Desarrollo de una base de Datos del Mundo Real para oncología. Análisis descriptivo del cáncer de mama en Argentina.

Autores: Streich G*, Blanco Villalba M**, Cid C***, Bramuglia GF****

(*) Médico Oncólogo, Jefe del Servicio de Oncología del Htal. Militar Central de Buenos Aires
(**) Médico Oncólogo, Director del Centro Médico Austral de Buenos Aires, Vce. Pte. de la Sociedad Argentina de Cancerología
(***) Licenciado en Sistemas, Jefe de I.T. de Argenomics y Fundación Investigar
(****) Bioquímico y Farmacéutico, Director Científico de Fundación Investigar

Resumen

Introducción: Los registros basados en Datos del Mundo Real (Real World Data – RWD) son los que se obtienen por fuera de los ensayos clínicos sistematizados y aleatorizados. Estos permiten recoger información de una gran cantidad de pacientes y posibilita la participación de un número importante de profesionales. PrecisaXperta es una plataforma web desarrollada para este fin con más de dos años de funcionamiento, parametrizada para oncología. El diseño de la misma permite construir una base de datos epidemiológica en tiempo real y exportable para su procesamiento. **Objetivo:** Describir las características y el funcionamiento de esta herramienta de registro de datos en línea, explicar cómo fue desarrollada y analizar la calidad de la información registrada, tomando como ejemplo los datos obtenidos de cáncer de mama. **Materiales y métodos:** Médicos, informáticos y analistas en Ciencia de los Datos, participaron en el desarrollo. Datos del paciente, antecedentes, nivel educativo, diagnóstico, estadificación, marcadores moleculares, calidad de vida, tipos de tratamientos, progresión y respuesta, imágenes, complicaciones, eventos adversos son algunos de los campos que se incluyen. El tratamiento de los datos en cuanto a su encriptación, anonimización, protección y validación también es explicado. Los datos seleccionados de cáncer de mama para su descripción fueron procesados con programas estadísticos de nivel medio, ya que aún no se cuenta con el número requerido para aplicar motores de Big Data. **Resultados:** De un total de 6892 tumores sólidos, 1892 eran de cáncer de mama y se seleccionaron 1654 que cumplían con un ¨data set mínimo¨ elaborado *ad hoc*. Los casos procedentes de 13 provincias, mostraron un sesgo de geolocalización acorde con el lugar del ejercicio de los profesionales de la red colaborativa. La falta de datos predominante se detectó en los marcadores moleculares (ki67) y la correlatividad en algunas líneas de tratamiento. También se detectaron inconsistencias en fechas y esquemas terapéuticos. La curaduría de datos permitió excluirlos. La edad de las pacientes fue 55,3 ± 11,88 años. Al momento del diagnóstico el predominio estuvo en los estadios I: 36.48% y II 30.06%, con receptores hormonales positivos en 1424 (89,96%) casos. Los tratamientos predominantes fueron los hormonales (61,54%) y target dirigidos con un 30,85% para los Her2(+) y 39,14% para los Her2 (-) acompañados en la mayoría de los casos (85,9%) por algún período de quimioterapia. La inmunoterapia estuvo mucho menos representada (0,36%). Los datos fueron procesados, homogeneizados, agrupados y presentados y hechos accesibles en forma adecuada para su aplicación a los análisis de Datos del Mundo Real. **Conclusiones:** PrecisaXperta cumple con este propósito de sistematizar la información para facilitar su carga con su interfaz simple e intuitiva. Del análisis de los datos obtenido en cáncer de mama, se desprende que algunos campos deberán ser obligatorios para poder mejorar la calidad de la información. Los resultados que describen los cánceres de mama registrados, nos dan una visión de superficie de la población afectada y nos prepara para diseñar futuros estudios cuando contemos con Big Data local. Este tipo de desarrollo, con

mejoras continuas y resultados *online*, permitirán con su difusión, que los profesionales participantes cuenten con información de lo que sucede en el mundo real, teniendo disponible de manera democrática, la epidemiología para poder estudiar, publicar e investigar con esos datos.

Palabras clave: RWD; datos del mundo real; base de datos; oncología; cáncer de mama; big data.

Summary

Introduction: *Registries based on Real World Data (RWD) are those obtained outside of systematized and randomized clinical trials. They allow the collection of information from a large number of patients and enable the participation of a significant number of professionals. PrecisaXperta is a web platform developed for this purpose with more than two years of operation, parameterized for oncology. Its design allows the construction of an epidemiological database in real time and exportable for processing.* ***Objective****: To describe the characteristics and operation of this online data recording tool, explain how it was developed and analyze the quality of the information recorded, taking as an example the data obtained for breast cancer.* ***Materials and methods****: Physicians, computer scientists and data science analysts participated in the development. Patient data, history, educational level, diagnosis, staging, molecular markers, quality of life, types of treatments, progression and response, imaging, complications, adverse events are some of the fields included. Data treatment in terms of encryption, anonymization, protection and validation is also explained. The selected breast cancer data for description were processed with medium-level statistical programs, since the number required to apply Big Data engines is not yet available.* ***Results****: From a total of 6892 solid tumors, 1892 were breast cancer and 1654 were selected that complied with a ¨data set minimum¨ elaborated ad hoc. Cases from 13 provinces showed a geolocation bias according to the place of practice of the professionals in the collaborative network. The predominant lack of data was detected in molecular markers (ki67) and correlativity in some lines of treatment. Inconsistencies in dates and therapeutic schemes were also detected. Data curation made it possible to exclude them. The age of the patients was 55.3 ± 11.88 years. At the time of diagnosis the predominance was in stage I: 36.48% and II 30.06%, with positive hormone receptors in 1424 (89.96%) cases. The predominant treatments were hormonal (61.54%) and target directed with 30.85% for Her2(+) and 39.14% for Her2(-) accompanied in most cases (85.9%) by some period of chemotherapy. Immunotherapy was much less represented (0.36%). Data were processed, homogenized, pooled and presented and made accessible in a form suitable for application to Real World Data analyses.* ***Conclusions:*** *PrecisaXperta fulfills this purpose of systematizing the information to facilitate its loading with its simple and intuitive interface. From the analysis of the data obtained in breast cancer, it is clear that some fields should be mandatory in order to improve the quality of the information. The results describing the registered breast cancers, give us a surface view of the affected population and prepares us*

