Public Biomedical Datasets 2026: All of Us vs UK Biobank

Published August 1, 2026 36 min read

Public Biomedical Datasets 2026: All of Us vs UK Biobank

Executive Summary

Public biomedical datasets have become the backbone of life-sciences artificial intelligence (AI) development, and by August 2026 the landscape has consolidated around five dominant resources: the **All of Us Research Program**, **UK Biobank**, the **National COVID Cohort Collaborative (N3C)** and its clinical-data-commons peers, **MIMIC-IV**, and the **Surveillance, Epidemiology, and End Results (SEER)** cancer registry program. As of its June 2026 data release, All of Us reported data from more than **747,000 participants** and a total enrolled cohort exceeding **883,000** people, making it, in the National Institutes of Health's (NIH) own description, the world's largest integrated genomic and electronic health record (EHR) database (Source: www.nih.gov). UK Biobank, by contrast, is smaller by headcount at **503,317 volunteers** recruited between 2006 and 2010 (Source: www.ukbiobank.ac.uk), but it charges tiered access fees running from **£3,000** to **£9,000** for the first three years of a project (Source: www.ukbiobank.ac.uk), a stark contrast to All of Us, which is free to registered researchers regardless of institution (Source: www.nih.gov).

This report compares these programs and their close relatives, including **PCORnet** (more than **53.6 million** unique patients across 78-plus health systems as of mid-2026) (Source: pcornet.org) and the commercial federated network **TriNetX** (more than **309 million** patient records across 14,200-plus clinical sites as of early 2026) (Source: trinetx.com), across access requirements, cost, scale, and demonstrated AI/machine-learning (ML) suitability. **N3C**, stewarded by the National Center for Advancing Translational Sciences (NCATS), grew to an estimated **22.8 million** patient records by late 2024 (Source: lhncbc.nlm.nih.gov) but went fully offline on **September 27, 2025** amid an HHS platform-consolidation directive, with reopening targeted for 2026 (Source: ncats.nih.gov), a disruption that matters for any organization planning research timelines around it. **MIMIC-IV**, the free, credentialed critical-care dataset from the Massachusetts Institute of Technology (MIT) Laboratory for Computational Physiology and Beth Israel Deaconess Medical Center (BIDMC), covers over **65,000 ICU patients** and **200,000 emergency department patients** (Source: physionet.org) and requires only free human-subjects training and a signed data use agreement (Source: mimic.mit.edu). SEER, funded continuously by the National Cancer Institute (NCI) since 1973, covers approximately **45.9%** of the US population through its registries and remains free for its standard research files (Source: seer.cancer.gov) (Source: seer.cancer.gov).

The comparative analysis finds no single "best" dataset for AI training; instead, fitness for purpose depends on the modeling task. Genomic and pharmacogenomic discovery work favors All of Us and UK Biobank, both of which now include whole genome sequencing (WGS) data for large fractions of their cohorts and have anchored discoveries such as the **PIEZO1**-varicose vein association from UK Biobank exome data (Source: www.nature.com). Critical-care and clinical-notes AI, including sepsis-prediction and mortality models, rely almost exclusively on MIMIC-IV and its ICU peer, the eICU Collaborative Research Database (Source: physionet.org). Population-scale outcomes and pandemic-response research has depended on N3C, whose harmonized EHR data powered long-COVID prediction models reaching an area under the receiver operating characteristic curve (AUROC) of **0.92** (Source: pmc.ncbi.nlm.nih.gov), while oncology-focused survival modeling continues to draw on SEER. Regulatory relevance is real but modest: a peer-reviewed analysis of US Food and Drug Administration (FDA) approvals from January

2019 through June 2021 found real-world evidence (RWE) present in **116** of the approvals studied, with 65 of those RWE studies influencing the agency's final decision (Source: doi.org), and a later 2025 study of 218 labeling-expansion approvals found real-world evidence for 55 of them, of which only 3 were identified in documents available from FDA itself and 52 were found instead via ClinicalTrials.gov and PubMed (Source: pmc.ncbi.nlm.nih.gov), underscoring that real-world data (RWD) remains a supplement to, not a replacement for, controlled trial evidence.

