics and Distinguished Professor of Operations Management
Rotman School of Management, University of Toronto
Founder and CEO, SiMLQ, WWW.SiMLQ.com

Dmitry Krass
Professor of Operations Management and Statistics and Sydney C. Cooper Chair in Business and Technology
Rotman School of Management, University of Toronto
Founder and CSO, SiMLQ, WWW.SiMLQ.com

Abstract

We wish to highlight the potential benefit and high feasibility of employing the digital twin approach in services, including public and government service providers. To illustrate these advantages, we focus on the complex set of processes and services embodied by the U.S. federal court system. The court system, pivotal for administration of justice, is often characterized by congestion which affects social welfare, economic development, and access to justice. This highly important system possesses multiple operational complexities, such as constrained resources, uncertain demand, long in-process waiting time, "impatient" clients, and service times measured in years. Nevertheless, empirical research on court operations remains limited, and designing and implementing improvement interventions is extremely complex. For these reasons, we suggest leveraging the digital twin approach to improve access to administration of justice, with a focus on the U.S. federal courts. Our approach employs Natural Language Processing (NLP) and Artificial Intelligence (AI) tools to transform U.S. federal district court case dockets into a detailed "event log," to which process- and queue-mining techniques can be applied to explore judicial congestion and highlight ways to ameliorate it. By scraping, labeling, and analyzing millions of docket entries through an operation management lens, we are able to observe the case flow in the system and assess congestion impacts on case processing. After illuminating the judicial workflow, it is possible to create a

[1] This document is approved for public dissemination. The document contains no business-proprietary or confidential information. Document contents may be reused by the government in the National Digital Twins R&D Strategic Plan and associated documents without attribution.

digital twin of this complex service system, not only to predict changes in the state of the system but also to estimate the effect of various interventions to improve access to justice.

1. Background

The court system is a unique environment, aimed to produce justice. However, it suffers severe congestion worldwide. Drawing an analogy to the healthcare system, another social service system characterized by the "Iron Triangle" of access, quality, and cost containment (Kissick, 1994), congestion in court systems poses a significant threat not only to quality and cost, but also to access to justice. Timely and efficient resolution of disputes is a vital aspect of ensuring access to justice (Vitkauskas and Dikov, 2017).

Unfortunately, the demand for judicial services continues to grow worldwide, resulting in severe congestion and delays (Church et al., 1978; Dakolias, 1999; Decker et al., 2011; CEPEJ, 2015; CEPEJ, 2016; Voigt, 2016). Moreover, evidence from various countries indicates that court congestion is not solely driven by exogenous supply and demand trends but also by other endogenous factors such as inefficient processes, judicial passivity, and mismanagement (Mitsopoulos and Pelagidis, 2010; Dalton and Singer, 2014; Castro and Guccio, 2015; Moffett et al., 2016; Peyrache and Zago, 2016).

The complexity of the court system, as discussed by Azaria and Shamir (2023), can be described on three levels: operational, organizational and incentives. As for the operational complexity, first and foremost, the court system is usually a geographically decentralized system with numerous employees in various positions, serving a wide range of individuals. Secondly, this system is characterized by a high degree of process and procedural variety, with high uncertainty regarding case flow and timing of events within each case.

Moreover, the court system possesses distinctive attributes that set it apart from other complex service systems. Court systems endure limited managerial flexibility, as major changes to processes and procedures must be authorized through legislation or regulation and, in the U.S., may implicate litigants' Constitutional and statutory rights. Within their jurisdiction, judges and court administrators have limited authority to introduce operational changes to the established local routines.

As for the organizational complexity of the court system, it arises from one of the fundamental principles of democracy, which necessitates the independence and autonomy of judges (Tacha, 1995). This independence relates not only to external influences such as government, media, and public opinion, but also to influences within the judicial system. Moreover, interference with case management could even be viewed as a threat to judicial independence (Agmon-Gonnen, 2005). In addition, as in many other organizations in the social sphere, the efficiency of the system relies mainly on the expertise of judges (Christensen and Szmer, 2012), who are traditionally evaluated and rewarded based on judicial quality rather than efficiency (Posner, 2004).

As for the complexities due to incentives, the nature of the court system is its conflicting views and that these views are represented by different parties. The plaintiff and the defendant have

completely different incentives and objectives, the lawyers have complex laws they need to adhere to and often need to explain these to their client, and the court's objective is different altogether. Moreover, the complexities of the process implies that even when incentives are aligned- finding this alignment may be extraordinarily difficult if not impossible.