to design future studies when we have local Big Data. This type of development, with continuous improvements and online results, will allow with its dissemination, that the participating professionals have information of what happens in the real world, having available in a democratic way, the epidemiology to be able to study, publish and investigate with these data. Keywords: RWD; real-world data; database; oncology; breast cancer; big data.

Keywords: *RWD; real world data; database; oncology; breast cancer; big data.*

Introducción

Las herramientas de Inteligencia Artificial (IA) están en continua evolución y permiten ampliar la comprensión de fenómenos multivariables. Son empleadas en un sin número de áreas, donde la cantidad de datos es tan grande que es imposible analizarlos por métodos estadísticos tradicionales. En medicina y epidemiología han permitido la construcción de modelos predictivos en diferentes especialidades. La oncología y en especial el cáncer de mama, por su alta incidencia, son áreas donde interesantes aplicar estos métodos en bases de datos sistematizadas para tal fin.

Los Datos del Mundo Real (RWD) y su correlato en la búsqueda de evidencia médica (Evidencia del Mundo Real – RWE) [1, 2] son un nuevo enfoque que permite involucrar en una investigación a un número importante de profesionales que aceptan participar, mediante la recolección de información epidemiológica anonimizada de sus pacientes y así para generar una base de datos sistematizada que organiza, jerarquiza y facilita su análisis. Además la sistematización del ingreso de la información permite definir su confiabilidad, así como, evaluar los sesgos y la factibilidad de obtención de evidencia relevante [3,4].

PrecisaXperta es una es una herramienta de adquisición de datos del mundo real, desarrollada por Fundación Investigar sobre una plataforma web, con datos sistematizados para patologías oncológicas. Con más de 2 años de uso, se decide evaluar el contenido de la información recabada, haciendo foco en una patología. Además, teniendo en cuenta el funcionamiento de la interfaz de carga, se analizan cambios que redunden en una mejora la calidad de la información.

Objetivo:

El objetivo del presente trabajo es describir cómo fue desarrollada la plataforma de registro de datos PrecisaXperta y cómo es su funcionamiento. También analizar la calidad de los datos consignados y proponer mejoras en los métodos de registro, teniendo en cuenta la información obtenida sobre los datos de cáncer de mama en Argentina recolectados durante el período que va desde agosto de 2020 a noviembre de 2021.

Materiales y métodos

Con el objeto de conseguir una interfaz amigable para favorecer la adherencia de profesionales a compartir información epidemiológica sobre Datos del Mundo Real, analistas en ciencia de los datos, oncólogos y programadores fueron convocados en una primera etapa para comenzar con el desarrollo de una plataforma web que fuera de fácil acceso y sistematizada para oncología. La estructura básica debía contar con datos del paciente, antecedentes, descripción del tumor, estadificación, etapas y líneas de tratamiento, respuesta y progresión, complicaciones, imágenes, otros medicamentos, calidad de vida y eventos

adversos. En el trabajo multidisciplinario, el consenso fue incluir opciones múltiples o simples, precargadas para que el uso del teclado se minimizara. Definidos así los campos, los datos y los métodos de carga, la plataforma fue puesta en operación, invitando a participar a especialistas con la consigna de registrar de manera libre datos anónimos de sus pacientes. Se convino con los profesionales que la epidemiología registrada por el grupo quedaría disponible democráticamente para ser utilizada por aquellos participantes que lo desearan.

Participaron en el aporte de esta información 38 médicos especialistas distribuidos en 13 provincias de Argentina. Del total de datos registrados en PrecisaXperta (correspondientes a 6892 pacientes al momento del cierre de esta publicación), 1892 tenían el diagnóstico de cáncer de mama, de los cuales se seleccionaron 1654 que fueron utilizados para el análisis ya que cumplían con el criterio de ¨data set mínimo¨ definido por los autores. Los descriptores considerados fueron: edad, peso, antecedentes, co-morbilidades, geolocalización (referida a la provincia donde habita la paciente), el estadio (TNM) al momento del diagnóstico, receptores hormonales, marcadores moleculares y tipo de tratamiento recibido.

Para el análisis de la calidad de la información, se decidió por un tumor sólido de alta prevalencia y como criterios de inclusión de selección de casos para evaluar, se tomaron aquellos que tuvieran como diagnóstico cáncer de mama y que cumplieran con el data set mínimo registrado, reflejado en las tablas adjuntas. Fueron excluidos los casos que no cumplieran con estas condiciones.

El análisis de los casos fue procesado mediante Infostat**,** software para análisis estadístico de aplicación general. Este permitirá en un futuro, al aumentar el volumen de datos, emplear estadísticas avanzadas como métodos de modelación y análisis multivariado.

Resultados

Se describen los datos registrados de la afección seleccionada (Cáncer de mama) para evaluar la calidad de la información obtenida. En la Tabla 1 se observan los datos antropométricos de las pacientes incluidas. La edad promedio de algo más de 55 años, nos coloca delante de un escenario de pacientes postmenopáusicas en su mayoría. Talla y peso corresponden en promedio a un índice de masa corporal medio (IMCm) de 25.2, valor calculado automáticamente por el sistema. El 11% de las pacientes del grupo estudiado registra la obesidad como antecedente, dato valioso por su correlato con algunos tumores malignos [(5)]. Cuando se evalúan las pacientes, en la misma se observan como antecedentes más frecuentes, la diabetes, la obesidad y el tabaquismo.

