METODOLOGÍAS DE VINCULACIÓN PROBABILÍSTICA: DESAFÍOS Y OPORTUNIDADES PARA EL SISTEMA ESTADÍSTICO NACIONAL CHILENO

Mayo 2023

Prof Katie Harron, UCL Great Ormond Street Institute of Child Health

Dr. Nicolás Libuy, UCL Social Research Institute, Centre for Longitudinal Studies

¿POR QUÉ SE VINCULAN DATOS?

VINCULACIÓN DE REGISTROS PARA DATOS DE SALUD

Cada persona del mundo crea un Libro de la Vida.

Este Libro comienza con el nacimiento y termina con la muerte.

En sus páginas se recogen los acontecimientos de la vida.

Se denomina vinculación de registros al proceso de reunir las páginas de este Libro en un volumen.

Dunn, 1946

“DE LA CUNA A LA TUMBA”

Atención comunitaria

Consulta médica general

Admisión hospitalaria

Consulta médica general

NACI-MIENTO

MUERTE

Maternidad

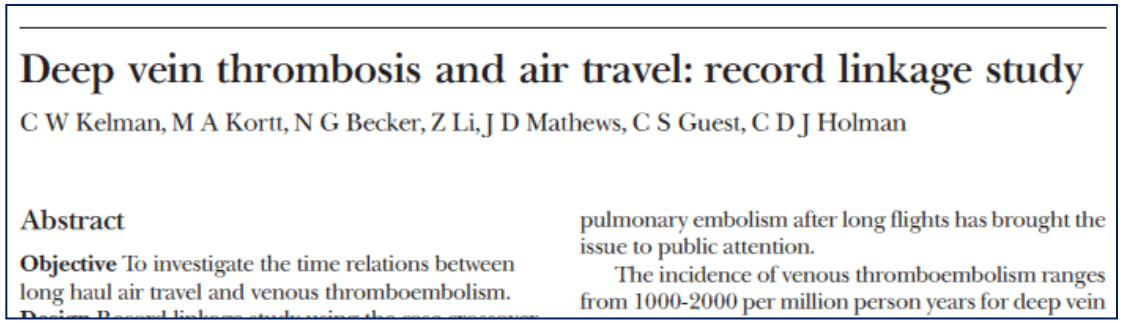
Supervisión de la salud infantil

Prescripción

Accidente & emergencia

Maternidad

Registro de cáncer



Deep vein thrombosis and air travel: record linkage study

C W Kelman, M A Kortt, N G Becker, Z Li, J D Mathews, C S Guest, C D J Holman

Abstract

Objective To investigate the time relations between long haul air travel and venous thromboembolism.

pulmonary embolism after long flights has brought the issue to public attention.

The incidence of venous thromboembolism ranges from 1000-2000 per million person years for deep vein

Datos electrónicos sobre llegadas y salidas de vuelos

Datos sobre hospitaliza-ciones

Addiction

RESEARCH REPORT

doi:10.1111/j.1360-0443.2012.04066.x

A record-linkage study of drug-related death and suicide after hospital discharge among drug-treatment clients in Scotland, 1996–2006

Elizabeth L. C. Merrall[1], Sheila M. Bird[1,2] & Sharon J. Hutchinson[2,3]

Registros de tratamientos contra las drogas (Base de datos escocesa sobre el abuso de drogas)

Defunciones (ICD, Escocia) episodios en hospital (GROS, Escocia), diagnósticos de hepatitis C (Health Protection, Escocia)

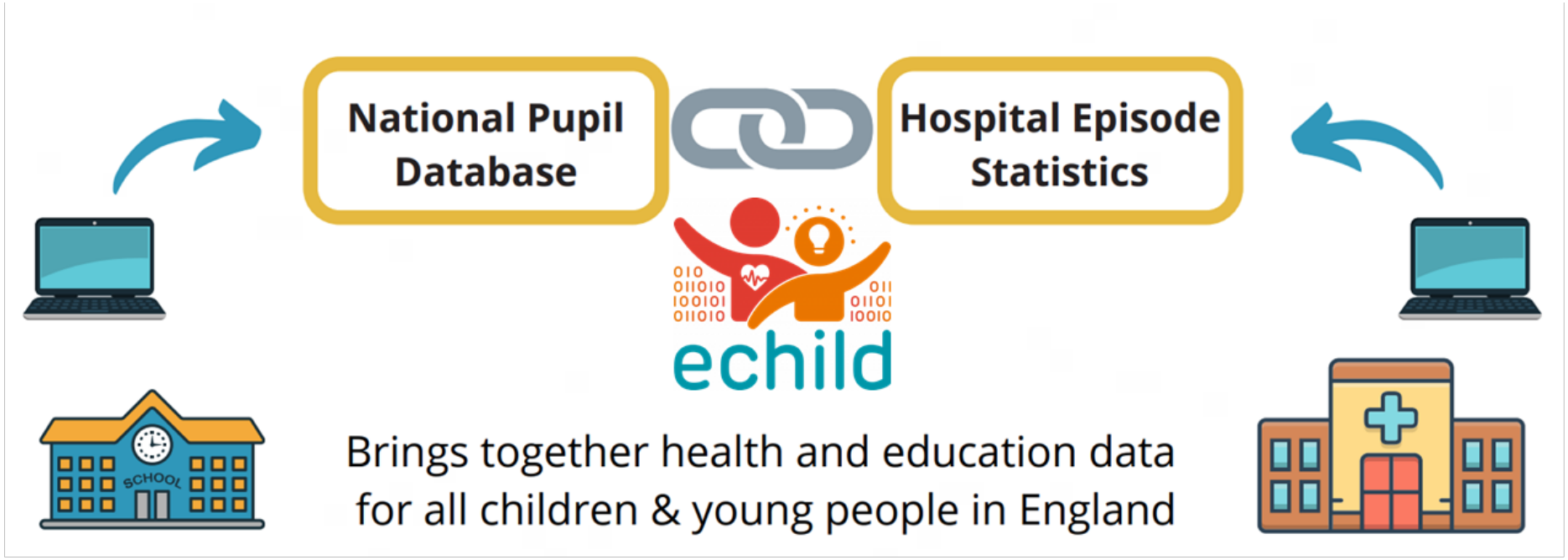
Teenage Pregnancy in England

CAYT Impact Study: Report No. 6

Claire Crawford
Jonathan Cribb
Elaine Kelly

Base de datos nacional de alumnos (NPD, Inglaterra)

Datos sobre concepciones de la Oficina Nacional de Estadística (ONS)



The ECHILD Database is

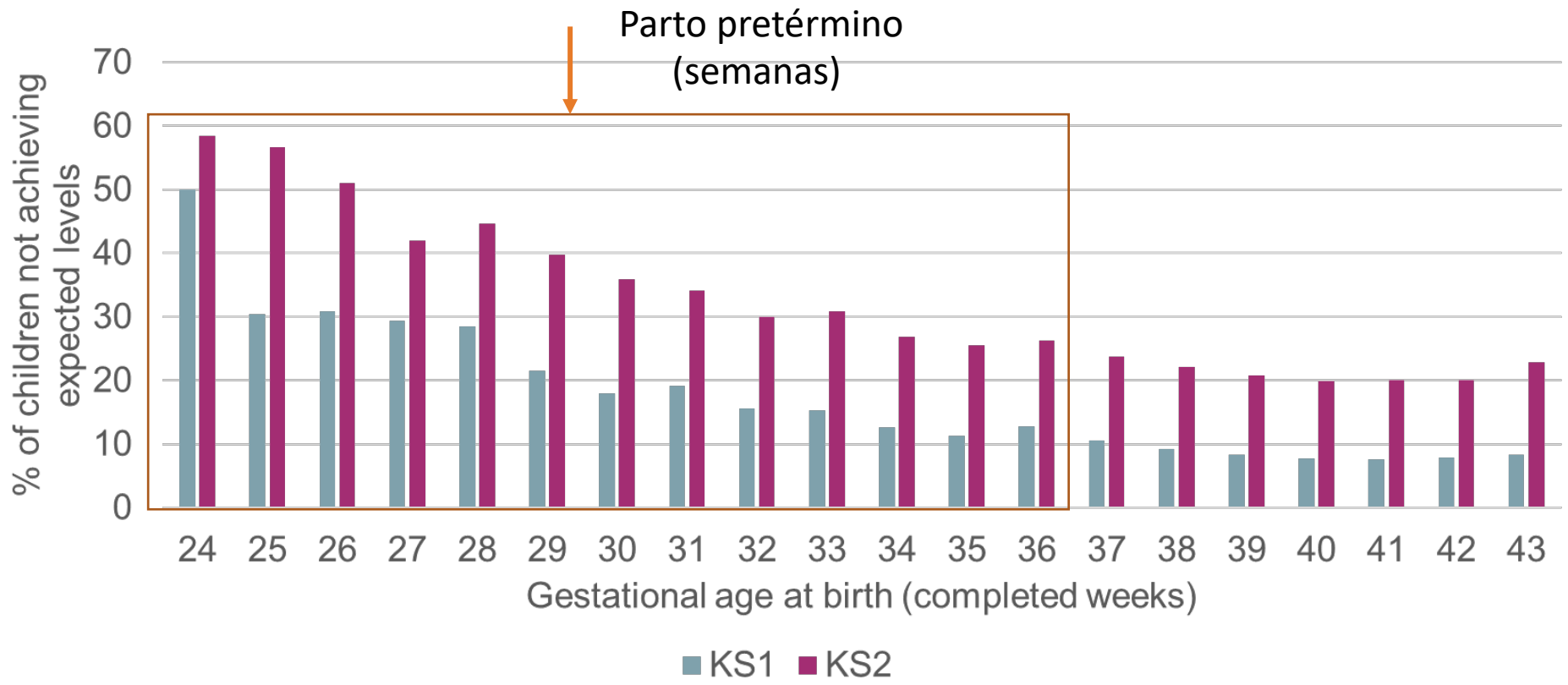
DE-IDENTIFIED

Linked data for

14.7 million pupils

Information from birth to age 24

Cómo se relacionan las características del nacimiento con los resultados del desarrollo



Libuy et al. Gestational age at birth, chronic conditions and school outcomes: a population-based data linkage study of children born in England, *International Journal of Epidemiology*, 2022; dyac105

Este trabajo se ha elaborado utilizando datos estadísticos de la ONS. El uso de los datos estadísticos de la ONS en este trabajo no implica la aprobación de la ONS en relación con la interpretación o el análisis de los datos estadísticos. Este trabajo utiliza conjuntos de datos de investigación que pueden no reproducir exactamente los agregados de las Estadísticas Nacionales.

PRINCIPALES DESAFÍOS PARA LA VINCULACIÓN DE DATOS

Privacidad y seguridad:
reunir más información sobre una persona aumenta el riesgo de identificación

Calidad de los datos:
falta de un identificador único / información personal exacta o completa / diferencias en el formato de los datos

¿CÓMO SE HACE?