For the reasons stated above, the management, improvements and optimization of court systems is not only immensely important, but also complex and expensive. For example, when the U.S. federal courts experiment with operational change, they often do so via pilot programs that test various interventions, though pilots can be slow, expensive, and limited in scope (one example of such pilot program can be found in https://www.fjc.gov/content/321837/mandatory-initial-discovery-pilot-project-overview).

In order to address some of these complexities, costs and uncertainty, we propose to employ the digital twin approach to mimic the structure, context, and behavior of the judicial system, based on reliable and accessible data. This unique application will enable multiple channels of improving access to justice: (1) advancing scientific research in law as well as operations management; (2) enhancing operational efficiency by allowing better monitoring of performance and assessing effects of interventions; (3) enabling predictive analytics and effective optimization, and (4) allowing the analysis of impacts of different interventions on the digital twin with less costs and less complexity than a pilot program; this will allow the system to focus pilots on interventions with a high potential benefit.

2. Digital Twin of a Federal District Court

Digital twins are designed to mimic "a natural, engineered, or social system (or system-of-systems)." Similar to its application in manufacturing, where a digital twin mirrors the physical production processes to enhance efficiency, predict failures, and optimize performance, a digital twin in a service system can simulate various scenarios, predict outcomes, and improve decision-making. The uniqueness of applying a digital twin to a service system lies, amongst other things, in the complexity and variability of human interactions, customer behavior, and service delivery processes. Unlike manufacturing, where processes are often more standardized and predictable, service systems require the digital twin to account for dynamic and often unpredictable human and exogenous factors. This necessitates more sophisticated modeling and real-time data integration to capture the nuances of service operations.

The first step in creating a federal court digital twin is data generation. This data must be in the form of an event log, as described below, and should be very granular, reflecting every important event in the system. The Federal Judicial Center produces a database of all cases filed and terminated in the federal district courts. But granular detail on each case's progression through a pathway of litigation events is not recorded. To transform the available data into the case-level event log format required for the approach described here, it is necessary to mine the text of each case's docket sheet. (Docket sheets are the chronological record of a case's litigation activity, available from the federal courts' Public Access to Court Electronic Records (PACER) system.)

The National Science Foundation's Convergence Accelerator program funded the SCALES Open Knowledge Network project (see https://scales-okn.org/), a multidisciplinary, multi-university research group. This group created a public repository of U.S. district court docket sheets along with standardized per-case litigation event labels. In total, the SCALES project holds approximately 1.3 million docket sheets and court documents from criminal and civil cases filed in the 94 U.S. district courts. The SCALES project used an English-based Large Language Model (Microsoft's large DeBERTaV3) trained also on 11 million docket entries and fine-tuned for classification tasks with manually annotated docket entries. This trained model allowed the SCALES team to generate standardized event labels from the unstructured docket text. For example, in one civil case, a docket entry might read, "The plaintiff and defendant have agreed to settle," whereas a docket entry in another case might read, "The parties reached a mutual resolution of all claims." Both entries are reporting the parties' settlement; the SCALES labeling workflow identifies both as 'settlement,' despite the substantial difference in wording.

Leveraging SCALES' standardized litigation event labels, along with event dates and judge information from the docket sheets, enables the creation of event logs that serve as the foundation for generating a digital twin of the court system. For this purpose, we define an "event" as a timestamped occurrence in a case that impacts the flow, with respect to either the routing (meaning, what would be the next occurrence) or the timing of the next occurrence.

An important component of the efficacy of digital twins in the service industry is better planning and design of the data collected, e.g., to make the creation of an event log straight-forward. Modern information systems typically gather transactional data on every customer, event, and activity that is being processed by that system. Such data includes at least a case identifier (unique customer or visit id), the event or activity name, and the event or activity timestamp; a set of additional attributes, such as customer type, resources required, or activity capacity, may also exist. This granular event data is collected by the system and may be stored in various system databases and be dispersed along numerous tables. Event data allows the inference of the process structure and dynamics. Moreover, enriched event data can be transformed into event log data. A high level of data granularity that facilitates the creation of an event log is maintained by many service providers. An example of an event log for the court system is given in Table 1.