El origen de los casos según la procedencia geográfica (Figura 5), refleja una mayor contribución de acuerdo al tamaño de la población de cada provincia, con la excepción de la de Santa Fe que está sobre representada, mientras que con Buenos Aires sucede lo contrario. Sobre el total, el 80% de los casos pertenecen a 3 regiones (Santa Fe, Buenos Aires y Capital Federal) que evidencia un sesgo en la representatividad, que se origina en la distribución geográfica de los investigadores participantes.

En la Tabla 2, observamos el número de pacientes según estadios al diagnóstico del cáncer de mama. En PrecisaXperta, encontramos que casi el 100% de los tumores habían sido estadificados junto con los marcadores moleculares registrados de estas pacientes. Si bien no se ven representados el 100% de los casos incluidos, muy pocas pacientes quedaron sin

clasificar para hormonales y factor de crecimiento epidérmico (Her2). No así sucedió con el marcador de proliferación celular ki67, con un número de registros sensiblemente menor, lo que imposibilitó efectuar la clasificación molecular junto con la estadificación TNM.

Se analizan las diferentes formas de presentación que llegaron al diagnóstico de cáncer de mama en nuestro grupo de observación. Podemos apreciar en la Tabla 1 que en el 31% de los casos se presentó alguna manifestación clínica que llevó a la paciente a la consulta y en más del 50% (representado por la sumatoria de la detección de un nódulo, el hallazgo incidental o una imagen sospechosa), refleja que las estrategias de *screening* clínico y mamográfico, permiten un diagnostico precoz de la enfermedad [6,7] En la Tabla 3, se presenta la frecuencia con la que en algún momento de su evolución, las pacientes recibieron un tipo de tratamiento según grupo terapéutico (Hormonoterapia, Quimioterapia, Target Dirigidos o Inmunoterapia). En cada uno se indica las drogas más frecuentemente empleadas. Se observa que en esta patología, la quimioterapia y la hormonoterapia encabezan la frecuencia de uso, seguidas de los tratamientos blanco dirigidos y en muy pocos caso vemos registros de inmunoterapia. Los datos de evaluación de calidad de vida encontrados reflejan la buena tolerancia y performance de los tratamientos inclusive en estadios avanzados de la enfermedad. (Tabla 4)

Discusión

Contar con información de la evolución en la población general de los tratamientos por fuera de los estudios sistematizados se ha tornado crucial para la adecuada toma de decisiones en la investigación, el desarrollo de nuevas moléculas y la implementación de políticas de Salud. Las dificultades para contar con esos datos son bien conocidas. Bases de datos anárquicas, no sistematizadas, incompletas y engorrosas para su carga son alguna de ellas, sin dejar de mencionar la poca adherencia de los profesionales a colaborar con los registros, ya que esta práctica aumenta la carga de trabajo en sus actividades rutinarias. El sistema empleado en esta publicación contribuye a minimizar esta problemática.

El incremento exponencial en la generación de datos en salud, ha traído aparejado el desarrollo de herramientas para el manejo de la Big Data médica. Millones de datos son generados alrededor del mundo y de su sistematización y ordenamiento dependerá la utilidad que podamos darle a los mismos. La interconectividad asociada al acceso a internet y su almacenaje virtual, han sido cruciales para este crecimiento. Se considera que para el 2025 la medicina será el área donde más se desarrollará la Big Data en el mundo [8-10].

La plataforma web PrecisaXperta, es una base de datos en línea, con un entorno gráfico sumamente amigable, pensada para que el médico incorpore de manera simple y rápida información del mundo real de pacientes oncológicos. Facilita el registro de la información, interconectando todos los datos epidemiológicos para construir estadísticas en tiempo real.

Por otra parte, los datos generados por la medicina de precisión y la bioinformática, con la información procesada para estudios genómicos en general y los aplicados en oncología clínica en particular, son una muestra más de la importancia que significa contar con datos clínicos confiables del mundo real y construir modelos predictivos con inteligencia artificial.[11]

Tratamiento del dato en PrecisaXperta

La información que registra esta base de datos parte de un algoritmo de anonimización. El mismo de manera automática convierte datos filiatorios y de identificación personal en un código alfa-numérico de 8 dígitos irreversible. Sólo el profesional actuante a cargo del paciente

con su usuario y contraseña, único y personal, puede hacer el recorrido inverso y transformar la codificación en los datos del paciente. Queda claro que la información recibida y procesada son datos epidemiológicos y no incluye datos sensibles de los pacientes.

Para garantizar la seguridad y protección de los datos, todo el contenido y procesamiento es encriptado mediante el método de AES (Advanced Encryption Standard, más conocida como Rijndael), empleado por el gobierno de USA para proteger los datos generales. A su vez, por normativa regulatoria de la República Argentina, todas las bases de datos generadas por Fundación Investigar son registradas en el Registro Nacional de Protección de Datos Personales (RNPDP) que depende del Ministerio de Justicia de la Nación y que es el organismo que otorga la certificación correspondiente (ley Nro. 25.326 y su reglamentación aprobada por decreto Nro. 1558/01). Para reforzar la anonimización, la geolocalización mínima a discriminar es la provincia (no la ciudad o dirección o la institución), para evitar el entrecruzamiento del registro con información que pudiera vincularse con la identificación de pacientes.

En cuanto al manejo ético de la información, se ha considerado como prioritaria la privacidad, la protección de datos sensibles y la garantía de la anonimización irreversible [12, 13]. El beneficio de contar con una base de datos democrática y transparente para los investigadores participantes, la seguridad en la protección y la responsabilidad de la veracidad de los datos son consideraciones inherentes al cumplimiento de esta premisa [12, 14].