COMPARACIÓN DE REGISTROS

Registro	Nombre	Apellido	Fecha Nac.	Sexo
1	Jane	Smith	6/7/1984	F
2	Jane	Smyth	7/6/1984	[falta dato]

Patrón de concordancia	Concuerda	No concuerda (pero similar)	No concuerda (pero similar)	Falta dato
Patrón de concordancia	1	0	0	FD

MÉTODOS DETERMINISTAS

Lista secuencial de normas

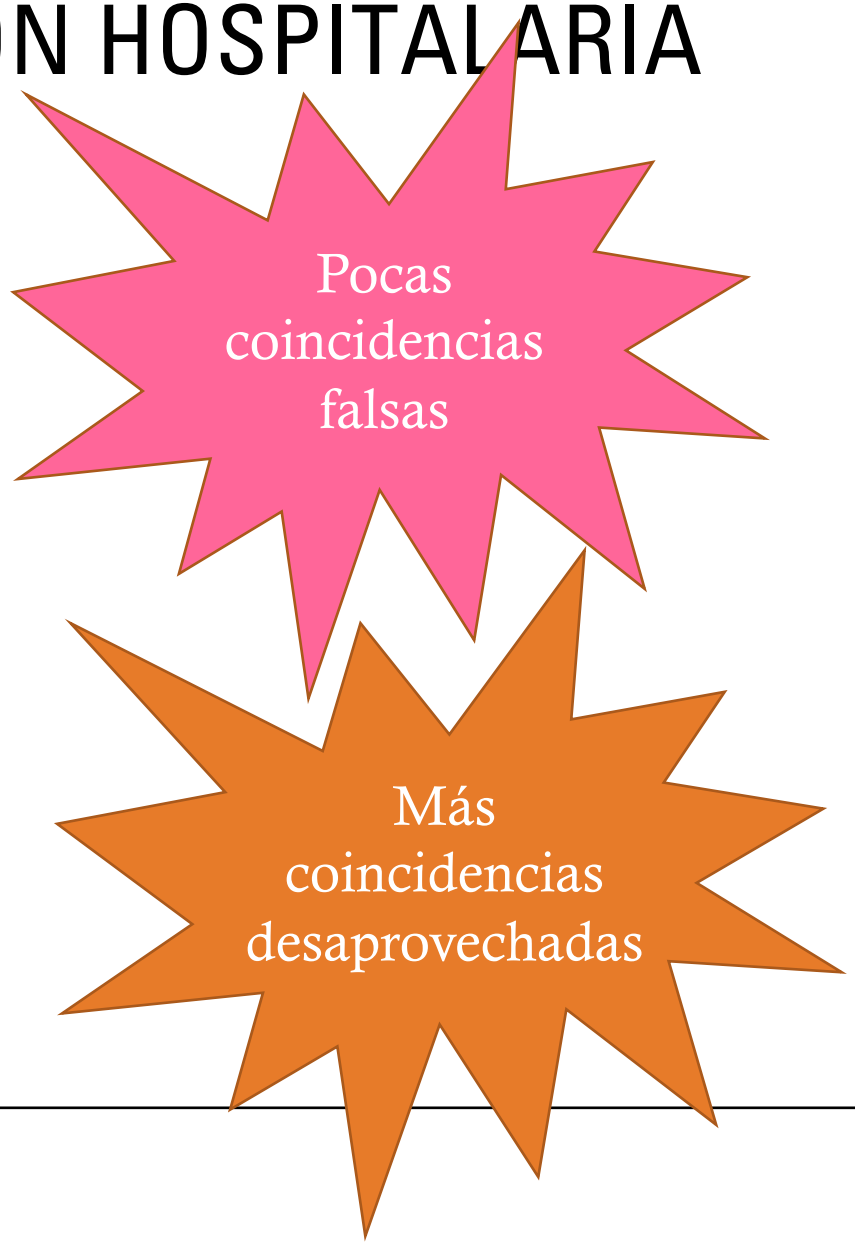
- Útil con identificadores únicos bien completados
- Fácil y rápido de aplicar
- Permite una pequeña cantidad de errores tipográficos preestablecidos
- Los criterios cumplidos pueden registrarse para cada par de comparación, lo que puede proporcionar una indicación de la confianza en el vínculo

EJEMPLO: DATOS SOBRE ADMISIÓN HOSPITALARIA

1
- Sexo
- Fecha de nacimiento
- Numero, servicio de salud

2
- Sexo
- Fecha de nacimiento
- Código postal
- Número, hospital

3
- Sexo
- Fecha de nacimiento
- Código Postal



ERRORES DE VINCULACIÓN

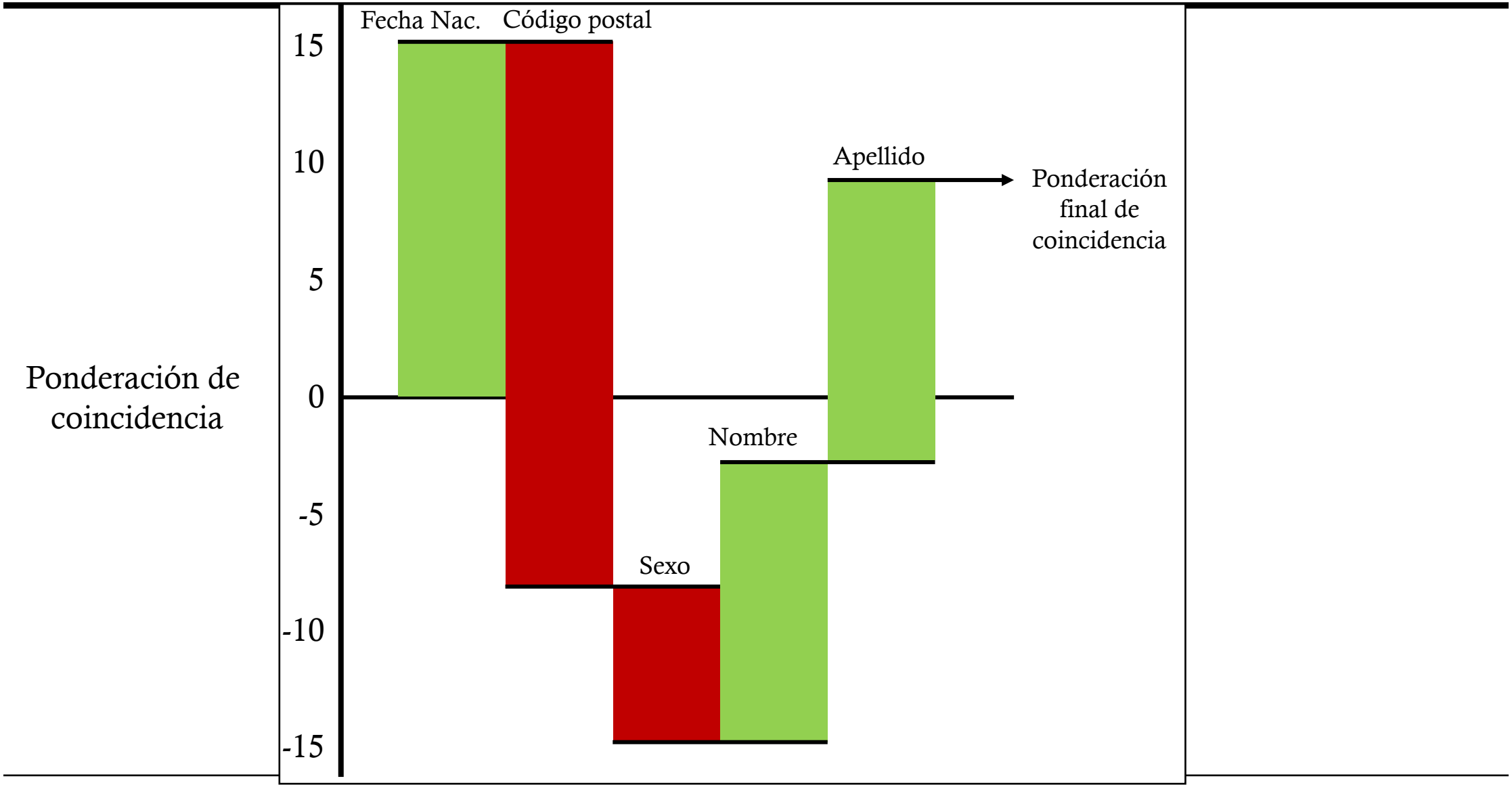
		Estado de coincidencia	
		Coincidencia (un par de la misma entidad)	Discrepancia (un par procedente de entidades diferentes)
Estado de vínculo	Vínculo	Coincidencia identificada	Coincidencia falsa
	Sin vínculo	Coincidencia desaprovechada	Discrepancia identificada