Table 1: One case event log (partial)

Case ID	Date	Event	Attribute	Judge	Number of Plaintiffs	Number of Defendants
01-cv-0XXX6	4/12/2017	complaint	opening	SJID0125	7	3
01-cv-0XXX6	4/12/2017	case assigned		SJID0125	7	3
01-cv-0XXX6	4/14/2017	summons issued	scheduling	SJID0125	7	3
01-cv-0XXX6	4/28/2017	summons returned		SJID0125	7	3
01-cv-0XXX6	5/2/2017	status hearing reset	scheduling	SJID0125	7	3
01-cv-0XXX6	5/8/2017	motion for time extension	unopposed	SJID0125	7	3
01-cv-0XXX6	5/8/2017	motion granted	hearing held	SJID0125	7	3
01-cv-0XXX6	5/11/2017	order		SJID1007	7	3
01-cv-0XXX6	6/27/2017	settlement	dispositive	SJID1007	7	3
01-cv-0XXX6	6/28/2017	dismiss without prejudice	dispositive	SJID1007	7	3
01-cv-0XXX6	6/28/2017	case dismissed	dispositive	SJID1007	7	3

The path from the SCALES data to the creation of an event log that can be used for process mining, and later for the creation of a digital twin is not trivial. It requires a panel of multidisciplinary experts from law, data science, operations research and computer science. By interlacing these disciplines, we can construct a toolkit which will eventually transform the natural language html dockets to comprehensive workable even logs. These event logs are the building block of understanding this complex process. These event logs can bridge the gap between the designed process, the process observed by domain experts, the process captured in the raw data, and the real process. After bridging these gaps, a digital twin can be generated and transform the process into the optimal process.

To demonstrate how process mining helps to close these gaps, we created event logs for a complete set of all civil and criminal cases filed in the U.S. District Court for the Northern District of Illinois. (This is the federal district court that encompasses Chicago and the surrounding areas.) In all, our data set covers all cases filed in 2002-2018 and consists of events harvested from the docket sheets of 178,523 cases. The process map of a sample of 1,000 civil cases from one of the courts in our data, created using SiMLQ (see Section 3), is shown in Figure 1. This map demonstrates the complexity of the flow of cases.