Por procesos informáticos preestablecidos de programación, la carga se encuentra validada, en tal modo que cualquier modificación, reedición o corrección queda registrada con usuario, día y hora, permitiendo así su seguimiento.

El ingreso de los datos se realiza con una interfaz amigable que no requiere curva de aprendizaje ya que la secuencia de carga posee la lógica de la descripción médica de un caso. En la Figura 1, se puede observar una de las pantallas de carga, en este caso de un evento adverso, basada en la CTCAE (Internacional Common Terminology Criteria for Adverse Events, Version 5.0), donde su descripción no requiere del uso del teclado, permitiendo avanzar en la secuencia rápidamente.

Algoritmo y sistematización de la carga de datos

Para una mejor comprensión sobre contenido de PrecisaXperta, la Figura 2 muestra en una síntesis, el mapa de sitio de la plataforma. Allí se pueden observar los diferentes campos que la componen y las diferentes alternativas en complejidad con que se puede completar la información.

Uno de los objetivos indirectos de este artículo es identificar dentro de estas variantes de carga, cuáles deberían ser los campos obligatorios que permitan reflejar la epidemiología del área de estudio. Así se encontraron con que la secuencia de ingreso del paciente (alta para la base de datos), los antecedentes de enfermedades pre-existentes, los marcadores moleculares para su clasificación y la estadificación deberían cumplir con esta condición.

PrecisaXperta posee un sistema de estadificación inteligente que permite, con sólo indicar la condición T (tumor) N (ganglios) y M (metástasis), calcular el estadio de manera automática, contando con ayudas de las guías de la NCCN en el caso que así se requiera. Este complemento, denominado TNMsmart, facilita esta tarea y hace que, en comparación con otras bases de datos disponibles, difieran el porcentaje de pacientes estadificados (CIE-O – 50

del código internacional) ya que en la mayoría de los registros epidemiológicos, por la dificultad de su realización, entre un 40 a 50% de los tumores se encuentran sin este dato [5].

La plataforma de adquisición de datos sufre periódicamente, procesos de monitoreo de manera remota. Esta modalidad, aceptada por la FDA a partir de la incorporación en casi todos los estudios de investigación de los CRFe (formularios de registro de casos electrónicos), consiste en solicitarle al investigador que envíe documentación de manera digital que respalde la fuente del dato. El monitoreo de los datos ingresados y la verificación de su validez es una característica necesaria para garantizar la veracidad y transparencia de la información [15, 16].

En la Figura 3 se observa una de las pantallas de estadísticas que proporciona el sistema a modo de ejemplo. Esta se construye de manera automática con los datos planos anonimizados preparados para la exportación a herramientas de análisis estadísticos y motores de inteligencia artificial. Las actualizaciones de los datos son en tiempo real y todos los investigadores tienen acceso a esa información. Uno de los atractivos más importantes para que los profesionales participen en este proyecto es la visualización, parametrizada con los filtros personalizados, de la información clínico-epidemiológica generada y la posibilidad de utilizarla para estudios y publicaciones personales.

También, PrecisaXperta genera documentación accesoria de forma automática que ayuda al profesional en su actividad asistencial. Resumen de lo registrado, formularios de auditorías y un consentimiento informado de conformidad para compartir datos anónimos personales son algunos de los elementos que pueden llevarse al papel para poder contar con ese documento impreso.

La evolución de los pacientes se registra en la llamada ¨Línea de tiempo¨. La misma consiste en un par de ejes cartesianos donde el horizontal representa el tiempo transcurrido y el vertical las opciones de carga del seguimiento de los pacientes. Sincronizadas con las fechas correspondientes, aparecen en la pantalla las etiquetas inteligentes (Smart-labels), íconos que indican lo consignado y que contienen detalles del dato a guardar (ver figura 4).

Calidad de los datos encontrados de cáncer de mama

Si bien esta plataforma de registro de Datos del Mundo Real está configurada para cualquier tumor sólido, del total de datos encontrados correspondientes a 6892 pacientes, se eligieron para su análisis sólo aquellos que tenían como diagnóstico cáncer de mama. El motivo fue hacer foco una afección (la más prevalente de nuestra base) y hacer un análisis descriptivo de la información obtenida, su calidad, su validación e inconsistencias, y poder de esta manera proponer mejoras que serán fundamentales en el futuro para el manejo de la Big Data.

En cuanto a los datos antropométricos consignados, los mismos coinciden con la población media femenina para ese rango etario. [17] Aunque al grupo mayoritario (23,13%) se lo caracteriza como que no presenta Antecedentes, la aparición de un número importante de pacientes obesas (10,88%) junto con el dato del IMCm de pre-obesidad (25.2Kg/m^2), podría dar lugar a futuros análisis pormenorizados entre el exceso de peso y el cáncer de mama en nuestro medio.

La estadificación se encuentra completa en la mayoría de las pacientes, gracias al sistema inteligente (TNM smart) que funciona de manera automática para todos los tumores sólidos. Cuando se comparan los estadios al diagnóstico entre los datos de RITA-INC y los del presente análisis descriptivo, se encuentran algunas diferencias. El 36.48% de esta base se encuentra en

estadio I vs 9.9% en RITA-INC, 30.06% vs 22,9% son los de estadio II, 19.61% vs 16,2% corresponden al III y 7,56% vs 7,1% para el estadio IV respectivamente. Las diferencias podrían ser atribuibles a la disparidad en los totales estadificados o también reflejar que el sistema público (fuente de RITA-INC preponderante) trata una mayor proporción de casos avanzados comparados con los de PrecisaXperta que provienen de pacientes con seguro médico[5-7].