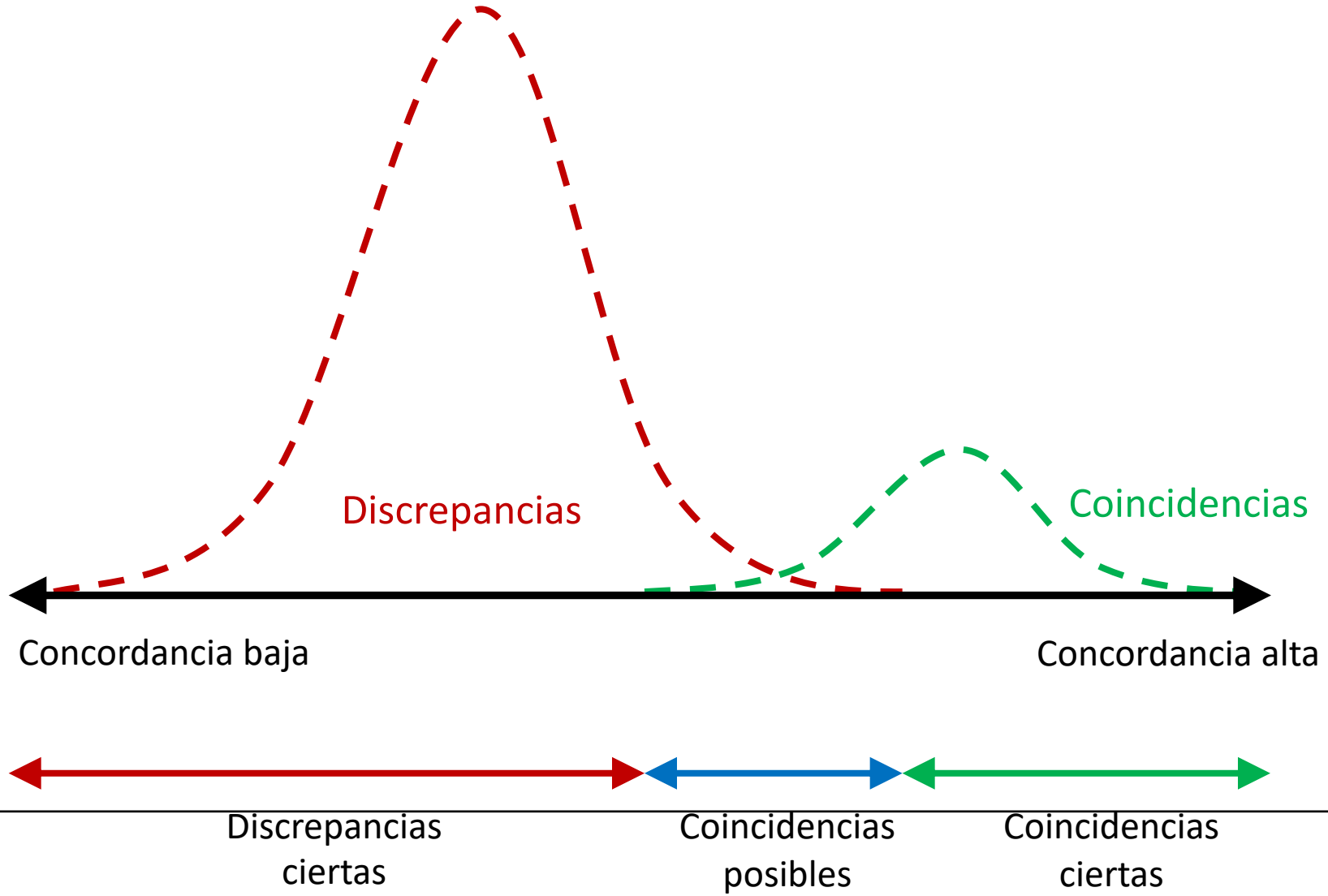
LÍMITES DE LA VINCULACIÓN DETERMINISTA

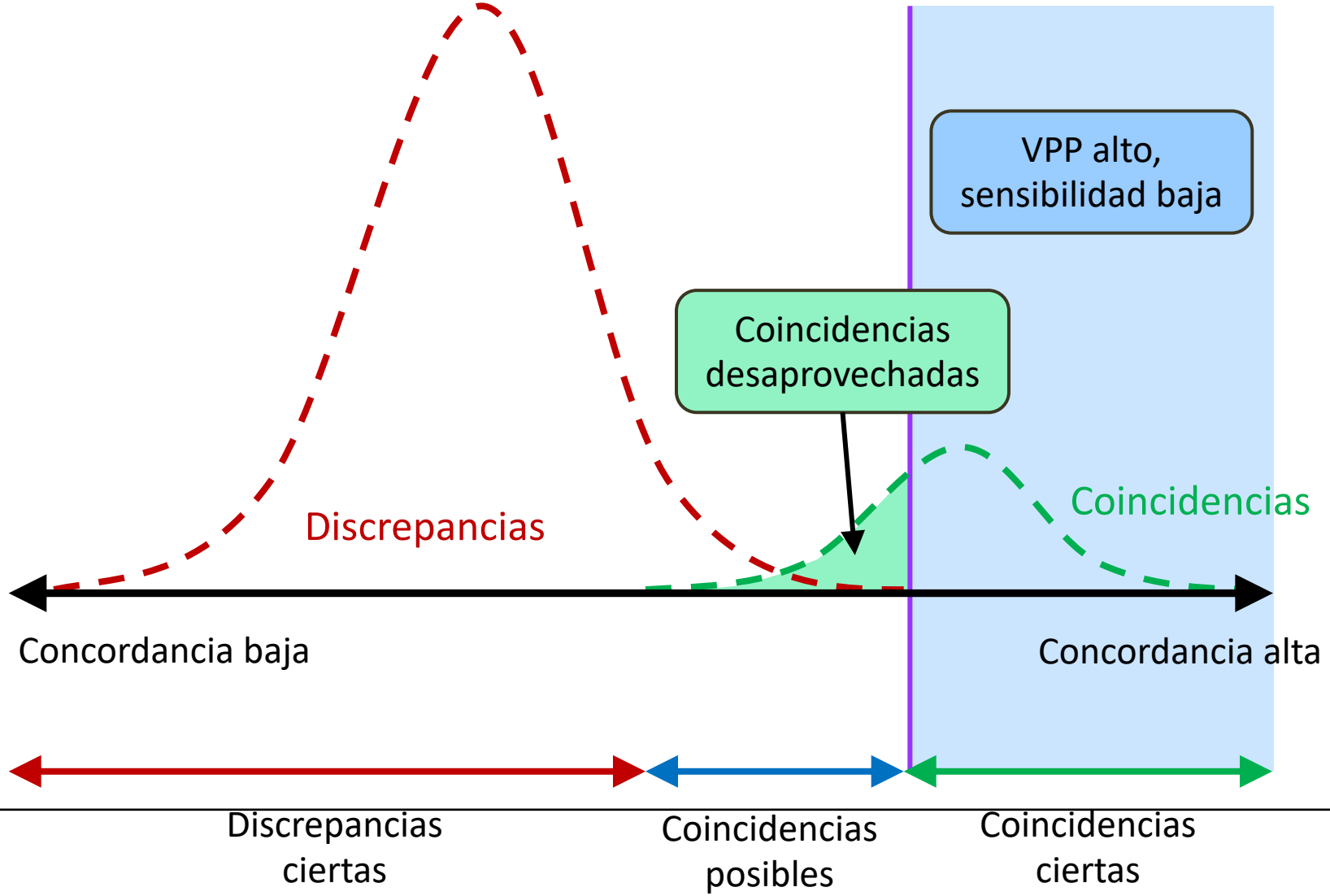
- Cuanto peor sea la calidad de las variables de concordancia, más variables se necesitarán. Con más variables de concordancia, el número de reglas posibles puede llegar a ser enorme.
- Con mediciones de concordancia no binarias, llega a ser infinito…
  - Concordancia parcial
    - Comparators de cadenas (¿Qué similitud hay entre "Smith" y "Smyth"?)
  - Concordancia en función del valor (basada en la frecuencia)
    - Concordancia sobre "Zhang" frente a concordancia sobre "Smith”
- No es fácil clasificar u ordenar las normas.

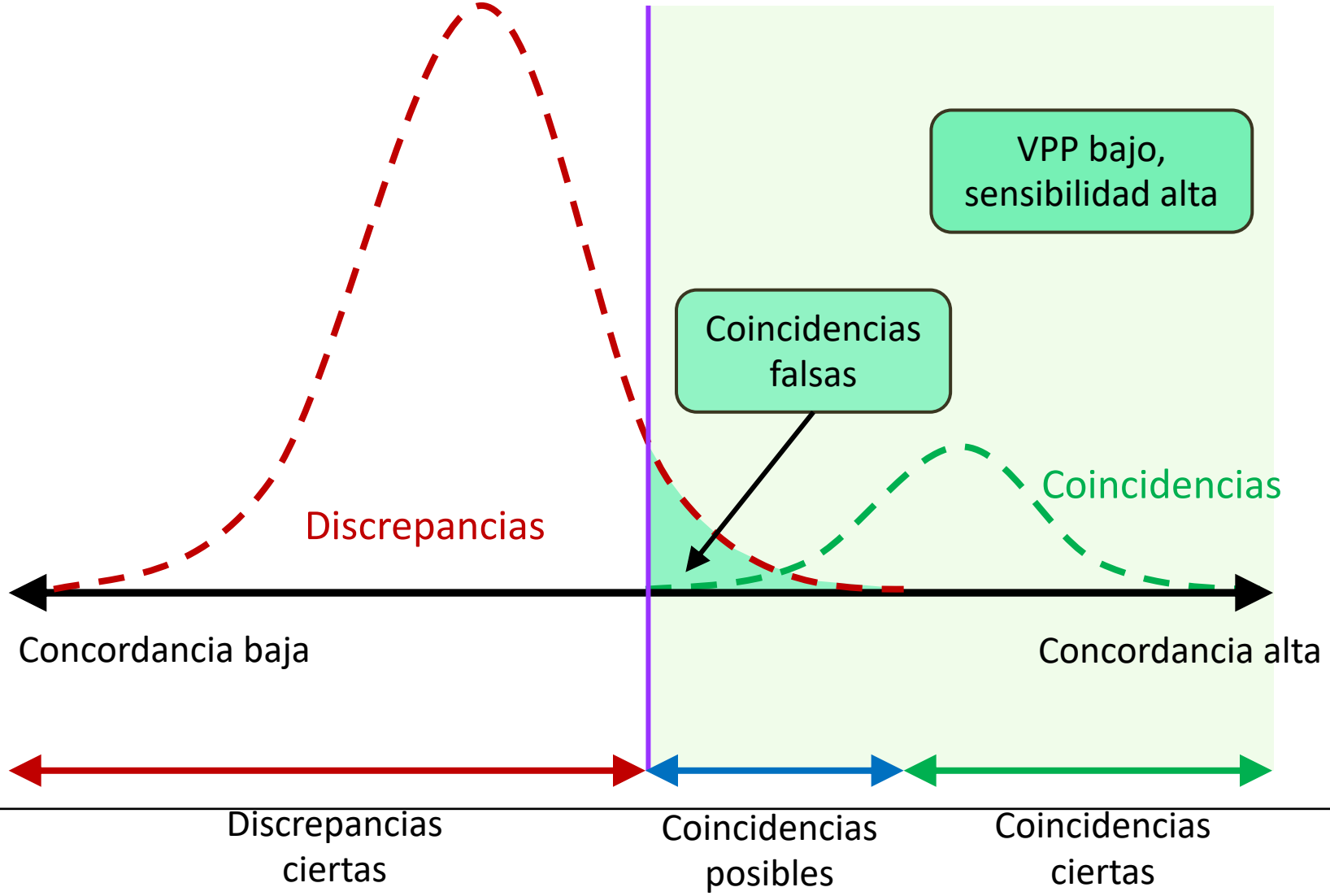
…PERO TENEMOS VINCULACIÓN PROBABILÍSTICA

- Propuesto por primera vez por Newcombe en 1959
- Utiliza las probabilidades de que dos registros coincidan en identificadores especificados, dado que coinciden, y dado que no coinciden
- Estas probabilidades se utilizan para obtener una **ponderación de coincidencia** para cada par de comparación
- La ponderación de las coincidencias tiene en cuenta el grado de diferenciación de un determinado identificador y la precisión con la que se registra


Ponderación de coincidencia
15
10
5
0
-5
-10
-15
Fecha Nac.
Código postal
Sexo
Nombre
Apellido
Ponderación final de coincidencia


Discrepancias
Coincidencias
Concordancia baja
Concordancia alta
Discrepancias ciertas
Coincidencias posibles
Coincidencias ciertas


VPP alto,
sensibilidad baja
Coincidencias
desaprovechadas
Discrepancias
Coincidencias
Concordancia baja
Concordancia alta
Discrepancias
ciertas
Coincidencias
posibles
Coincidencias
ciertas


VPP bajo,
sensibilidad alta
Coincidencias
falsas
Discrepancias
Coincidencias
Concordancia baja
Concordancia alta
Discrepancias
ciertas
Coincidencias
posibles
Coincidencias
ciertas

RESUMEN

- La vinculación **determinista** utiliza reglas:

  SI [patrón] ENTONCES [vincular/no vincular]

- La vinculación **probabilística** asigna ponderaciones (puntuaciones) a las coincidencias, por ejemplo:

	Nombre	Fecha Nac.	Sexo	Ponderación coincidencia
Patrón 1	Concuerda	Concuerda	Concuerda	20
Patrón 2	Concuerda	Falta	Concuerda	10
Patrón 3	Falta	Concuerda	Discrepa	0
Patrón 4	Discrepa	Discrepa	Discrepa	-20

- Ambos tipos de vinculación tienen muchas variantes y a menudo se combinan

Doidge, J. and K. Harron (2018). "Demystifying probabilistic linkage." Int J Popul Data Sci **3**(1).