Cost and governance discipline separate these resources as much as scale does. Market researchers disagree sharply on the size of the broader RWE solutions market, with 2026 estimates ranging from **$2.7 billion** (Source: www.researchandmarkets.com) to figures several times larger from other firms, a discrepancy this report presents transparently rather than resolving artificially. For organizations building AI systems on top of these datasets, including life-sciences consultancies such as intuitionlabs.ai that integrate multiple real-world data sources into client analytics pipelines using platforms like Databricks and Snowflake (Source: intuitionlabs.ai), the practical decision is rarely which single dataset to use but how to combine access tiers, cost structures, and governance obligations across several of them.

Introduction and Background

Life-sciences AI development in 2026 depends on a small set of publicly accessible, large-scale biomedical datasets that supply the labeled, longitudinal, and genomically linked data that proprietary electronic health record (EHR) systems rarely expose to outside researchers. Unlike a decade ago, when most machine learning research in health care used small, single-institution datasets, today's landscape is anchored by a handful of national and international programs designed explicitly for external reuse: the NIH's **All of Us Research Program**, the United Kingdom's **UK Biobank**, NCATS's **N3C**, MIT's **MIMIC-IV** critical-care database, and NCI's **SEER** cancer registry system. Around these five sit a second tier of clinical data commons and networks, including **PCORnet** and the commercial network **TriNetX**, that extend similar EHR-based research capability into different governance and business models.

This distinction matters because "public" does not mean uniformly free, and "biomedical dataset" does not mean a single data type. All of Us is free of charge to any registered researcher (Source: www.nih.gov), while UK Biobank charges published access fees exclusive of value-added tax (VAT) that were introduced on **1 April 2021** and apply per project per tier (Source: community.ukbiobank.ac.uk). MIMIC-IV is free but credentialed, requiring completion of a human-subjects training course and a signed Data Use Agreement (DUA) before any file is released (Source: mimic.mit.edu). SEER's headline research files are open to any user with a valid email address (Source: seer.cancer.gov), yet its linked SEER-Medicare product is not a public-use file and carries direct cost-recovery fees (Source: healthcaredelivery.cancer.gov). N3C, meanwhile, requires an institutional DUA with NCATS before any individual researcher application is even considered (Source: ncats.nih.gov).

These datasets also diverge in data type. All of Us and UK Biobank both combine WGS, EHR data, and survey or physical-measurement data, positioning them for genomic discovery and precision-medicine research. MIMIC-IV and the related eICU Collaborative Research Database focus narrowly on hospital and ICU episodes, with structured vitals, laboratory results, medications, and, in MIMIC-IV's case, linked chest radiographs, discharge notes, and electrocardiograms. N3C, PCORnet, and TriNetX aggregate structured EHR data across dozens to thousands of health systems using common data models, primarily supporting population-scale outcomes research rather than genomics. SEER is narrower still, a cancer-specific registry tracking incidence, stage, treatment, and survival rather than raw clinical encounters.

For AI and machine learning teams, this heterogeneity is the central practical fact: no dataset in this comparison is a universal substitute for the others, and building a defensible real-world evidence or model-training strategy typically means combining two or more of them under different governance regimes. This report examines each major dataset's capabilities, access process, cost, adoption, and limitations; presents a feature-comparison matrix and published performance benchmarks; and closes with the quantitative and case-study evidence needed to choose among them as of August 2026.

All of Us Research Program

Capabilities

The **All of Us Research Program** is an NIH initiative whose stated mission is to collect and study data from **one million or more** people living in the United States to accelerate precision-medicine research (Source: allofus.nih.gov). As of its June 30, 2026 data release, the program had grown to a total enrolled-participant count exceeding **883,000**, an increase of more than 114,000 from the prior release (Source: www.nih.gov), while data from over **747,000 participants** was actually available to researchers (Source: www.nih.gov). That release, designated **Controlled Tier Dataset v9 (CDRv9)**, included more than **535,000** whole genome sequences linked to nearly **482,000** EHRs (Source: www.nih.gov) and marked the program's first entry into multiomics, adding proteomics data for nearly 10,000 participants, RNA-sequencing data for nearly 9,000, and long-read WGS for over 14,500 (Source: www.nih.gov).

Data access runs through the cloud-based **Researcher Workbench**, split into a **Registered Tier** and a **Controlled Tier**. The Registered Tier includes EHR data, wearables such as Fitbit device data, surveys, and physical measurements (Source: www.researchallofus.org), while the Controlled Tier adds genomic data, including short-read WGS, long-read WGS, structural variants, and genotyping arrays, along with unshifted event dates (Source: www.researchallofus.org). Only the three most recent Curated Data Repository (CDR) versions remain available in the Workbench at any time (Source: support.researchallofus.org).