La estratificación Luminal A, Luminal B, Basal Like y No Like y Triple Negativo ya forma parte de la rutina en el manejo del cáncer mamario En este análisis descriptivo se encontró con que la mayoría de los casos contaban con marcadores hormonales y de Her2 y un porcentaje muy bajo contenía el marcador Ki67, lo que significó no contar con la información completa para la clasificación molecular. Se detectó entonces la necesidad de considerar datos obligatorios al set de marcadores moleculares y así poder registrar la clasificación molecular completa [18-23].

En la distribución de tipos de tratamiento, en el grupo analizado se observa que más del 60% de las pacientes recibieron quimio y hormonoterapia, basadas fundamentalmente en ciclofosfamida, adriamicina y paclitaxel para los primeros y trastuzumab, palbociclib y pertuzumab en los otros. Las pacientes cuyos marcadores moleculares las hacían candidatas a tratamientos targets, los recibieron, evidenciando que en el país las condiciones de acceso a los medicamentos de alto costo son buenos para esta patología. Cuando se cruzaron datos sobre la cobertura de salud en relación al acceso a estas drogas, no se verificó diferencia significativa entre la seguridad social, los seguros de salud y las pacientes hospitalarias sin cobertura, con baja representatividad en la presente muestra [24].

Otros datos de interés encontrados con respecto a esta afección, ponen de manifiesto la relevancia de los registros de la vida real, ya que generalmente no son considerados como objetivos en los estudios controlados [25]. La evaluación de la calidad de vida demuestra la buena tolerancia a los medicamentos, inclusive en los estadios más avanzados. Salvo el síntoma dolor que se hace presente en casi todas las etapas del tratamiento pero de muy baja intensidad, la actividad cotidiana se mantiene sin requerimiento de asistencia salvo en los estadios IV. También se ha verificado una relación entre el mayor nivel educativo y las estadificaciones más bajas (I y II) al diagnóstico de la enfermedad, independiente de la cobertura de salud que posea la paciente. Se infiere entonces que la información recibida a través de campañas de prevención, impactan de manera más efectiva en los niveles culturales más altos y son determinantes en el diagnóstico precoz, ya que el acceso a la atención médica se encuentra asegurado en el país [24]. Los Datos de la Vida Real, hasta conseguir el volumen de información necesaria para implementar motores de algoritmos predictivos con el procesamiento de Big Data, permiten una mirada poblacional descriptiva de gran utilidad al momento de avaluar aspectos relacionados con las tendencias y los tratamientos en oncología.

Conclusiones

Esta constituye la primera recolección de datos de la vida real en pacientes que padecen cáncer, mediante una plataforma web sistematizada para la carga de datos en oncología clínica. La misma fue concebida de manera multidisciplinaria, con una interfaz gráfica amigable e intuitiva. La información registrada con respecto a la edad, las co-morbilidades, la estadificación, los marcadores moleculares, la calidad de vida y los tipos de tratamientos aplicados, ofrecen una visión limitada pero prometedora en el uso de PrecisaXperta como plataforma integradora de recolección de datos validados para la especialidad. Del análisis de la patología elegida para evaluar la calidad de la información (cáncer de mama), surge que algunos campos deberían ser considerados obligatorios. Se pone en evidencia que los

marcadores moleculares (en especial el ki67) así como la secuencia fechas-tratamiento-control-respuesta deberían también considerarse con criterios de obligatoriedad y estos datos consignarse por lo menos cada 90 días para consolidar la evolución de la enfermedad [26-28]. Fundación Investigar pretende generar Big Data epidemiológica y desarrollar algoritmos predictivos basados en inteligencia artificial para ponerlos a disposición de la comunidad médica y colaborar con las terapéuticas de los pacientes oncológicos.

Agradecimientos: *Fundación Investigar agradece a Instituciones y empresas que colaboran con la promoción y difusión de la base de Datos de la Vida Real en oncología PrecisaXperta y a todos los investigadores que participan en este proyecto.*

Conflictos de Intereses

Dr. Streich, Guillermo declara no poseer ningún conflicto de intereses relacionado con la publicación
Dr. Blanco Villalba Marcelo declara no poseer ningún conflicto de intereses relacionado con la publicación
Lic. Cid, Christian declara haber colaborado con el desarrollo de la plataforma en Fundación Investigar
Dr. Bramuglia, Guillermo declara ser el Director Científico de Fundación Investigar

Fundación Investigar en una ONG, sin fines de lucro, que promueve la investigación en áreas de la Salud Humana.
No se han recibido aportes ni financiación alguna para la realización de este trabajo.