¿CÓMO EVALUAMOS LA CALIDAD?

MEDICIONES DEL ERROR DE VINCULACIÓN (IDEAL)

		Estado de coincidencia	
		Coincidencia (un par de la misma entidad)	Discrepancia (un par procedente de entidades diferentes)
Estado de vínculo	Vínculo	Coincidencia identificada a	Coincidencia falsa b
	Sin vínculo	Coincidencia desaprovechada c	Discrepancia identificada d

Sensibilidad (proporción de coincidencias vinculadas correctamente)=a/(a+c)

Especificidad (proporción de discrepancias no vinculadas)=b/(b+d)

Tasa de concordancia = proporción de registros de un conjunto de datos que se han vinculados

MEDICIONES DEL ERROR DE VINCULACIÓN (IDEAL)

		Estado de coincidencia	
		Coincidencia (un par de la misma entidad)	Discrepancia (un par procedente de entidades diferentes)
Estado de vínculo	Vínculo	Coincidencia identificada a	Coincidencia falsa b
	Sin vínculo	Coincidencia desaprovechada c	Discrepancia identificada d

Valor predictivo positivo (VPP o precisión): proporción de vínculos que son coincidencias verdaderas=a/(a+b)

Tasa de coincidencias falsas (1-PPV): proporción de vínculos que son coincidencias falsas.=b/(a+b)

¿MINIMIZAR LAS COINCIDENCIAS FALSAS O DESAPROVECHADAS?

1. Envío de invitaciones para pruebas:

   Es importante capturar todas las posibles coincidencias verdaderas

   → El objetivo es maximizar la **sensibilidad**

2. Administración de tratamientos contra las drogas:

   Es importante evitar coincidencias falsas y garantizar que todos los registros vinculados son coincidencias verdaderas

   → El objetivo es maximizar la **especificidad**

3. Estimación de la frecuencia o prevalencia mediante la vinculación de un conjunto de personas a un conjunto de acontecimientos:

   Los vínculos desaprovechados conducen a un sesgo negativo, los vínculos falsos conducen a un sesgo positivo.

   → El objetivo es minimizar **errores totales**

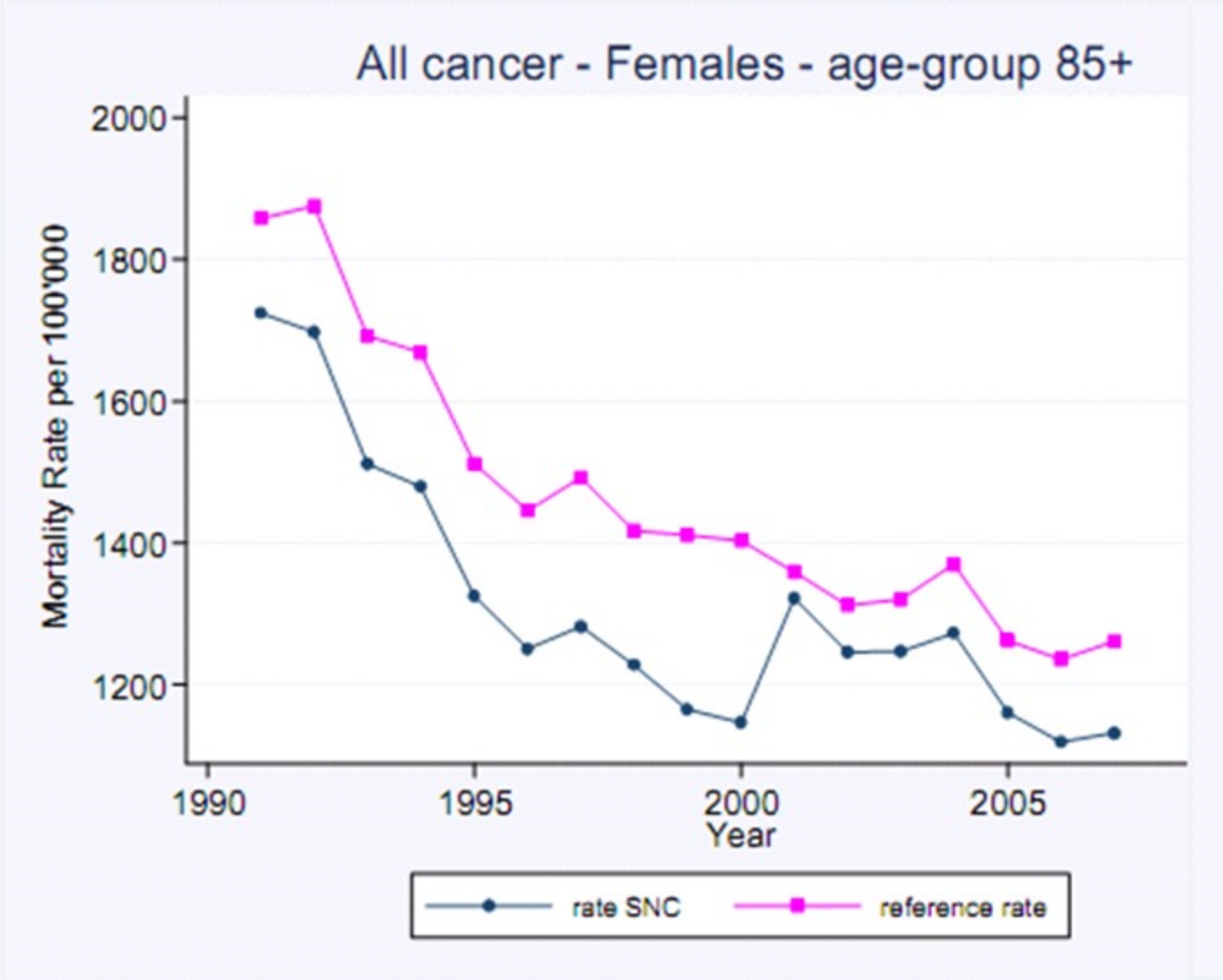
EVALUACIÓN DE LA CALIDAD DE LOS VÍNCULOS

1. Comparaciones de registros vinculados/desvinculados
2. Comparaciones con datos de referencia externos
3. Datos de referencia de alto nivel
4. Comprobaciones de plausibilidad

	Matched pairs	ISC residuals	MDC residuals
Maternal factors	*n* = 250 186	*n* = 2596	*n* = 3798
Mean age (years)	29.6	28.9	30.0
Married	78.7	73.4	NA
Australian-born mother	72.6	77.9	75.7
Birth in private hospital	22.0	27.1	28.9
Caesarean delivery	23.1	20.7	28.9
Diabetes	4.4	3.2	4.8
Hypertension	7.1	7.9	8.3
Stillbirth[a]	0.5	4.6	3.2
Baby factors	*n* = 253 538	*n* = 1570	*n* = 3157
Birthweight (g)			
<1000	0.4	0.8	4.4
1000–1999	1.7	3.9	7.9
2000–2999	18.5	22.5	27.8
3000–3999	66.9	59.9	48.8
4000–4999	12.4	12.1	10.5
≥5000	0.2	0.3	0.3
Plurality			
Singletons	96.7	95.4	95.5
Twins	3.2	4.6	4.2
Death in hospital	0.2	0.9	2.8
Preterm birth[b]	6.5	9.7	26.3
Transfer to another hospital	5.3	11.9	10.4

Ford JB, Roberts CL, Taylor LK (2006) Characteristics of unmatched maternal and baby records in linked birth records and hospital discharge data. Paediatr Perinat Ep 20 (4):329-337

2. COMPARACIONES CON DATOS DE REFERENCIA EXTERNOS



Schmidlin K et al (2013) Impact of unlinked deaths and coding changes on mortality trends in the Swiss National Cohort. BMC Med Inform Decis Mak 13 (1):1

3. COMPARACIONES CON DATOS DE REFERENCIA DE ALTO NIVEL

- La mayoría de las mediciones de error de vinculación pueden estimarse utilizando datos de **alto nivel** en los que se conoce el verdadero estado de coincidencia de los pares de comparación.

- Puede tratarse de
  - un subconjunto con identificadores únicos de alto nivel
  - una muestra con revisión administrativa

- Requiere que el subconjunto/muestra sea **representativo** del conjunto de datos

4. COMPROBACIONES DE PLAUSIBILIDAD

Utilizar pruebas de que dos registros no pertenecen a la misma persona para identificar coincidencias falsas, por ejemplo:

- Dos registros de nacimientos o defunciones
- Admisión tras defunción
- Vinculación de los registros de cáncer de próstata con los registros hospitalarios femeninos

	Infants (n = 733,770)		
	Not (n = 773,446)	Simultaneous Admission (N = 324)	p
Male	51.7%	56.8%	.07
Preterm[a]	7.9%	15.1%	<.001
White[a]	75.8%	66.8%	(ref)
Mixed[a]	4.6%	6.0%	.09
Asian[a]	11.1%	18.4%	<.001
Black[a]	5.3%	4.4%	.83
Chinese[a]	0.6%	1.0%	.26
Other[a]	2.7%	3.5%	.22
Multiple birth[a]	3.5%	3.8%	.75

Hagger-Johnson et al (2014) Identifying possible false matches in anonymized hospital administrative data without patient identifiers. Health Serv Res DOI:10.1111/1475-6773.12272

PUNTOS IMPORTANTES

- En la medida de lo posible, los métodos de vinculación deben adaptarse a la finalidad y los requisitos de los datos vinculados.
- Incluso pequeños errores de vinculación pueden sesgar considerablemente los resultados.
- Existen varias formas de evaluar el error de vinculación.
- La transparencia sobre los métodos de vinculación es muy importante.
- Existen directrices para facilitar la presentación de informes transparentes sobre los estudios de vinculación.