Adoption

Access is free of charge to registered researchers at any institution, a design choice the NIH frames explicitly as an equity measure: "**All of Us** data is available to registered researchers at no cost, giving scientists at rural universities the same access as those at major research institutions" (Source: www.nih.gov). Cloud compute is not entirely free, however: each registered Researcher Workbench user receives **$300** in initial credits (Source: www.researchallofus.org), which expire 365 days after the researcher completes Responsible Conduct of Research (RCR) training and signs the Data User Code of Conduct (Source: support.researchallofus.org); after that, further Google Cloud Platform (GCP) usage is billed to the researcher's own account. RCR training is mandatory for Registered Tier access, with an additional module required before Controlled Tier data can be used, renewed annually (Source: support.researchallofus.org). More than **1,200 institutions** now hold a master DUA with the program, which simplifies onboarding for their researchers (Source: www.genomeweb.com). By June 2026 the program's data had fueled more than **1,400 peer-reviewed publications** by nearly **23,000 researchers** (Source: www.nih.gov).

A distinctive strength of the program is its diversity mandate: **86%** of participants, more than 645,000 people, come from communities historically underrepresented in biomedical research (Source: www.nih.gov). A 2024 Nature paper describing an earlier data release documented **245,388** clinical-grade genome sequences, of which 77% came from historically underrepresented participants, and reported more than **275 million** previously unreported genetic variants identified in the cohort (Source: www.nature.com) (Source: www.nature.com). The same study replicated known disease associations across ancestries with high fidelity, evaluating 3,724 genetic variants tied to 117 diseases (Source: www.nature.com).

Strengths and Limitations

All of Us's principal strengths are its cost model, its diversity, and its unusually broad data modalities within a single governed platform. Its principal limitation is a growing gap between consent and completeness: although **98%** of participants agree in principle to share EHR data, more than **300,000** participants have no EHR data at all in the database, a completeness gap the program is actively working to close (Source: www.statnews.com). A separate academic bias analysis found the cohort skews older, more female, more educated, more insured, less White, and less healthy than the general US population, a pattern consistent with volunteer self-selection common to all such cohorts (Source: www.medrxiv.org). Funding volatility is also a real constraint: a **75%** cut to the program's budget in the prior year sharply slowed enrollment growth, moving consented participants from roughly 860,000 to roughly 873,000 between April 2025 and January 2026 (Source: www.genomeweb.com), a reminder that public-dataset roadmaps are subject to appropriations risk as much as scientific planning.

UK Biobank

Capabilities

UK Biobank is a prospective cohort of **500,000** United Kingdom participants who agreed to make their health-related data available for research, recruited between ages 40 and 69 (Source: www.ukbiobank.ac.uk). Baseline recruitment ran from 2006 to 2010 and enrolled exactly **503,317 volunteers** (Source: www.ukbiobank.ac.uk). The resource has since expanded far beyond baseline questionnaires: on **30 November 2023**, UK Biobank released whole genome sequencing data for all 500,000 participants, described by the organization as the biggest whole-genome dataset in the world (Source: community.ukbiobank.ac.uk). Subsequent releases added proteomics data for 54,000 participants and metabolomics data for 300,000 participants in 2023, and by 2025 expanded metabolomics to the full 500,000-participant cohort and imaging data to 100,000 participants (Source: www.ukbiobank.ac.uk). A separate imaging enhancement, begun in 2014, re-invited 100,000 of the original 500,000 participants for brain, cardiac, and abdominal magnetic resonance imaging (MRI), bone-density scanning, and carotid ultrasound (Source: www.nature.com); that imaging effort reached a milestone of 100,000 completed scans in **July 2025**, generating more than one billion individual images and supporting more than 1,300 peer-reviewed papers (Source: www.ukbiobank.ac.uk). Linked EHR data includes primary-care records for roughly 230,000 participants (through 2016 or 2017 depending on data supplier) and hospital inpatient, cancer, and death-registry data for the full 500,000-person cohort (Source: www.ukbiobank.ac.uk).