Referencias

1. Fernandes L E, Epstein C G, Bobe A M, Bell J, Stumpe M C, Salazar M E, Salahudeen A A, Pe Benito, R A, McCarter C, Leibowitz B D, Kase M, Igartua C, Huether R, Hafez A, Beaubier N, Axelson M D, Pegram M D, Sammons S L, O'Shaughnessy J A, Palmer, G. A. (2021) Real-world Evidence of Diagnostic Testing and Treatment Patterns in US Patients with Breast Cancer with Implications for Treatment Biomarkers From RNA Sequencing Data. Clinical breast cancer, 21(4), e340–e361. https://doi.org/10.1016/j.clbc.2020.11.012
2. Grimberg F, Asprion,PM, Schneider B, Miho E, Babrak L, Habbabeh, A (2021) The Real-World Data Challenges Radar: A Review on the Challenges and Risks regarding the Use of Real-World Data. Digital biomarkers, 5(2), 148–157. https://doi.org/10.1159/000516178
3. Makady A, de Boer A, Hillege H, Klungel O, Goettsch W, (on behalf of GetReal Work Package 1) (2017). What Is Real-World Data? A Review of Definitions Based on Literature and Stakeholder Interviews. Value in health: the journal of the International Society for Pharmacoeconomics and Outcomes Research, 20(7), 858–865. https://doi.org/10.1016/j.jval.2017.03.008
4. Sajjadnia Z, Khayami R, Moosavi M R (2020) Preprocessing Breast Cancer Data to Improve the Data Quality, Diagnosis Procedure, and Medical Care Services. Cancer informatics, 19, 1176935120917955. https://doi.org/10.1177/1176935120917955
5. RITA Registro Institucional de Tumores de Argentina Instituto Nacional del Cáncer. Ministerio de Salud, Argentina. RESULTADOS, AVANCES Y DESAFÍOS Período 2012-2018. https://bancos.salud.gob.ar/sites/default/files/2021-05/2021-05-26-publicacion-RITA-digital2.pdf
6. Viniegra M, Paolino M, Arrosi S (2010) Organización Panamericana de la Salud. Cáncer de mama en Argentina: organización, cobertura, y calidad de las acciones de prevención y control. Informe final julio 2010: diagnóstico de situación del Programa Nacional y Programas Provinciales 1–141. http://www.msal.gov.ar/inc/publicaciones
7. MacKinnon J A, Duncan RC, Huang Y, Lee D J, Fleming L E, Voti L, Rudolph M, Wilkinson, J D (2007) Detecting an association between socioeconomic status and late stage breast cancer using spatial analysis and area-based measures. Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology, 16(4), 756–762. https://doi.org/10.1158/1055-9965.EPI-06-0392).
8. Mehta, N., & Pandit, A. (2018). Concurrence of big data analytics and healthcare: A systematic review. International journal of medical informatics, 114, 57–65. https://doi.org/10.1016/j.ijmedinf.2018.03.013
9. Rakesh Raja, Indrajit Mukherjee, Bikash Kanti Sarkar, "A Systematic Review of Healthcare Big Data", Scientific Programming, vol. 2020, Article ID 5471849, 15 pages, 2020. https://doi.org/10.1155/2020/5471849

10. Mallappallil, M., Sabu, J., Gruessner, A., & Salifu, M. (2020). A review of big data and medical research. SAGE open medicine, 8, 2050312120934839. https://doi.org/10.1177/2050312120934839
11. Ristevski, B., & Chen, M. (2018). Big Data Analytics in Medicine and Healthcare. *Journal of integrative bioinformatics*, *15*(3), 20170030. https://doi.org/10.1515/jib-2017-0030
12. Ienca, M., Ferretti, A., Hurst, S., Puhan, M., Lovis, C., & Vayena, E. (2018). Considerations for ethics review of big data health research: A scoping review. *PloS one*, *13*(10), e0204937. https://doi.org/10.1371/journal.pone.0204937
13. Price, W. N., 2nd, & Cohen, I. G. (2019). Privacy in the age of medical big data. Nature medicine, 25(1), 37–43. https://doi.org/10.1038/s41591-018-0272-7
14. Mittelstadt, B. D., & Floridi, L. (2016). The Ethics of Big Data: Current and Foreseeable Issues in Biomedical Contexts. *Science and engineering ethics*, *22*(2), 303–341. https://doi.org/10.1007/s11948-015-9652-2.
15. Yang, A., Troup, M., & Ho, J. (2017). Scalability and Validation of Big Data Bioinformatics Software. Computational and structural biotechnology journal, 15, 379–386. https://doi.org/10.1016/j.csbj.2017.07.002
16. Miller, D. D., & Brown, E. W. (2018). Artificial Intelligence in Medical Practice: The Question to the Answer?. *The American journal of medicine*, *131*(2), 129–133. https://doi.org/10.1016/j.amjmed.2017.10.035
17. Ministerio de Salud y Desarrollo Social. 4taEncuesta Nacional de Factores de riesgo, 2020. https://bancos.salud.gob.ar/sites/default/files/2020-01/4ta-encuesta-nacional-factores-riesgo_2019_principales-resultados.pdf
18. García Fernández, A., Chabrera, C., García Font, M., Fraile, M., Lain, J. M., Gónzalez, S., Barco, I., González, C., Torres, J., Piqueras, M., Cirera, L., Veloso, E., Pessarrodona, A., & Giménez, N. (2015). Differential patterns of recurrence and specific survival between luminal A and luminal B breast cancer according to recent changes in the 2013 St Gallen immunohistochemical classification. Clinical & translational oncology : official publication of the Federation of Spanish Oncology Societies and of the National Cancer Institute of Mexico, 17(3), 238–246. https://doi.org/10.1007/s12094-014-1220-8
19. Sporikova, Z., Koudelakova, V., Trojanec, R., & Hajduch, M. (2018). Genetic Markers in Triple-Negative Breast Cancer. Clinical breast cancer, 18(5), e841–e850. https://doi.org/10.1016/j.clbc.2018.07.023
20. Zhang, A., Wang, X., Fan, C., & Mao, X. (2021). The Role of Ki67 in Evaluating Neoadjuvant Endocrine Therapy of Hormone Receptor-Positive Breast Cancer. Frontiers in endocrinology, 12, 687244. https://doi.org/10.3389/fendo.2021.687244
21. Cirqueira M B, Moreira M A, Soares L R, Cysneiros M A, Vilela M H, Freitas-Junior R (2015) Effect of Ki-67 on Immunohistochemical Classification of Luminal A to Luminal B Subtypes of Breast Carcinoma. The breast journal, 21(5), 465–472. https://doi.org/10.1111/tbj.12441
22. Davey M G, Hynes S O, Kerin M J, Miller N, Lowery A J (2021) Ki-67 as a Prognostic Biomarker in Invasive Breast Cancer. Cancers, 13(17), 4455. https://doi.org/10.3390/cancers13174455
23. Røge R, Nielsen S, Riber-Hansen R, Vyberg M (2021) Ki-67 Proliferation Index in Breast Cancer as a Function of Assessment Method: A NordiQC Experience. Applied immunohistochemistry & molecular morphology: AIMM, 29(2), 99–104. https://doi.org/10.1097/PAI.0000000000000846
24. Valencia D, Granda P, Pesce V et al. (2021) Argentina's National Program for Control of Breast Cancer: Time 1, Patient Navigation, and Patient Cancer Education Experience. J Canc Educ). https://doi.org/10.1007/s13187-021-02011-4
25. Jaksa A, Wu J, Jónsson P, Eichler H G, Vititoe S, Gatto N M (2021) Organized structure of real-world evidence best practices: moving from fragmented recommendations to comprehensive guidance. Journal of comparative effectiveness research, 10(9), 711–731. https://doi.org/10.2217/cer-2020-0228
26. Prat A, Pineda E, Adamo B, Galván P, Fernández A, Gaba L, et al. Clinical implications of the intrinsic molecular subtypes of breast cancer (2015) Breast. Nov;24 Suppl 2:S26-35.
27. Hammond ME, Hayes DF, Dowsett M, Allred DC, Hagerty KL, Badve S, et al. (2010) American Society of Clinical Oncology/College of American Pathologists Guideline Recommendations for Immunohistochemical Testing of Estrogen and Progesterone Receptors in Breast Cancer. J Clin Oncol. 28(16):2784-95.
28. Sorlie T, Tibshirani R, Parker J, Hastie T, Marron JS, Nobel A, et al. (2003) Repeated observation of breast tumor subtypes in independent gene expression data sets. Proc Natl Acad Sci U S A. 100(14):8418-23.