REFERENCIAS ÚTILES

- Doidge J, et al. (2020). Quality assessment in data linkage. Joined up data in government: the future of data linking methods Office for National Statistics. https://www.gov.uk/government/publications/joined-up-data-in-government-the-future-of-data-linking-methods
- Sayers, A., et al. (2015). "Probabilistic record linkage." Int J Epidemiol **45(3): 954-964.**
- Doidge, J. C., & Harron, K. (2018). Demystifying probabilistic linkage: Common myths and misconceptions. *International Journal of Population Data Science, 3*(1).
- Harron K, et al. (2017). A guide to evaluating linkage quality for the analysis of linked data. *Int J Epidemiol* **46**(5): 1699-1710.
- Benchimol, E. I. et al. The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) Statement (2015). *PLoS Med* 12, e1001885

DESAFÍOS Y OPORTUNIDADES PARA EL SISTEMA ESTADÍSTICO NACIONAL CHILENO

- Integrar micro-datos
- Sistematizar y validar procesos de integración
- Evaluar y mejorar continuamente
- Capacidades técnicas e infraestructura
- Aspectos legales y seguridad de la información

¿DÓNDE SE EVIDENCIAN ESTOS DESAFÍOS?

INE

- **Registro Estadístico de Población (REP)**
  - Caracterización de extranjeros sin RUN
  - Integración con Encuestas (Encuesta Nacional de Empleo, Encuesta de Presupuesto Familiares, etc.)
  - Deduplicación de registros
- **Registro Estadístico de Unidades Económicas (RUE)**
  - Integración de diferentes fuentes (RRAA, encuestas, registros de empresas internacionales) y deduplicación.
- **Censo(s) y registros históricos**
  - Deduplicación (personas, hogares, viviendas)
  - Vinculación entre Censos
- **Vinculación entre Registros Administrativos y Encuestas**

¿DÓNDE SE EVIDENCIAN ESTOS DESAFÍOS?

- **FONASA**
  - Deduplicación de registros para beneficiarios en sin RUN
- **Ministerio de Educación**
  - Elegibilidad de beneficios a partir de información administrativa desde otros ministerios
- **Ministerio de Salud**
  - Estimación de cobertura de vacunas para población extranjera
- **Servicio Nacional de Migraciones**
  - Flujos migratorios

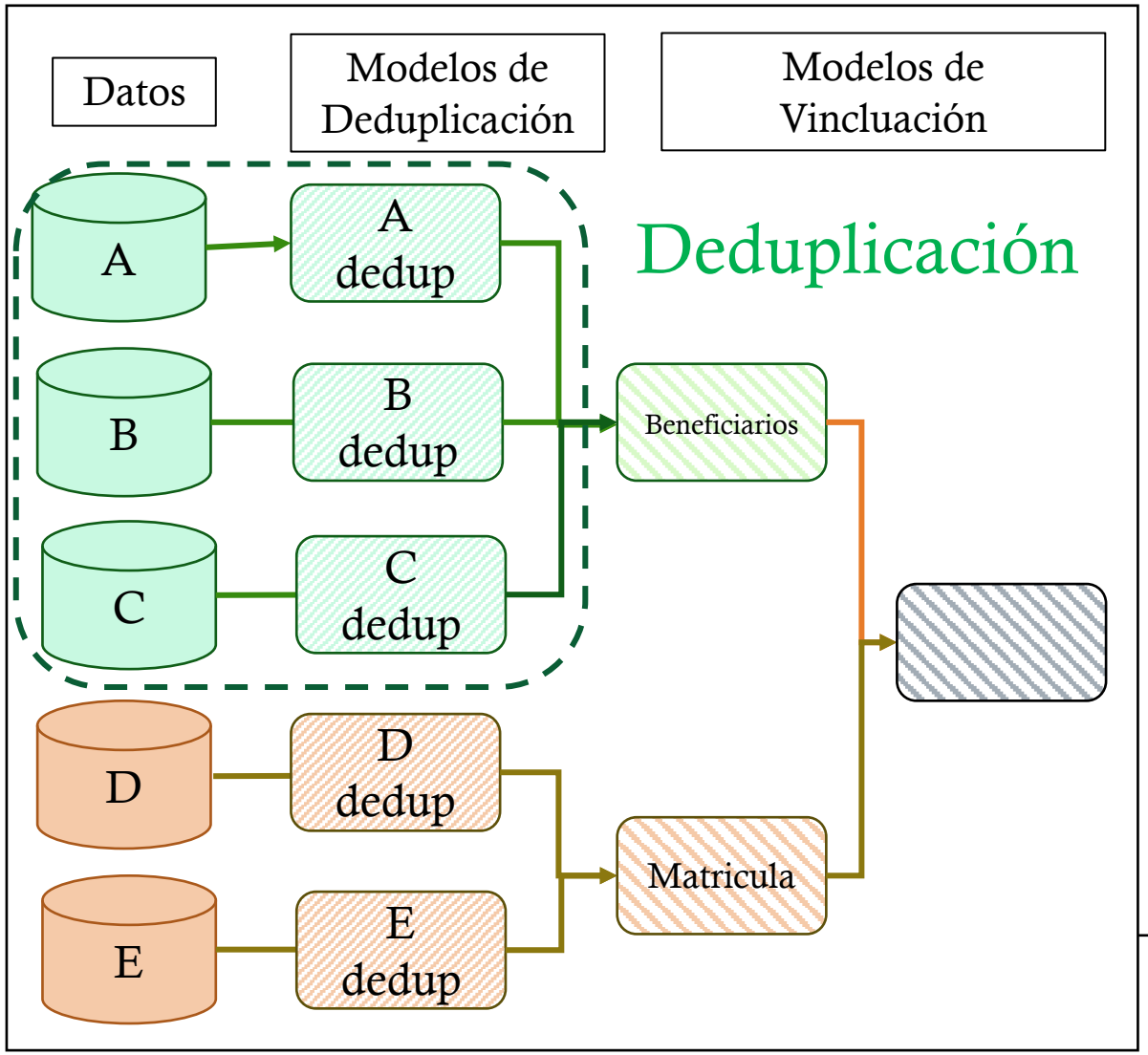
COLABORACIÓN ENTRE MINISTERIO DE EDUCACIÓN, MINISTERIO DE SALUD, FONASA Y ACADÉMICOS DE UCL

- Objetivo: Mejorar la integración de registros administrativos y encuestas mediante la implementación de métodos de vinculación probabilística.
  - Caso de estudio 1: Mineduc – Fonasa: Integrar registros para la población escolar sin RUN.
  - Caso de estudio 2: Mineduc – DEIS: Ajustar estimaciones de cobertura de vacunas en la población escolar.

CASO DE ESTUDIO 1: MINEDUC – FONASA

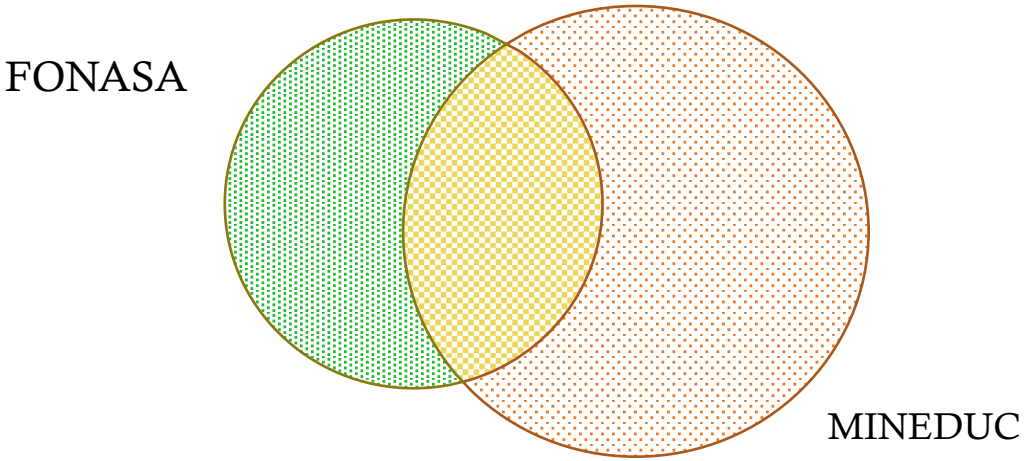
- Contexto:
  - Fonasa asigna Números de Identificación Provisorios (NIP) a personas migrantes sin RUT para que sean atendidas.
  - Mineduc un Identificador Provisorio Escolar (IPE)
- Objetivo:
  - Deduplicar e integrar mediante el modelo Fellegi-Sunter de vinculación probabilística aprovechando el desarrollo reciente de SPLINK

CASO DE ESTUDIO 1: MINEDUC – FONASA



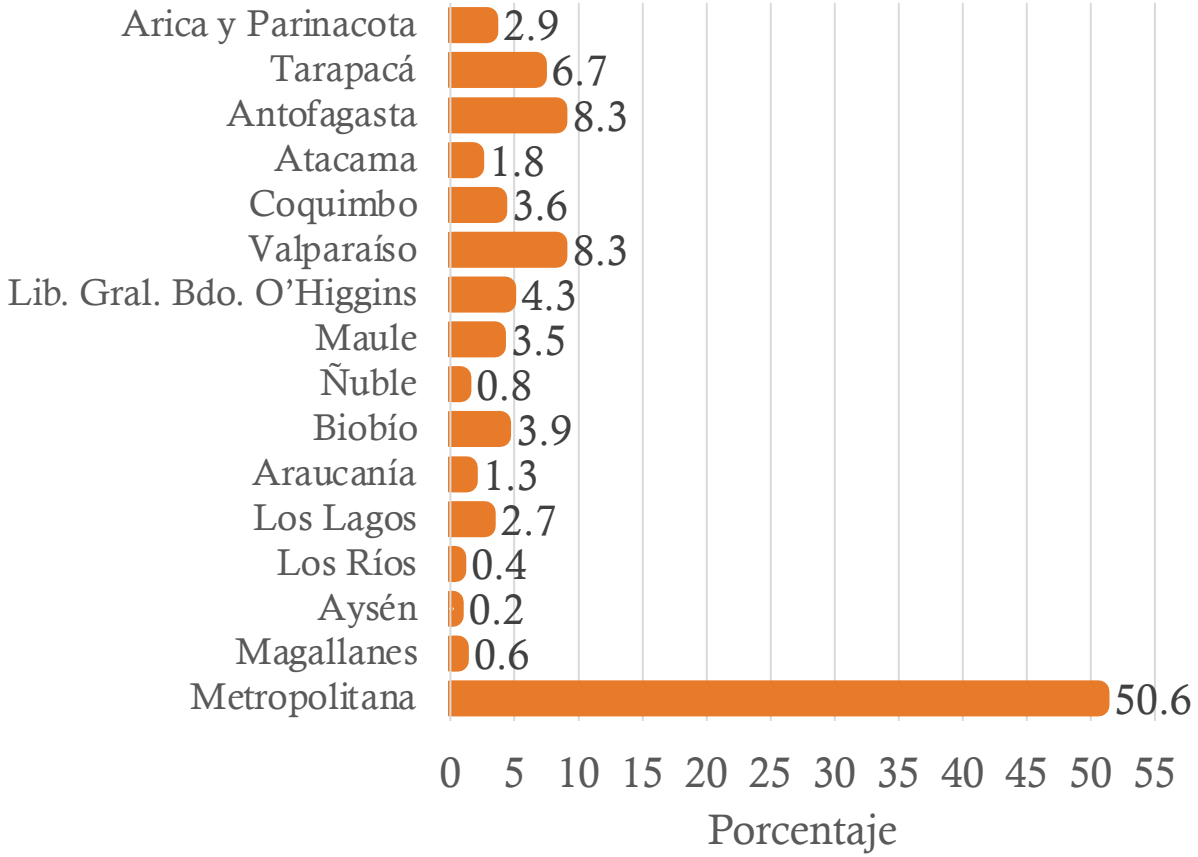
FONASA: ~15.6m registros; ~ 850.000 NIP, ~195.000 menores de 18 años.