Adoption

Data access follows an application-based process through UK Biobank's Access Management System, requiring affiliation with a recognized research organization, background eligibility checks, and a signed Material Transfer Agreement before data are released (Source: www.ukbiobank.ac.uk). Since January 2025, insurance companies are no longer approved for direct data access, a governance tightening reflected in the same policy page (Source: www.ukbiobank.ac.uk). Access is explicitly tiered by cost: **Tier 1** costs **£3,000** for the first three years plus **£1,000** per year for extensions; **Tier 3**, which is required for whole genome sequence data, whole exome sequence data, and imaging data, costs **£9,000** for the first three years plus **£3,000** per year to extend (Source: www.ukbiobank.ac.uk). These fees, exclusive of VAT, were introduced on 1 April 2021 and apply to access delivered through the UK Biobank Research Analysis Platform (UKB-RAP) (Source: community.ukbiobank.ac.uk). Adding an extra collaborating institution to an application costs a further **£1,000** for the first three years, while lower-income-country and student researchers can apply reduced rates of **£500** (Source: www.ukbiobank.ac.uk).

As of mid-2026, UK Biobank had paused new access applications while upgrading its Research Analysis Platform; its own summary states that over **22,000 researchers** have accessed the data and more than **18,000 peer-reviewed papers** have used it (Source: www.ukbiobank.ac.uk). The organization's FY2025 annual report states that participants have donated **30 petabytes** and 16 million biological samples in total, and that researchers in **60 countries** are actively working on the

data (Source: www.ukbiobank.ac.uk). As of April 2026, the project tracker listed **6,934** total approved research projects, of which 5,319 were current and 1,614 closed (Source: www.ukbiobank.ac.uk).

Strengths and Limitations

UK Biobank's greatest strength for AI training is depth: few other resources combine WGS, whole-exome sequencing, multi-organ imaging, proteomics, metabolomics, and decades of linked administrative health records in a single governed dataset. Industry partnership is a core part of that depth. A public-private consortium including eight biopharmaceutical companies, the UK Biobank Exome Sequencing Consortium, produced exome data on **200,643** participants specifically as a drug-discovery resource (Source: www.nature.com), and the first exome tranche of 49,960 participants, generated by the Regeneron Genetics Center in collaboration with GlaxoSmithKline, identified novel loss-of-function variants including a **PIEZO1**-varicose vein association and found that roughly **2%** of the studied population carried a medically actionable genetic variant (Source: www.nature.com). By January 2023, Regeneron's ongoing UK Biobank collaboration and related efforts had reached approximately **2 million** collaboration-participant exomes sequenced, and in 2025 the two organizations launched a Pharma Proteomics Project described as the world's most comprehensive study of proteins, funded by Regeneron and 13 other biopharmaceutical companies (Source: www.regeneron.com) (Source: www.regeneron.com).

The primary limitation is a well-documented "healthy volunteer" selection bias. Only **5.5%** of the 9.2 million UK adults aged 40 to 69 invited to the baseline assessment actually participated (Source: www.ukbiobank.ac.uk), and a peer-reviewed representativeness study found that, at ages 70 to 74, all-cause mortality was **46.2%** lower in men and **55.5%** lower in women among UK Biobank participants than in the general population (Source: pubmed.ncbi.nlm.nih.gov). The WGS project itself, spanning 2019 to 2023, cost approximately **£200 million** (roughly **$254 million**), co-funded by UK Research and Innovation, the Wellcome Trust, and four pharmaceutical companies (Source: www.genomeweb.com), illustrating that even a free-to-view registry can carry a substantial capital cost passed on to future users through access fees.

N3C and Clinical Data Commons

Capabilities

The **National COVID Cohort Collaborative (N3C)**, stewarded by NCATS with contributions from more than **90 institutions**, began in 2020 as an open-science community for pooling patient-level EHR data during the COVID-19 pandemic (Source: ncats.nih.gov) (Source: pmc.ncbi.nlm.nih.gov). NCATS has since said it is leveraging the original N3C infrastructure to expand into new disease domains and cover general real-world-data training needs beyond COVID-19 (Source: ncats.nih.gov). N3C harmonizes contributed EHR data into the OMOP Common Data Model (Source: ncats.nih.gov) and offers three data tiers: a Limited Data Set, a De-identified Data Set, and a Synthetic Data Set that resembles patient information statistically without containing actual patient data (Source: ncats.nih.gov). By one program blog account from March 2025, the repository had grown to roughly **23 million** patient records and 33 billion rows of data, calling it the largest longitudinal open-science clinical data resource globally (Source: ccos-cc.ctsa.io), a figure broadly consistent with a technical profile of Release 184 recording **22,854,489** patients and 746,939 deceased-patient records (Source: lhncbc.nlm.nih.gov).