Tablas e imágenes:

Figura 1: Ingreso de un evento adverso basado en el CTCAE v5.0 en PrecisaXperta

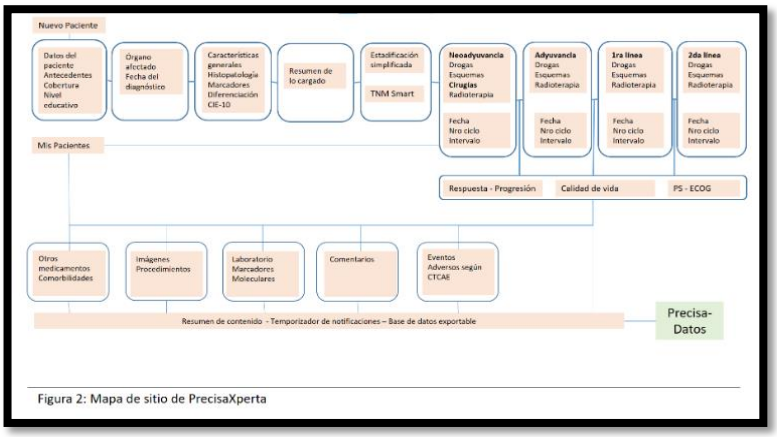

Figura 2: Mapa de sitio de PrecisaXperta

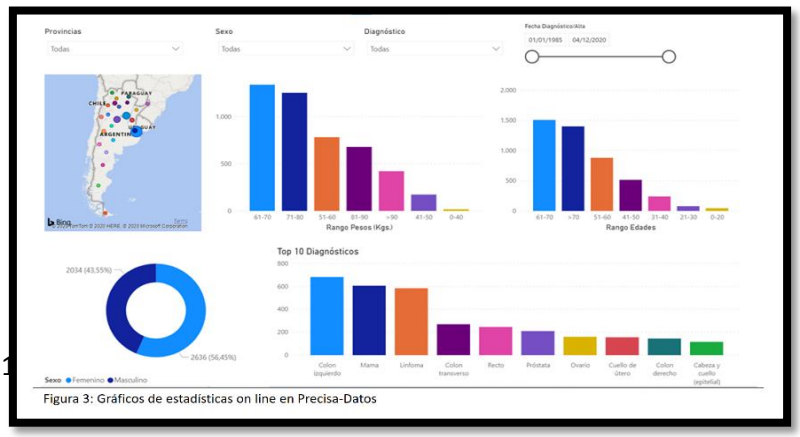

Figura 3: Gráficos de estadísticas on line en Precisa-Datos

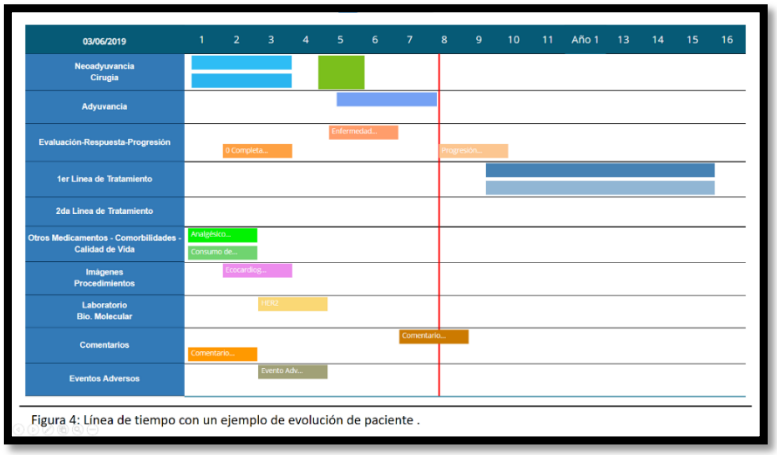

Figura 4: Línea de tiempo con un ejemplo de evolución de paciente .

