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 aim to address the lack of transparency and accessibility in Canadian immigration inadmissibility decisions. The data provided by IRCC is poorly structured and lacks documentation, while court decisions are primarily textual and difficult for legal professionals to analyze efficiently.

To tackle these issues, we apply data science methods, including data cleaning, exploratory analysis, statistical inference, and natural language processing, to extract and structure key information. Our goal is to uncover patterns and potential biases in decision-making. The final output will be a publicly accessible, interactive dashboard built in Python using Dash, aimed at informing legal professionals, policymakers, and the public, and promoting greater accountability in immigration governance.

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.

Federal court decisions, though publicly available, are presented in unstructured formats that limit systematic analysis. IRCC data, meanwhile, must be requested through a formal process with a fee and is often raw, inconsistently formatted, and lacking context. These barriers hinder legal practitioners, public interest groups, and policymakers from identifying patterns or advocating for fairer immigration processes (Shekarian (2025)).

To address these challenges, this project integrates legal domain expertise with data science methodologies. We will clean and assess the quality of the IRCC data to enable further analysis and to provide recommendations for improving data transparency and usability. For the analysis of textual court decisions, we will adapt Professor Sean Rehaag's methodology (Rehaag (2023)), which employs regular expressions and large language models (LLMs) to categorize cases into orders, notices, and judgments, and to extract key details such as case outcomes and presiding judges. The processed data is stored in a tidy format, enabling effective exploratory data analysis and statistical inference. While Professor Rehaag focused on stay of removal cases, our analysis will center on inadmissibility decisions in the context of immigration. Additionally, we will collaborate with Wei William (Will) Tao, who assisted in obtaining the necessary IRCC data, to apply and extend these techniques in support of increased transparency in Canada's immigration system.

2.2 Tangible Objectives

This project has four concrete goals:

1. **Analyze IRCC inadmissibility and litigation datasets** to identify trends based on country of citizenship, type of decision, and applicant status (temporary or permanent).
2. **Combine legal knowledge with data science techniques** and apply it on 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. By making the dashboard publicly accessible, it will encourage the **use of open and interpretable data** by transforming raw, complex, and difficult-to-access datasets into structured and contextualized insights. This will support greater transparency in immigration decision-making. Additionally, the project includes an evaluation of IRCC data quality, identifying inconsistencies and recommending areas for improvement.

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.
- A **report** documenting 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 that will guide the project. By providing tools to systematically analyze both structured administrative records and unstructured legal texts, we aim to uncover 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 (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, human rights and other 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).
- **Use data visualization** (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.

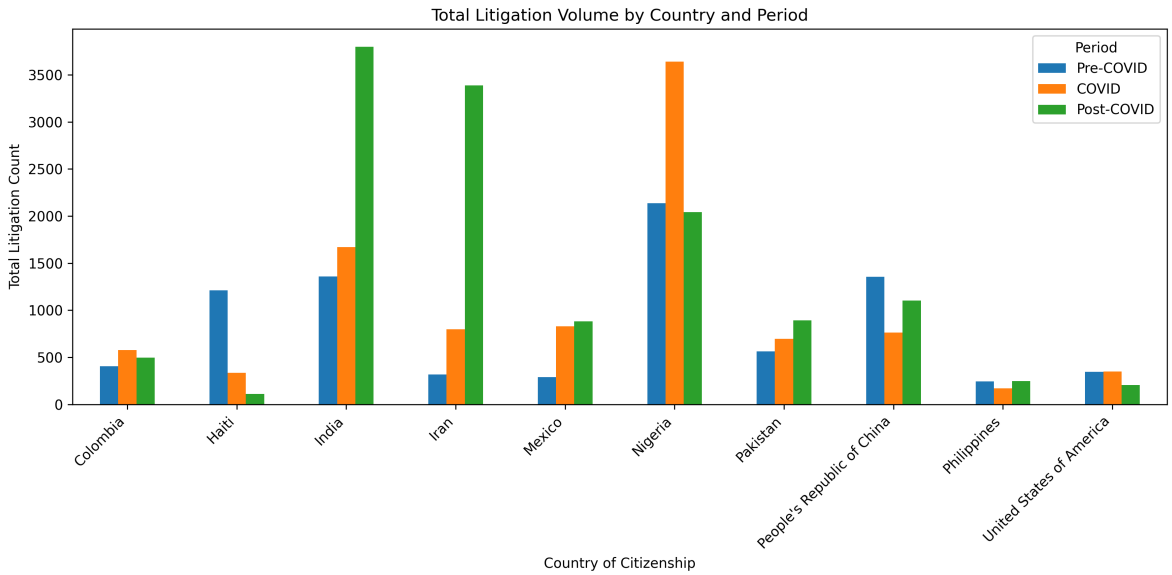

Figure 2: Differences in Litigation Counts by Country Before, During, and After COVID-19

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 (2025), 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 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.
  - Accuracy and precision of extracted information (e.g., judge names, outcomes) from federal court decisions.
- **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

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

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** (8 weeks). A working version of the interactive dashboard will be available by **June 9**, followed by a refinement and documentation phase. Workstreams on IRCC administrative data and federal court decisions will proceed **in parallel**, with responsibilities distributed among team members to maximize efficiency. Feedback loops with our partner will be integrated at multiple stages to ensure relevance, usability, and accuracy.

Week	Dates	Planned Activities
Week 1	Apr 28 – May 4	Set up development environment (Dash, Quarto); configure GitHub for version control. Review both IRCC and court datasets. Define initial schema for structured outputs. Align on scope with partner.
Week 2	May 5 – May 11	Begin parallel EDA: clean IRCC data (identify missing fields, inconsistencies) and pre-process court decisions (filter relevant cases). Develop evaluation framework and metadata structure.
Week 3	May 12 – May 18	Implement NLP pipeline to extract case-level metadata (e.g., outcome, judge) from legal texts. Continue IRCC analysis (grouping by country/status). Share preliminary findings with partner for feedback.
Week 4	May 19 – May 25	Build integration layer for dashboard. Merge cleaned data sources into backend structure. Validate early figures and metrics with partner. Draft summary of data quality (IRCC).

Week	Dates	Planned Activities
Week 5	May 26 – June 1	Develop dashboard interface: implement visualizations (filters by country, time, applicant type). Begin writing supporting text (methods, limitations).
Week 6	June 2 – June 8	Complete a **runnable dashboard prototype**. Conduct internal QA and usability check. Prepare structured summary of key insights for review. Share full preview with partner.
Week 7	June 9 – June 15	Incorporate mentor and partner feedback into visualizations and layout. Refine analysis outputs (especially court case categorization and trend interpretation). Improve dashboard interactivity.
Week 8	June 16 – June 25	Final QA, accessibility testing, documentation, and packaging for formal submission. Prepare reproducibility guide. Conduct final partner check-in.

4.1 Stretch Goals

If time allows, we will explore grouping court decisions based on similarities in language, tone, or legal reasoning. This may involve using basic sentiment analysis or clustering techniques to identify common patterns across cases. The goal is to better understand variations in how inadmissibility decisions are written and reasoned, which could reveal useful trends or inconsistencies.

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.

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.