Two related, non-federal networks extend the same clinical-data-commons concept. **PCORnet**, funded by the Patient-Centered Outcomes Research Institute (PCORI), describes itself as a large distributed "network of networks" connecting researchers to health data from roughly **47 million** people annually through eight Clinical Research Networks and more than 70 partner organizations (Source: pcornet.org), with a live population count of **53,658,361** unique patients, current as of this report's research date, standardized across more than 78 health systems using its own Common Data Model (Source: pcornet.org) (Source: pcornet.org). **TriNetX**, founded in 2013, is a commercial federated real-world data network (Source: trinetx.com) that, as of early 2026, reported a global network exceeding **309 million** patient lives across more than 14,200 clinical sites in over 20 countries, with more than 4,000 peer-reviewed publications citing its data (Source: trinetx.com).

Adoption

N3C access requires an institutional-level Data Use Agreement with NCATS before any individual researcher's application can proceed, and those institutional agreements remain valid for five years (Source: ncats.nih.gov) (Source: ncats.nih.gov). Individual data use requests, once approved by N3C's Data Access Committee, are valid for one year and renewable (Source: ncats.nih.gov). NCATS reports more than **5,000 citations** and an h-index of 33 generated from N3C-enabled research, with 1,589 contributing authors (Source: ncats.nih.gov). Materially, however, the platform experienced a significant operational disruption: it went fully offline on **September 27, 2025**, tied to a Department of Health and Human Services (HHS) platform-consolidation directive, with reopening to the broader research community targeted for 2026 (Source: ncats.nih.gov). Any organization planning research against N3C in 2026 needs to treat that timeline risk as a first-order planning input rather than a footnote.

PCORnet offers no-cost initial consultations, including feasibility reviews, to researchers considering the network (Source: pcornet.org). TriNetX markets feasibility and cohort-discovery querying across its federated network, built, in the company's own words, on the idea that better access to real-world data could transform research and improve patient lives (Source: trinetx.com); access is typically bundled into an institutional subscription, so individual researchers at subscribing sites query at no incremental cost, while custom, patient-level data pulls are billed separately, according to one academic medical center's public reference page (Source: researchroadmap.mssm.edu).

Strengths and Limitations

The clinical-data-commons model's chief strength is scale achieved through federation rather than a single recruitment effort: N3C, PCORnet, and TriNetX each aggregate tens to hundreds of millions of patient records without requiring any single institution to run its own biobank. The corresponding limitation is data harmonization risk. A peer-reviewed critique published in JAMA Network Open warned that combining data across institutions "creates complex issues related to standardization, obfuscates site heterogeneity, and, most importantly, disconnects the data from the people who know them best" (Source: doi.org). N3C remains in a transition period as NCATS restores features in phases after reopening the Data Enclave; meanwhile, TriNetX's commercial model means the deepest data access requires an institutional subscription rather than open registration, unlike All of Us, MIMIC-IV, or standard SEER data.

MIMIC-IV and Critical Care Datasets

Capabilities

MIMIC-IV (Medical Information Mart for Intensive Care) is a deidentified dataset of patients admitted to the emergency department (ED) or an intensive care unit (ICU) at Beth Israel Deaconess Medical Center (BIDMC) in Boston, produced by the MIT Laboratory for Computational Physiology (Source: physionet.org). Version **3.1**, the current release as of the research date, was published **October 11, 2024** (Source: physionet.org) and contains data for over **65,000** patients admitted to an ICU and over **200,000** patients admitted to the ED (Source: physionet.org). In version 3.0, a total of **364,627** unique individuals accounted for 546,028 hospitalizations and 94,458 unique ICU stays (Source: physionet.org), covering admissions from 2008 through 2022 (Source: physionet.org). All dates are deidentified by shifting them into a fictional future window between 2100 and 2200, a detail AI teams must account for when computing patient age or seasonal trends (Source: physionet.org).