Mineduc: ~3.6m alumnos por año, ~130.000 con IPE en 2022.



CASO DE ESTUDIO 1: MINEDUC – FONASA

Distribución regional de la población de estudiantes extranjeros con IPE en el sistema escolar chileno en el año 2022*



Fuente: Mineduc, *estimaciones preliminares

Múltiples razones por la que registros no vinculen:

- No todas las personas son beneficiarios FONASA
- Salidas y entradas del país
- Errores en el registro de información (ejemplo. Apellidos)
- Carácter longitudinal de los registros administrativos

SPLINK

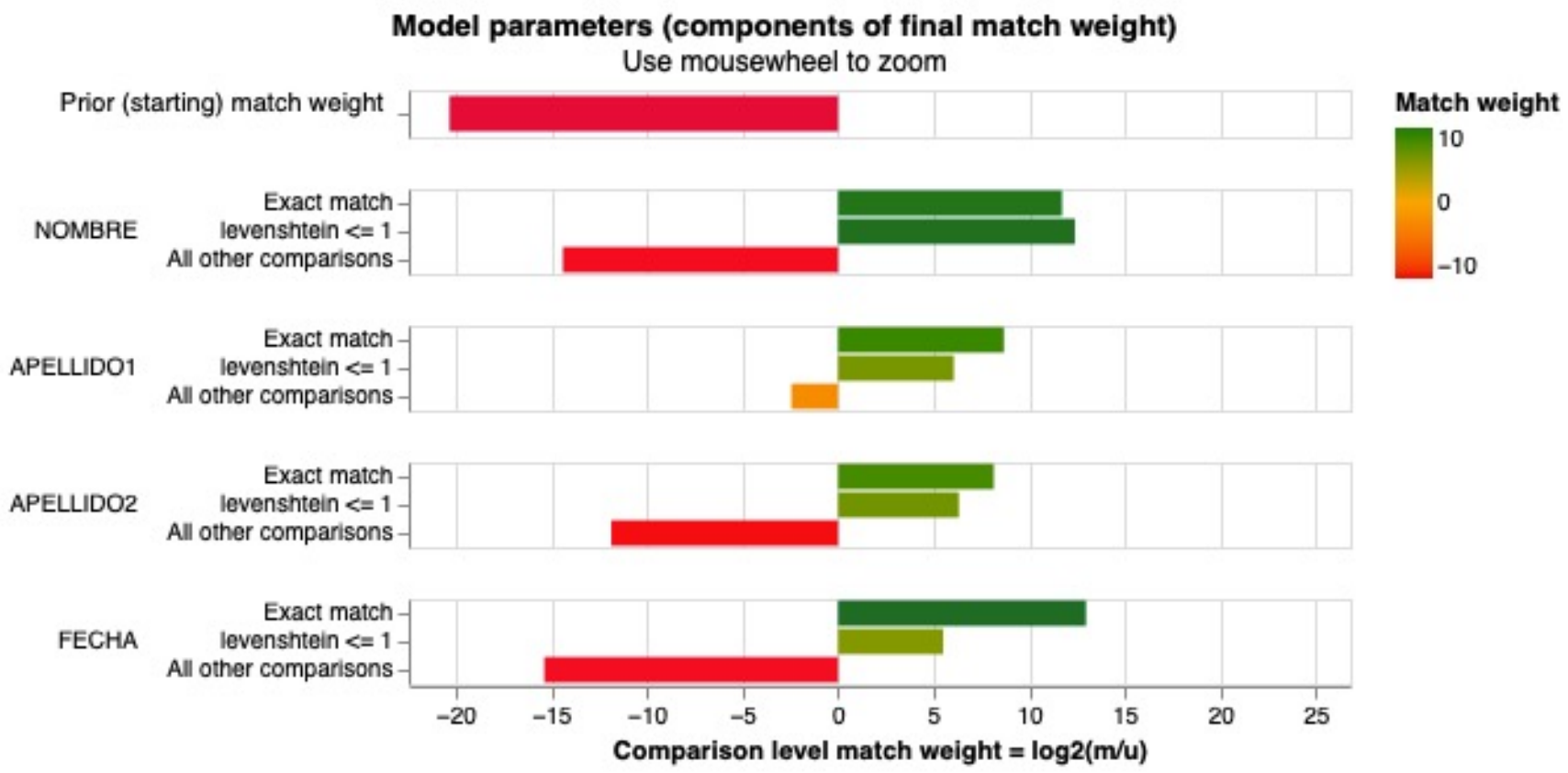
- Paquete de Python (PySpark) que impl
  vinculación de registros y permite estim
  Esperanza-Maximización (EM)
- Desarrollado por el Ministerio de Justi
  proyecto Data First del Administrative
- Trabaja a una escala mucho mayor y m
  actuales de código abierto (más de 100
  programa de R FastLink.

```
settings = {
    "link_type": "dedupe_only",

    "blocking_rules_to_generate_predictions": [
    "substr(l.NOMBRE, 1,1) = substr(r.NOMBRE, 1,1)
        and substr(l.APELLIDO1, 1,1) = substr(r.APELLIDO1, 1,1)
        and l.AGNOMES = r.AGNOMES",
    ],

    "comparisons": [
        cl.levenshtein_at_thresholds("NOMBRE",1),
        cl.levenshtein_at_thresholds("APELLIDO1",1),
        cl.levenshtein_at_thresholds("APELLIDO2",1),
        cl.levenshtein_at_thresholds("FECHA", 1),

    ],

    "retain_matching_columns": True,
    "retain_intermediate_calculation_columns": True,
    "max_iterations": 100,
    "em_convergence": 0.001,

}
linker = DuckDBLinker(df, settings)
```

PASOS DE IMPLEMENTACIÓN MODELO

- Dependiendo la bases de datos, definir y entender las variables a utilizar: Ejemplo: Nombres, Apellido Paterno, Apellido Materno, Fecha de Nacimiento.
- Definición del Modelo de **Deduplicación**
  - Definir las reglas de bloqueo (*blocking rules*)
    - Ejemplo: (Misma primera letra del nombre) Y (misma primera letra del apellido) Y (mismo mes y año de nacimiento).
  - Definir reglas determinísticas de comparación para variables
    - Coincidencia exacta, Jaro Winkler, Levenstein, Jaccard, geográficas, cercanía de fechas, etc.
- Entrenar el modelo (estimar probabilidades m y u)
- Estimar el modelo y definir umbrales de aceptación