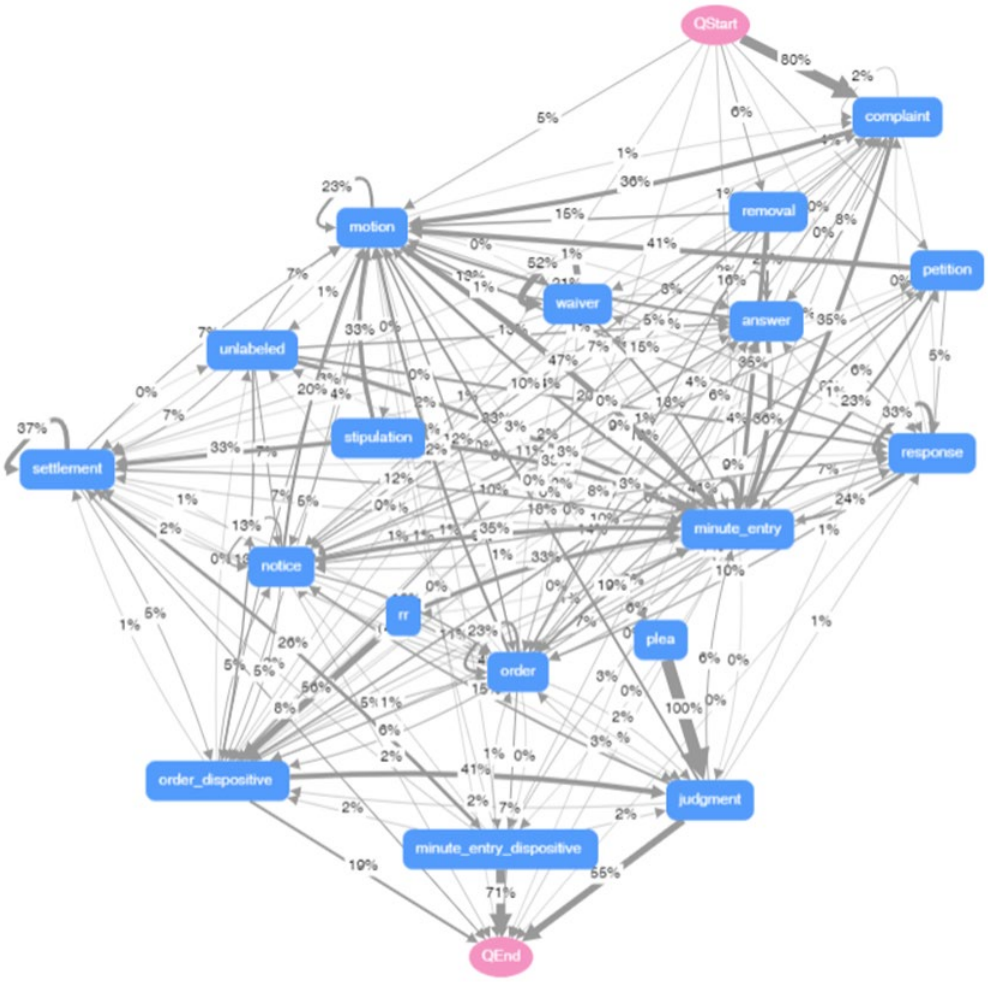

Figure 1: Process map of civil cases sample

In this map, each arrow signifies transition of cases from each event to all the relevant subsequent events. It also shows the portion of cases transitioning to each of the possibilities, which is also demonstrated by the width of the arrow. For example, 80% of the cases in this sample started by a complaint and 2% of them were followed by another complaint.

Now that the service process is mapped, we can combine it with queueing theory models to shed light on some of the interdependencies of this complex system. For example, customers (cases) compete for resources (judges) and not activities, waiting times are generated by the process, not exogenously, and so on. In operations management, the most fundamental method to analyze such complex service networks is simulation. So, we employ queue-mining and machine learning (ML) to create data-driven simulation that can automatically produce a digital twin of the system.

To conclude, using a digital twin in the Federal Court system could improve access to justice and overcome the unique complexities of the system in several aspects:

- A key part of the problem is when decisions at both micro and macro levels are made, the downstream impacts of these decisions on system delays are not visible. Thus parties making these decisions are either completely unaware of these impacts, or are, at best, guessing at what these might be.

- The main advantage of developing a digital twin is that it can make such impacts visible. Thus, a judge deciding whether to grant a party's motion for a time extension on a filing deadline (for example) can check the impact of this decision on the current and other cases. Similarly, when evaluating a system intervention (e.g., changing certain filing procedures) the impacts can be made visible to all affected parties.
- Once the impacts are quantified, an educated decision can be made on whether the expected additional delays are offset by the usefulness of the proposed action (or vice versa – whether the extra costs imposed on the parties in the present case and other cases in the judge's caseload are sufficiently offset by the improvements in Key Performance Indicators (KPIs).
- Experimentation in the U.S. courts often happens through various pilot programs that test different kinds of interventions. Those are expensive, slow, and limited in scope. The digital twin approach promises a much more nimble, real-time alternative to pilot programs.

3. SiMLQ

SiMLQ (see www.SiMLQ.com) is a recently-developed leading edge software that uses ML and AI to quickly generate data-driven simulation models of congested systems, where queueing plays an important role. The court systems example, as well as many other services, such as healthcare, license renewals, contact center, and insurance, are prime examples of such congested systems. The SiMLQ software (patent pending) automatically creates data-driven simulation models of congested systems. SiMLQ uses machine learning to quickly process event logs and generate a high fidelity simulation model of the process that generated this data. Creation of digital twins of congested processes can build upon SiMLQ's capabilities. SiMLQ's initial set-up and training would lead to a well calibrated simulation model of the system. Then, there are two main steps required to enable SiMLQ as a digital twin.

The first step is on the input side. An API should support the continuous upload of recent events into SiMLQ. This upload will update the current system status, allowing the system to depict the current process map and to initiate the simulation run of the current state of the system. Once the simulation model is calibrated and validated, various interventions designed to enhance system performance can be specified and their effect simulated. Then, comparing the impact of different interventions on KPIs in view of the cost of these interventions, would be used to prescribe optimal actions. The newly created data (after implementation of actions) would be immediately uploaded to the digital twin created by SiMLQ allowing it to stay up to date.

The second step is retraining. The continuous data input would keep the digital twin updated by supporting the retraining of the simulation model. SiMLQ's quick training allows it to be retrained much more often than standard MLOPS principles consider. Depending on the actions taken and the speed in which the process changes, different retraining periods, such as daily, can be implemented. Such training, with data generated after actions are implemented, is required to provide full functionality of a digital twin model.

