Increasing Public Data Transparency for Immigration Law in Canada

Ismail (Husain) Bhinderwala, Jessica Yu, Ke Gao, Yichun Liu

2025-05-04

Table of contents

1 Executive Summary

This project aims to improve transparency in Canada's immigration law by analyzing publicly available data on inadmissibility decisions and legal outcomes. Using datasets from Immigration, Refugees and Citizenship Canada (IRCC) and federal court decisions curated by the Refugee Law Lab, we will investigate how individuals are found inadmissible to Canada and how such decisions are reviewed in court.

To uncover patterns and potential biases, which are a key goal of this project, we will explore textual data on court decisions and statistical data on inadmissibility trends using data science methods, including exploratory data analysis, statistical inference, and natural language processing.The final product will be a public-facing dashboard built in Python using Dash, designed to present insights in a clear and interactive way for legal professionals, policymakers, researchers, and the general public. This initiative supports the broader goal of increasing accountability in immigration governance by making complex data more accessible and interpretable.

2 Introduction

2.1 Problem Statement and Importance

Inadmissibility decisions under Canadian immigration law determine who may enter or remain in Canada, based on factors like national security threats, human rights violations, criminal activity, or breaches of the *Immigration and Refugee Protection Act (IRPA)*, such as section A34(1) related to security-based inadmissibility.

While decisions from federal courts are publicly available, they are presented in unstructured formats that limit systematic analysis. In contrast, the data released by IRCC is often raw, inconsistently structured, and lacking contextual detail, making it difficult to access and interpret.

This lack of transparency hinders legal practitioners, public interest groups, and even policymakers from identifying trends, systemic issues, or inconsistencies in immigration decision-making. Without clear, interpretable data, it is difficult to advocate for fairer and more accountable processes.(Shekarian (2025))

To address these challenges, this project combines legal domain knowledge and data science methods. While the data provided by IRCC is often raw and inconsistently structured, data science techniques such as NLP and EDA can effectively uncover patterns and potential biases. Drawing inspiration from Professor Sean Rehaag's research on judicial decisions in refugee and immigration law (Rehaag (2023)), we aim to apply similar methodologies,where useing AI to extract and categorize data from judicial decisions as well as revealing patterns and biases in the decision-making process. For this project, we will audit and analyze the data obtained from

IRCC, referencing Professor Rehaag's code, and generate structured summaries based on the grounds of inadmissibility outlined in the legal introduction. In addition, we will collaborate with Will, who is assisting in obtaining the necessary data, to apply Rehaag's techniques and further enhance transparency in Canada's immigration process.

2.2 Tangible Objectives

This project has four concrete goals:

1. **Analyze IRCC inadmissibility and litigation datasets** to identify trends based on time, country of citizenship, type of decision, and applicant status (temporary or permanent).
2. **Apply legal analytics** which combines legal knowledge with data science techniques,to federal court decisions involving inadmissibility.We will use modern NLP methods to extract meaningful case-level information such as outcomes, judges, and legal reasoning from unstructured court decision texts. This helps us uncover trends, biases, and data gaps that affect the fairness of decision-making.
3. **Develop a public-facing dashboard** using Dash (a Python framework) to allow users, legal professionals, policymakers, and others, to explore key trends and findings interactively.This dashboard helps uncover patterns and insights related to inadmissibility decisions, ultimately supporting fairer and more informed decision-making in immigration cases.
4. **Promote open and interpretable data use** by transforming difficult-to-access raw data into structured, contextualized insights.

2.3 Final Data Product

The final deliverable will include:

- A **web-based interactive dashboard**, allowing users like lawyers,policymakers, and data scientists to filter and visualize patterns in immigration inadmissibility and court decisions.
- **Documentation** describing the data sources, analytical methods, and key limitations.
- **Reproducible Python scripts** for transparency and future use by researchers or advocacy groups.

3 Data Science Techniques

This section outlines the datasets, analytical methods, and evaluation criteria used in the project. By combining structured administrative records with unstructured legal texts, we aim to systematically analyze patterns in Canadian inadmissibility decisions.

3.1 Data Sources

The project will draw on three key datasets:

1. **IRCC A34(1) Refusals (2019–2024):**
   An excel file, containing records of applicants who are refused entry under section A34(1) of the **Immigration and Refugee Protection Act (IRPA)**, which relates specifically to security-based inadmissibility. As shown in **Figure 1**, the dataset includes the number of applicants refused entry under section A34(1) of IRPA by country of citizenship and residency status. The full dataset spans from 2019 to 2024.