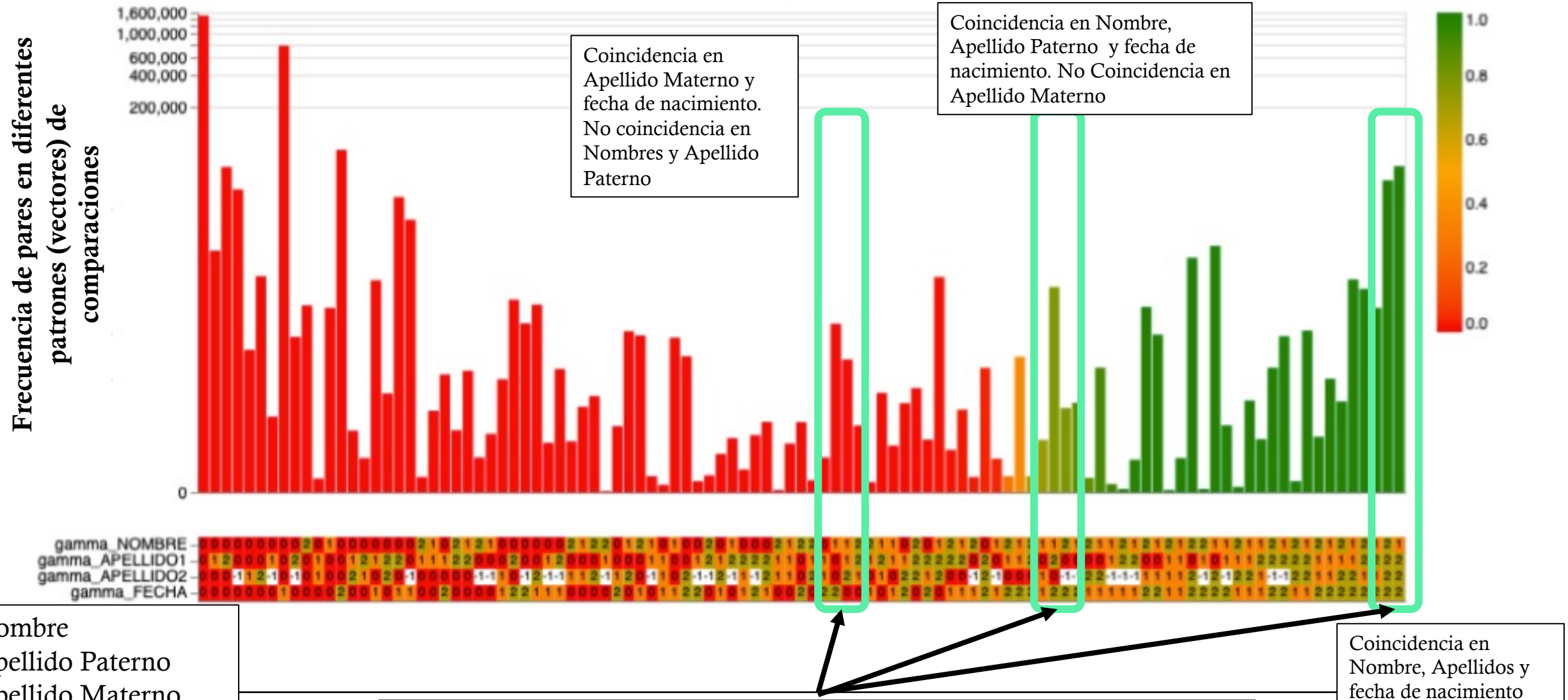
CASO DE ESTUDIO 1: MINEDUC – FONASA



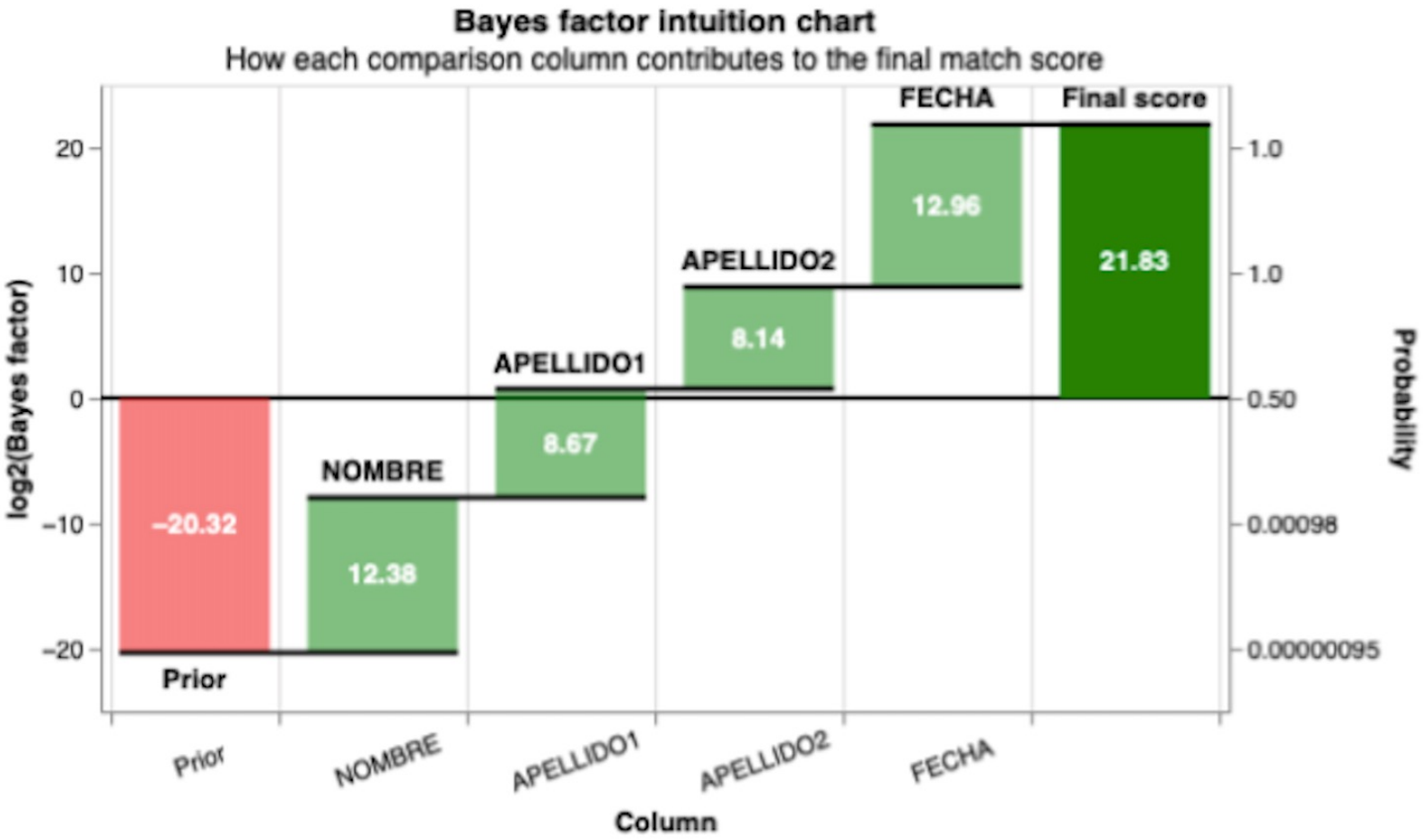
Parámetros del Modelo

- Luego de entrenar el modelo, los pesos de coincidencia se calculan a partir de los valores u y m estimados para cada variable a comparar.
- Los pesos de coincidencia (match-weight) resumen el modelo.

Conteo de valores en patrones (vectores) de coincidencia

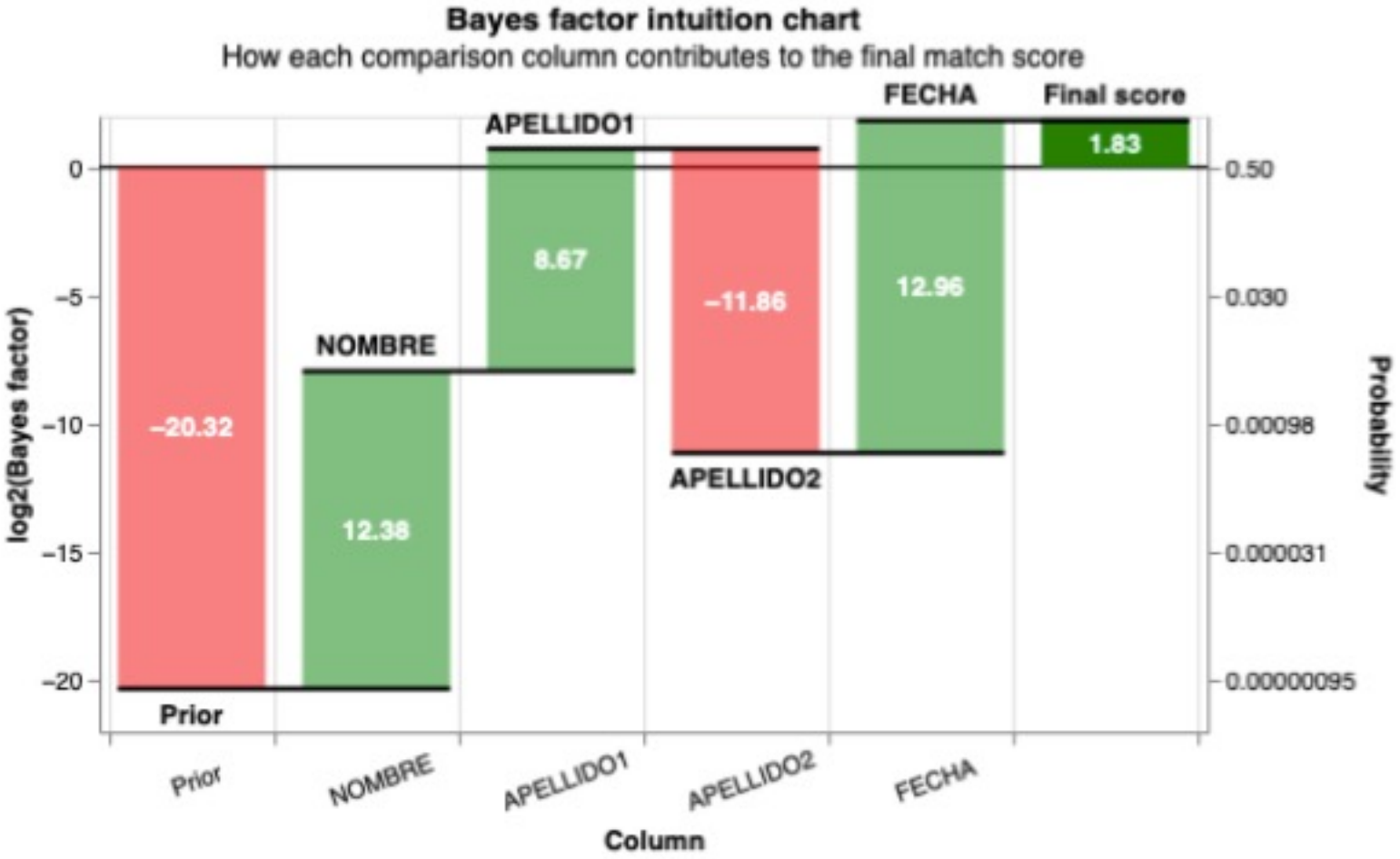


	NOMBRE	APELLIDO1	APELLIDO2	FECHA
☐	CONSUELO ESPERANZA	LUGO	DELGADO	2008-01-30
☐	COSUELO ESPERANZA	LUGO	DELGADO	2008-01-30



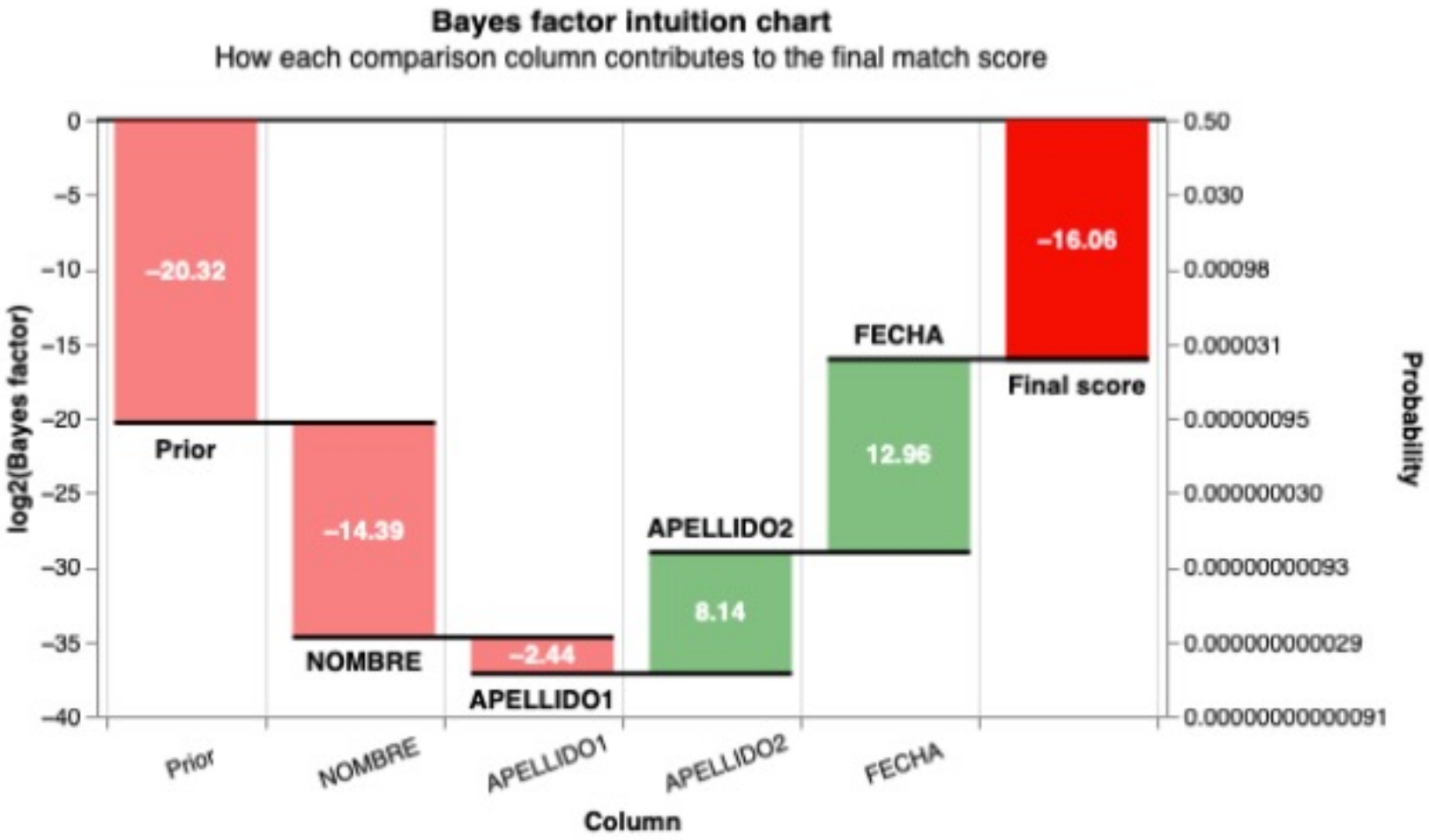
- Coincidencia en Nombre, apellidos materno, paterno y fecha de nacimiento