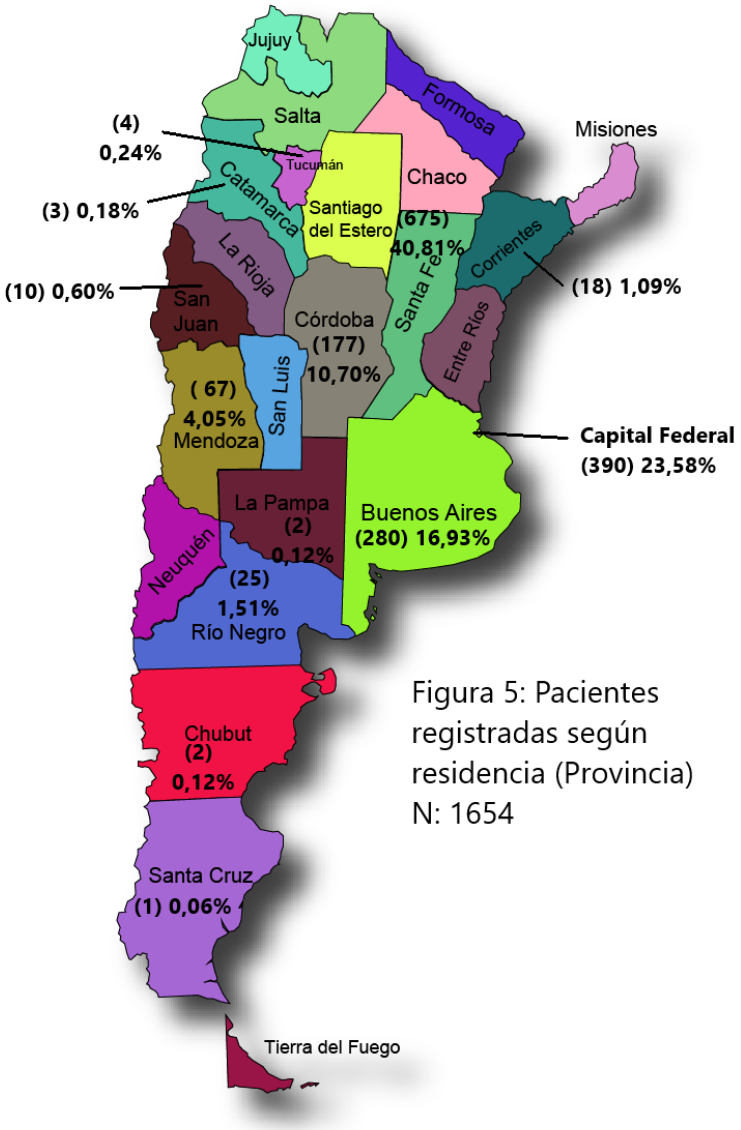

Figura 5: Pacientes registradas según residencia (Provincia)
N: 1654

Tablas nuevas editables

Tabla 1: Descripción de la muestra analizada (1654)

Antropometría	Media (ds)
Edad años)	55.3 (11.88)
Peso (Kgs)	67.8 (11.58)
Altura (cms)	163.8 (6.02)
Superficie Corporal (m^2)	1.69 (0.20)

Co-morbilidades. (n)	% del total
No registrada (391)	23.63
Diabetes (241)	14.57
Obesidad (180)	10.88
Tabaquismo (146)	8.82
Osteoporosis (123)	7.43
Depresión (89)	5.38
Hipertensión arterial (85)	5.13
Dispepsia (65)	3.92
Asma (41)	2.47
Otras* (293)	17.71
Formas de presentación	
Sintomático (510)	31%
Nódulo único (406)	24,70%
Hallazgo Incidental (254)	14.7%
Imagenológico (212)	12,90%
Biopsia (54)	3,30%
Múltiples nódulos (35)	2,10%
Cirugía (23)	1,40%
Siembra miliar (12)	0,70%
Metástasis cerebrales (7)	0,40%
Metástasis torácicas (5)	0,30%
Sin Datos (137)	8,30%

Tabla 2: Características tumorales de la muestra N: 1654

Estadificación. (n)	% del total
0 - in situ (177)	6,03
IA (157)	13,4
IB (281)	23,34
IIA (278)	22,92
IIB (85)	7,14
IIIA (98)	8,14

	IIIB (120)	9,97
	IIIC (18)	1,5
	IV (91)	7,56
Receptores Hormonales.	(n)	% del total
	R.H.Positivos (1424)	89,96%
	R.H.Negativos (159)	10,04%
Receptores del HER2.	(n)	% del total
	R. del HER2 Positivo (278)	19%
	R.del HER2 negativo (1209)	81,30%
Receptores de Proliferación celular.	(n)	% del total
	Marcador de P.Celular Ki67 Positivo (216)	13,60%
	Marcador de P.Celular Ki67 Negativo (135)	8,16%
	Marcador de P.Celular Sin Dato (1303)	78,78%

Tabla 3 Pacientes según tipo de tratamiento recibido* N:1654

Tratamientos	nro de pacientes	% del total	Drogas más Empleadas
Hormonoterapia	1018	61.5	Tamoxifeno; Letrozole; Anastrazole
Quimioterapia	1242	75.9	Ciclofosfamida; Paclitaxel; Adriamicina
Target Dirigidos	521	31.4	Trastuzumab; Palbociclib; Pertuzumab
Inmunoterapia	6	0.36	Atezolizumab

(*) Recibido en algún momento de su evolución, sin discriminar la etapa del tratamiento

Tabla 4: Evaluación de la Calidad de Vida según Estadios

Items considerados	IA - IB	IIA - IIB	IIIA - IIIB	IV
Actividad Cotidiana Completa	95%	96%	70%	20%
Actividad Cotidiana Limitada	5%	4%	20%	65%
Sin Dolor	70%	75%	60%	0%
Dolor No Invalidante	30%	25%	60%	sin datos
Movilidad Completa	95%	95%	92%	15%
Movilidad Limitada	5%	5%	8%	98%