This two-step process would allow SiMLQ to keep its digital twin up-to-date in terms of both capturing and describing the current situation as well as maintaining a current, well trained, digital twin to capture the impact of additional actions on the updated system.

4. Summary

As our discussion on the implementation of digital twins in the U.S. federal court system demonstrates, implementing digital twins in services, including public and governmental services, holds substantial promise for improving access and efficiency. In the court system, digital twins can improve justice by addressing congestion and inefficiencies inherent to the judicial process. This approach leverages AI, ML and NLP to transform court case data into detailed event logs. These logs enable advanced process- and queue-mining techniques, as used in SiMLQ, that can illuminate the flow of cases in the system and highlight issues and ways to address them.

The unique complexity of the court system, characterized by its operational, organizational, and incentive-based challenges, necessitates sophisticated solutions like the digital twin approach. By creating a virtual replica of the court system, stakeholders can simulate different scenarios and make data-driven decisions to enhance efficiency and reduce delays.

Key benefits of the digital twin approach include:

- Operational Efficiency: The ability to simulate interventions before implementing them can lead to more cost-efficient operational changes, as well as improve the efficiency of existing resources.
- Predictive Analytics: The digital twin enables predictive modeling, helping to foresee and mitigate potential issues before they escalate. In many services predicting longer than allowed delays for some customers can support an intervention that expedites their service.
- Cost Reduction: Both direct and indirect costs associated with delays can be minimized, benefiting the broader economy.
- Improved Access to Service: Supporting the delivery of services, such as justice, in a timely fashion improves not only the efficiency of the system but also the efficacy of the service (resolutions in the court system).

The collaborative effort involving legal experts, data scientists, and operations researchers underscores the interdisciplinary nature required to successfully implement digital twins in services. Moreover, the digital twin model offers scalability and adaptability, potentially benefiting public services, such as court systems, worldwide by providing a framework to address similar challenges of congestion and inefficiency. Overall, the adoption of digital twins in services is a forward-thinking strategy that is aimed to drive significant improvements in service delivery. In the judicial system this would be a better administration, ultimately contributing to a more efficient, equitable, and accessible legal system. In the healthcare services this would be an improvement of quality of life and longevity.

Finally, below we briefly address the following topics outlined in the RFI:

- Artificial Intelligence (AI): Machine learning(ML) and AI would support the creation of digital twins for service systems. The ability to use this twin to simulate the impact of different actions on the system and optimize such actions may be an important and effective avenue for using ML&AI in practice. ML&AI would be used in different phases of the creation of the digital twin- from making existing data accessible to the twin, see, e.g., the SCALES project, via the algorithms used to provide data driven representation of the process, and to the algorithms that generate, train, maintain, and analyze the digital twin.
- Business: The "return on investment" in improving access to justice can be observed on two levels - internal and external. First, the direct cost of the justice system can be affected, as improving efficiency can deal with expenditures which are now dedicated to dealing with system inefficiencies. Second, there are indirect costs to lengthy litigation which burden the economy. This means that improving access to justice has the potential to reduce the cost of doing business. Generally, digital twins of a service system can (1) improve access to these services, e.g., in access to healthcare where delays have very high costs, (2) improve equitability of service offerings to different populations, e.g., between rural areas and suburbs, and (3) reduce the cost of providing an adequate service level, e.g., by expediting the implementation of process improvements.
- Data: Our approach requires a well constructed event log which traces the path of every task (i.e. each customer, patient, or litigant, depending on the setting), with the activities and resources involved in each activity. Such data often exists in service systems. Therefore, the data gathering, governance, and planning methodology that would allow for the creation of digital twins in services may be cheaper than those required for the creation of digital twins for physical systems, as the latter often requires additional investments, e.g., in sensors or other recording devices, to collect real time data.
- Ecosystem: Creating a valuable digital twin to a complex system, as the court system, requires the collaboration of researchers and practitioners from several disciplines: legal subject-matter experts, queueing theorists, data scientists, ML & AI experts, and others. The combination of knowledge and complimenting perspectives can push forward this initiative, as well as develop knowledge relevant to other environments.
- International: While court systems worldwide are somewhat different in structure and in nature, they are very similar in dealing with the challenge of congestion. To that extent, comparing court systems and enriching the understanding of the common dynamics has significant potential for mutual benefits. The digital twin approach allows for testing different models based on other systems, without changing the actual system. This benefit of the scalability of digital twins in the service industry may increase their economic value.
- Long Term: Creating digital twins in the service industry is a challenging task that has a very high potential reward. A long term vision where managers of service processes can easily manipulate a digital twin to test the impact of different actions on Key Performance Indicators (KPI) and optimize their process faces several challenges that can only be addressed by combining resources of the government, private companies, and academia. This vision requires appropriate data management and governance, at the personal, interorganizational, country, and global levels (e.g., to support using data from a digital twin in one court system to the one in a different country); adequate computational power

to generate, maintain, and use the digital twins; and an improved understanding of the impact of managerial actions on people (both service providers and consumers) and the data generated by the system (that is than fed into the digital twin).