	NOMBRE	APELLIDO1	APELLIDO2	FECHA
☐	BENJAMIN ESTEBAN	VARGAS	ALFARO	1996-03-27
☐	BENJAMIN ETEBAN	VARGAS	AALARO	1996-03-27

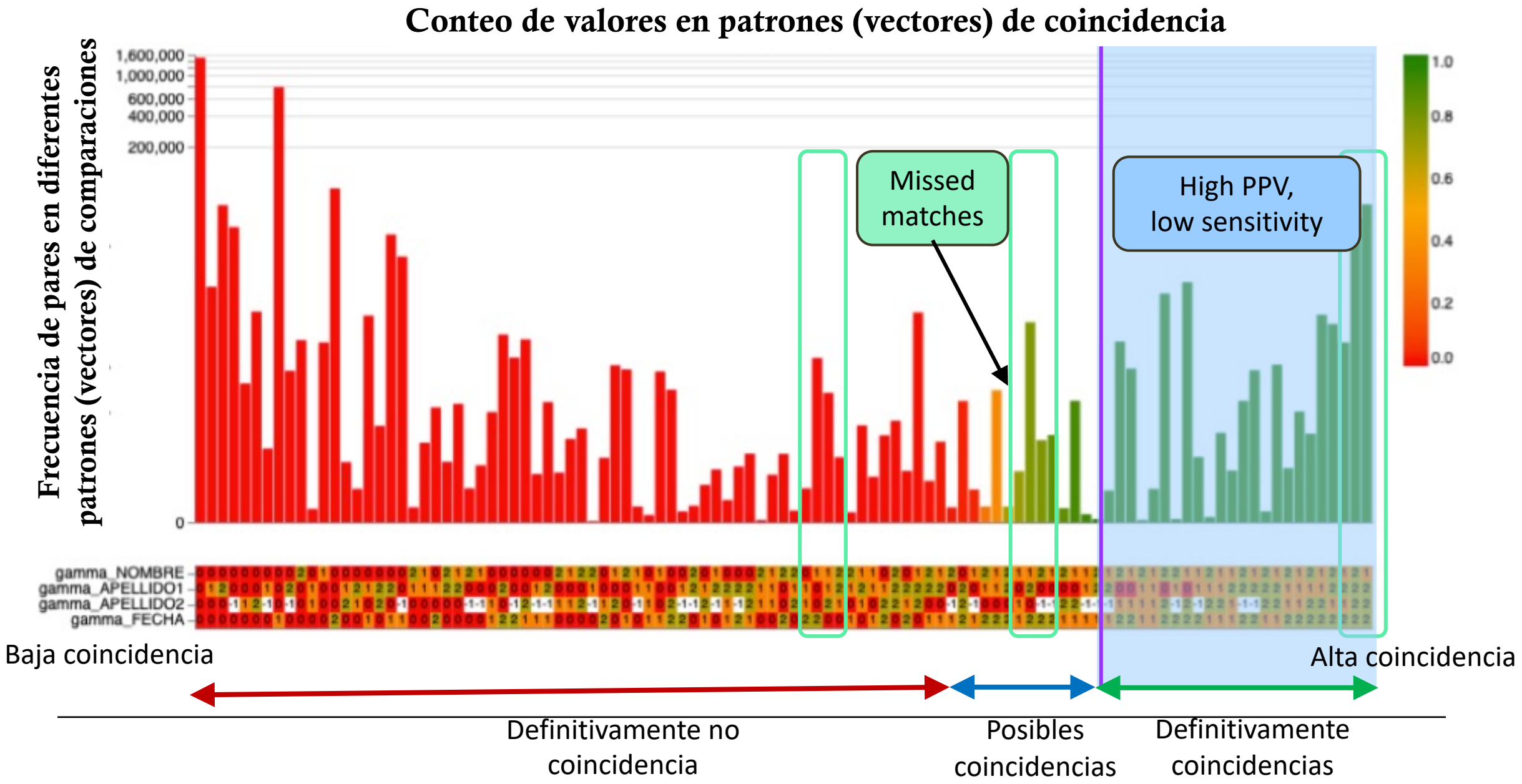


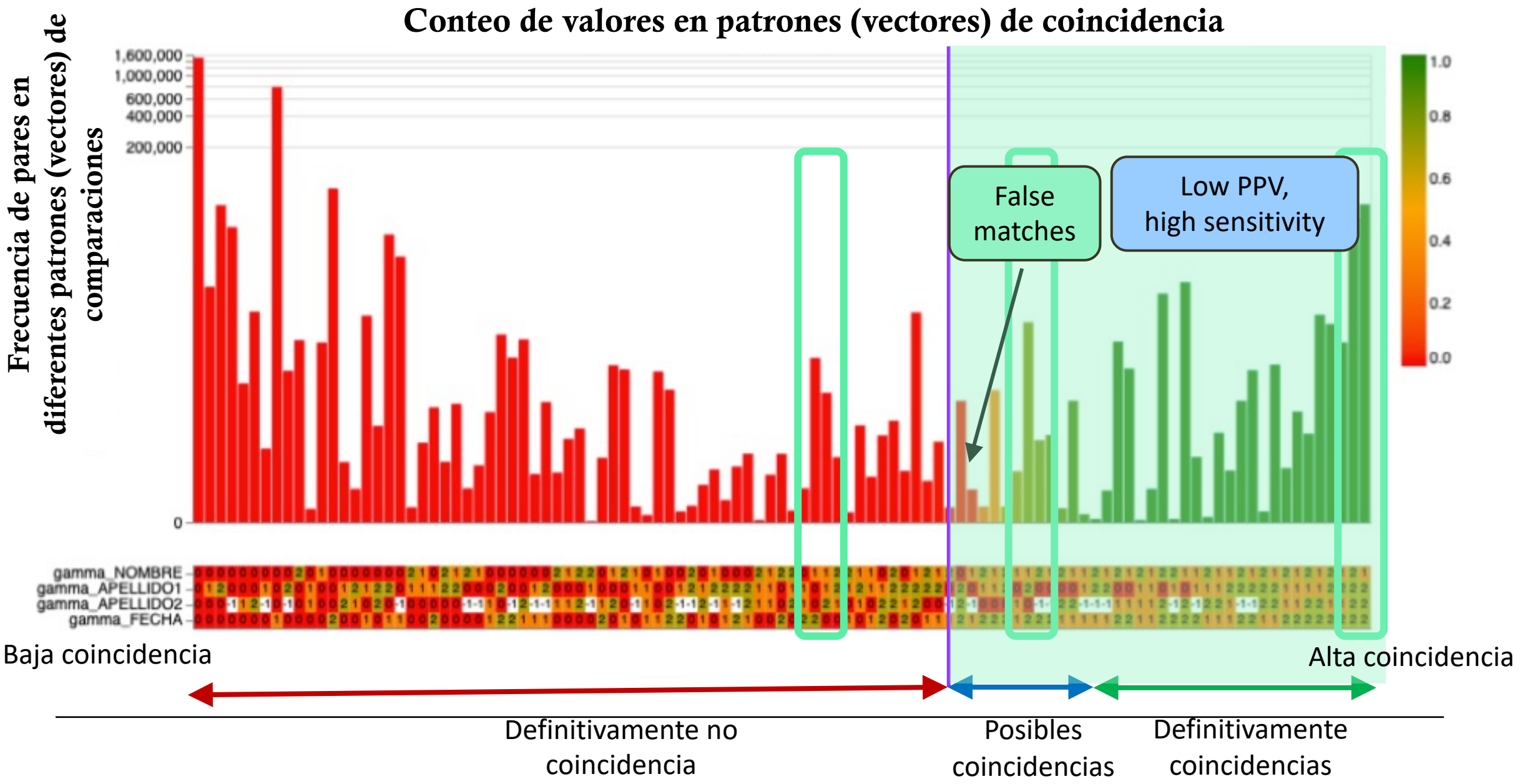
- Coincidencia en Nombre, Apellidos Paterno y fecha de nacimiento
- No Coincidencia en Apellido Materno

	NOMBRE	APELLIDO1	APELLIDO2	FECHA
☐	EDUARDO	ARELLANO	CONTRERAS	1995-11-18
☐	ELENA DEL CARMEN	ARAYA	CONTRERAS	1995-11-18



- Coincidencia en Apellido Materno y fecha de nacimiento.
- No coincidencia en Nombres y Apellido Paterno


Conteo de valores en patrones (vectores) de coincidencia
Frecuencia de pares en diferentes patrones (vectores) de comparaciones
1,600,000
1,000,000
600,000
400,000
200,000
0
1.0
0.8
0.6
0.4
0.2
0.0
Missed matches
High PPV, low sensitivity
gamma_NOMBRE
gamma_APELLIDO1
gamma_APELLIDO2
gamma_FECHA
Baja coincidencia
Alta coincidencia
Definitivamente no coincidencia
Posibles coincidencias
Definitivamente coincidencias


Conteo de valores en patrones (vectores) de coincidencia
Frecuencia de pares en diferentes patrones (vectores) de comparaciones
1,600,000
1,000,000
600,000
400,000
200,000
0
1.0
0.8
0.6
0.4
0.2
0.0
False matches
Low PPV, high sensitivity
gamma_NOMBRE
gamma_APELLIDO1
gamma_APELLIDO2
gamma_FECHA
Baja coincidencia
Alta coincidencia
Definitivamente no coincidencia
Posibles coincidencias
Definitivamente coincidencias

CONCLUSIONES

- Si bien en Chile existe un identificador único de personas (u otras unidades) que facilita la integración de registros administrativos, existen variados ejemplos en los que aún se requiere utilizar múltiples llaves que potencialmente no son únicas.
- El reciente desarrollo de SPLINK y la probada experiencia en otros países (Reino Unido) permite utilizar métodos de vinculación probabilística y determinística integrándolos dentro de los procesos que utilizan registros administrativos y encuestas.
- Al contar con un identificador único, el sistema Chileno está en una posición privilegiada para crear datos Gold-standard y entrenar modelos de vinculación probabilística.
- Se requiere aumentar las capacidades técnicas e infraestructura para utilizar estos métodos.

MUCHAS GRACIAS

Prof Katie Harron,
UCL Great Ormond Street Institute of Child Health
ECHILD
@klharron
k.harron@ucl.ac.uk

Dr. Nicolás Libuy,
UCL Social Research Institute
Centre for Longitudinal Studies
Institute of Education
@nicolibuy
nicolas.libuy.16@ucl.ac.uk