Country of Citizenship	Permanent Resident														Total
	COR Not Canada							COR Canada							
	2019	2020	2021	2022	2023	2024	Total	2019	2020	2021	2022	2023	2024	Total	
A34(1)	**7**		**2**	**3**	**2**		**14**	**7**						**7**	**21**
Afghanistan	1			1	1		3								3
Argentina								1						1	1
Egypt	1						1								1
Eritrea			1				1								1
Haiti					1		1								1
India				1			1								1
Iran	1						1								1
Pakistan	1						1								1
Philippines								1						1	1
Romania								1						1	1
Sri Lanka	1						1								1
Syria	2		1	1			4								4
United Kingdom and Overseas Territories								1						1	1
United States of America								3						3	3
A34(1)(a)	**7**			**6**	**3**	**1**	**17**	**5**		**1**			**2**	**8**	**25**

Figure 1: Distribution of Applicants Refused Under Section A34(1) of IRPA (2019–2024)

2. **IRCC Litigation Applications (2018–2023):**
   This structured dataset, provided in Excel,captures federal court applications challenging immigration decisions. Each row represents an aggregate litigation record based on unique combinations of decision year, country, and decision type. The dataset includes key variables such as the year of the leave decision, country of citizenship, case outcome (e.g., "Allowed", "Dismissed at Leave"), case type (e.g., "Removal Order", "Mandamus"), who filed the case ("Person Concerned"), tribunal type (usually "Federal Court"), geographic office, and the number of litigation instances for each combination of fields.

3. **Canadian Legal Decisions (2001–2024):**
   An unstructured dataset of legal texts from federal court and tribunal decisions related to immigration, compiled by the Refugee Law Lab(Lab (2025)).Our analysis will filter this dataset to specifically examine inadmissibility cases, focusing on the grounds for inadmissibility under the Immigration and Refugee Protection Act (IRPA) and excluding refugee claims. This will allow us to concentrate on cases related to national security, criminality, and other security-based inadmissibility grounds.

Each dataset provides a different perspective, administrative and judicial, allowing a well-rounded examination of inadmissibility in Canada.

3.2 Analytical Approach

3.2.1 Data Preparation and Quality Checks

Before conducting any analysis, we will assess each dataset's completeness and clarity. This involves:

- Identifying missing or inconsistent entries (e.g., inconsistent country names).
- Reviewing metadata to understand how variables were defined.
- Verifying that categories (e.g., outcomes, statuses) are clearly and consistently applied.

This step ensures that both legal and data science conclusions are grounded in well-understood data. By following a structured data preparation and quality check process, along with performing "**sanity checks**" **with domain experts**, we will produce an **informal data quality audit**. This will help domain experts interact more precisely with IRCC, **identify potential data inconsistencies**, and **address issues** that may impact **transparency** and decision-making accuracy.

3.2.2 Structured Data Analysis (IRCC Datasets)

For the two IRCC datasets, we will:

- **Restructure the data** into a consistent format (also known as "tidy" data).(Shekarian (2025))
- **Use data visualization tools** (bar charts, heatmaps, and trend lines) to explore:
  - Differences in litigation counts by country or region.
  - Disparities between temporary vs. permanent applicants
  - Differences in decisions made by Regional office vs. Other offices
  - Potential biases in decision outcomes based on applicant demographics or country of origin.
- **Form hypotheses** regarding whether certain countries or applicant groups exhibit higher rates of inadmissibility. As shown in **Figure 2**, the differences in litigation counts by country before, during, and after the COVID-19 pandemic suggest potential trends and disparities that warrant further investigation.