References

Agmon-Gonnen, M. (2005). Judicial Independence: The Threat from Within. *Israel Law Review*, *38*(3), 120-153. https://doi.org/10.1017/S0021223700012838

Azaria, S. & Shamir, N. (2023). *Operations management in court systems : lessons learned from the delay of Justice*. [online] Tel Aviv University. Available at: https://tau.primo.exlibrisgroup.com/permalink/972TAU_INST/quev9q/alma9933587306104146

Castro, M.F. & Guccio, C. (2015). Bottlenecks or Inefficiency? An assessment of first instance Italian courts' performance. *Review of Law & Economics*, *11*(2), 317-354. https://doi.org/10.1515/rle-2015-0030

CEPEJ (European Commission for the Efficiency of Justice) (2015). *The 2015 EU Justice Scoreboard*. [online] CEPEJ. Available at: https://ec.europa.eu/info/sites/default/files/justice_scoreboard_2015_en.pdf

CEPEJ (European Commission for the Efficiency of Justice) (2016). *European Judicial Systems. Efficiency and Quality of Justice, CEPEJ STUDIES No. 23, Edition 2016 (2014 data)*. [online] CEPEJ. Available at: https://rm.coe.int/european-judicial-systems-efficiency-and-quality-of-justice-cepej-stud/1680786b58

Christensen, R.K. & Szmer, J. (2012). Examining the efficiency of the US courts of appeals: Pathologies and prescriptions. *International Review of Law and Economics*, *32*(1), 30-37. https://doi.org/10.1016/j.irle.2011.12.004

Church, T., Carlson, A., Lee, J.L. & Tan, T. (1978). Justice delayed: The pace of litigation in urban trial courts. *State Court Journal*, *2*(4), 3–8. https://www.ojp.gov/pdffiles1/Digitization/52162NCJRS.pdf

Dakolias, M. (1999). *Court Performance Around the World: A Comparative Perspective*. Washington, DC: The World Bank.

Dalton, T. & Singer, J.M. (2014). Bigger Isn't Always Better: An Analysis of Court Efficiency Using Hierarchical Linear Modeling. *Pace Law Review*, *34*(3), 1169-1189.

Decker, K., Mohlen, C. & Varela, D.F. (2011). *Improving the Performance of Justice Institutions*. Washington, DC: The World Bank.

Kissick, W.L. (1994). *Medicine's dilemmas: infinite needs versus finite resources*. Yale University Press.

Mitsopoulos, M. & Pelagidis, T. (2010). Greek appeals courts' quality analysis and performance. *European Journal of Law and Economics*, *30*(1), 17-39. https://doi.org/10.1007/s10657-009-9128-4

Moffett, K.W., Maltzman, F., Miranda, K. & Shipan, C.R. (2016). Strategic Behavior and Variation in the Supreme Court's Caseload Over Time. *Justice System Journal*, *37*(1), 20-38. https://doi.org/10.1080/0098261X.2015.1067156

Peyrache, A. & Zago, A. (2016). Large courts, small justice!: The inefficiency and the optimal structure of the Italian justice sector. *Omega*, *64*, 42–56. https://doi.org/10.1016/j.omega.2015.11.002

Posner, R.A. (2004). Judicial behavior and performance an economic approach. *Florida State University Law Review, 32*(4), 1259-1280.

Tacha, D. (1995). Independence of the judiciary for the third century. *Mercer Law Review, 46*(2), 645-660.

Vitkauskas, D. & Dikov, G. (2017). *Protecting the Right to a Fair Trial Under the European Convention on Human Rights*, 2nd edition. Council of Europe.

Voigt, S. (2016). Determinants of judicial efficiency: A survey. *European Journal of Law and Economics*, *42*(2), 183–208. https://doi.org/10.1007/s10657-016-9531-6