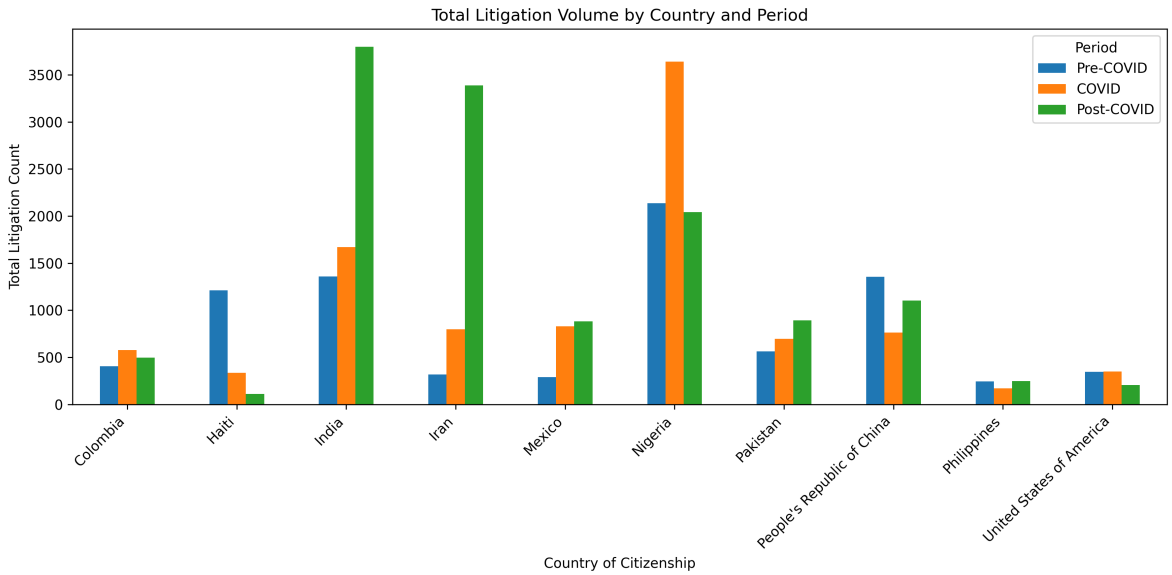

Figure 2: Differences in Litigation Counts by Country Before, During, and After COVID-19

While this analysis cannot establish causality, and it remains uncertain whether statistical inference can be reliably performed, it can highlight patterns that warrant further legal or policy investigation. Acknowledging these limitations is crucial, as it reflects the need for careful interpretation and further validation in future work.

3.2.3 Unstructured Legal Text Analysis

For the court decision texts from "Refugee-Law-Lab / Canadian-Legal-Data - Datasets at Hugging Face" (n.d.), we will:

- **Classify cases** based on the type of inadmissibility using regular expression by keyword-matching.
- **Extract key information** such as judge names, outcome (granted or denied), and city of filing using large language models (LLMs).
- **Analyze patterns** across judges, regions, or years to detect potential biases or inconsistencies in rulings.

A stretch goal includes **semantic analysis**, examining the reasoning within judgments to see how legal language evolves over time or varies by case type.

3.2.4 Comparative Analysis

Across all datasets, we will conduct basic time-based trend analysis to identify:

- Whether inadmissibility findings have increased or decreased over time.
- If certain court outcomes correspond with changes in policy or global events.
- How legal decisions align (or diverge) from administrative trends.

3.2.5 Evaluation Metrics and Success Criteria

To assess whether the project has met its goals, we will use the following criteria:

General Metrics:

- **Coverage Metrics**: Percentage of cases successfully categorized from legal text.
- **Accessibility**: Can both legal and non-technical users interact with and understand the dashboard?
- **Reproducibility**: Are our data processing and analysis steps transparent and replicable

Specific Goals:

- **Partner Expectation**: A functional, clear dashboard and evidence-based insight into inadmissibility trends.

We will also gather feedback from mentors and the capstone partner to ensure that the final product aligns with stakeholder expectations.

4 Timeline

The project runs from **April 28 to June 25**, spanning 8 weeks. A runnable version of the data product (interactive dashboard) will be ready by **June 9**, followed by a refinement phase leading to final submission.

Week	Dates	Task Description
Week 1	Apr 28 – May 4	Set up environment and version control (Dash, Quarto, GitHub); review datasets
Week 2	May 5 – May 11	Perform EDA on IRCC datasets; clean and filter legal text dataset
Week 3	May 12 – May 18	Extract metadata from legal decisions; continue EDA on litigation data
Week 4	May 19 – May 25	Integrate court data with dashboard; finalize IRCC analysis
Week 5	May 26 – June 1	Finalize dashboard layout; implement initial visuals and filters
Week 6	June 2 – June 8	Complete runnable dashboard draft and review internally

Week	Dates	Task Description
Week 7	June 9 – June 15	Incorporate mentor/partner feedback; refine dashboard and visualizations
Week 8	June 16 – June 25	Final documentation, QA, polish, and formal submission

Parallel Work: IRCC and legal datasets will be analyzed concurrently to ensure timely delivery.

References

Lab, Refugee Law. 2025. "Luck of the Draw III: Canadian Immigration Law Analysis." https://github.com/Refugee-Law-Lab/luck-of-the-draw-iii/tree/main.

"Refugee-Law-Lab / Canadian-Legal-Data - Datasets at Hugging Face." n.d. https://huggingface.co/datasets/refugee-law-lab/canadian-legal-data.

Rehaag, Sean. 2023. "Luck of the Draw III: Using AI to Extract Data about Decision-Making in Federal Court Stays of Removal." *Queen's LJ* 49: 73.

Shekarian, Siavash. 2025. "IRCC: Better Immigration Policy for Democracy." https://www.linkedin.com/posts/siavashshekarian_ircc-betterimmigrationpolicy-democracy-activity-7325134722454466561-cwSJ?utm_source=share&utm_medium=member_desktop&rcm=ACoAADi1H8IBc6VZz76Ro-UHLeNQAv4cCb8xzc0.