Beyond structured tabular data, PhysioNet hosts several linked MIMIC modules relevant to multimodal AI: **MIMIC-CXR** contains 227,835 imaging studies and 377,110 images for roughly 65,000 patients presenting to the BIDMC emergency department between 2011 and 2016 (Source: www.nature.com); **MIMIC-IV-Note** contains 331,794 deidentified discharge summaries from 145,915 patients and 2,321,355 deidentified radiology reports from 237,427 patients (Source: physionet.org); and **MIMIC-IV-ECG** contains approximately **800,000** diagnostic 12-lead electrocardiograms across nearly **160,000** unique patients matched to the clinical database (Source: physionet.org).

Adoption

Accessing MIMIC requires completing human-subjects training and signing a Data Use Agreement, a straightforward three-step credentialing process built into the PhysioNet platform (Source: mimic.mit.edu). PhysioNet recommends the CITI Program's "Data or Specimens Only Research" course for that training (Source: physionet.org), and that course is free of charge when taken under the "Massachusetts Institute of Technology Affiliates" option (Source: physionet.org). PhysioNet itself offers free web access to its large collections of recorded physiologic signals (Source: archive.physionet.org), and the credentialed license explicitly prohibits attempts to re-identify individuals or institutions referenced in restricted data (Source: www.physionet.org), and, notably, prohibits sharing credentialed data with third parties, including sending it through application programming interfaces (APIs), a restriction directly relevant to teams hoping to pipe MIMIC data into hosted large language model (LLM) services (Source: physionet.org).

Adoption metrics are substantial: MIMIC-IV v3.1 alone recorded **22,787** unique registered viewers of that version and **51,381** across all MIMIC-IV versions (Source: physionet.org). Across the whole PhysioNet platform, more than **15,000** scientific publications cited PhysioNet resources in the most recent year, with registered users from more than **180 countries** (Source: news.mit.edu). A dedicated bibliometric analysis of MIMIC-specific research found **2,769** MIMIC-related publications between 2004 and 2024, with **40.6%** annual growth, and identified AI and machine-learning research, including sepsis and mortality prediction, as the dominant research theme (Source: www.sciencedirect.com).

Strengths and Limitations

MIMIC-IV's strengths are cost, richness of linked modalities, and its unusually long track record of reproducible AI benchmarking. Its central limitation, shared with the comparable **eICU Collaborative Research Database**, is single-institution or narrow-network scope: eICU comprises over 200,000 patient-unit encounters for over 139,000 unique patients admitted between 2014 and 2015 across 335 units at 208 hospitals throughout the United States, and requires the same proof of human-research training and signed data use agreement as MIMIC (Source: physionet.org) (Source: physionet.org). Predecessor dataset MIMIC-III similarly covered over 40,000 ICU patients at BIDMC between 2001 and 2012 (Source: mimic.mit.edu), underscoring that the entire MIMIC lineage represents US critical care at one health system rather than a nationally representative sample, which limits external generalizability of any model trained purely on it, a limitation the field increasingly addresses by validating models across MIMIC, eICU, and other cohorts simultaneously.

SEER and Cancer Registry Data

Capabilities

SEER, the Surveillance, Epidemiology, and End Results Program, is the NCI's authoritative source of information on cancer incidence and survival in the United States (Source: seer.cancer.gov). It has been continuously funded by the NCI since 1973 (Source: seer.cancer.gov) and is structured around 17 Core registries and 11 Research Support registries spanning 22 geographic areas (Source: seer.cancer.gov). SEER registries currently cover approximately **45.9%** of the US population (Source: seer.cancer.gov), though coverage varies notably by racial and ethnic group, spanning **39.6%** of Whites, **43.5%** of African Americans, and **64.9%** of Hispanics (Source: seer.cancer.gov). A separate 2024 metrics factsheet described the 17 core registries alone as representing **48%** of the US population and collecting approximately **950,000** new cancer cases per year, a discrepancy from the 45.9% figure that likely reflects different registry-count snapshots and is worth flagging rather than reconciling artificially (Source: seer.cancer.gov). SEER routinely collects patient demographics, primary tumor site, tumor morphology, stage at diagnosis, first course of treatment, and follow-up for vital status (Source: seer.cancer.gov). The current data release is the **November 2025 submission**, spanning diagnosis years 1975 through 2023 with a follow-up cutoff date of December 31, 2023 (Source: seer.cancer.gov), and the required SEER*Stat analysis software's latest release, version **9.0.43**, dates to March 27, 2026 (Source: seer.cancer.gov).

Adoption

Standard SEER Research Data is accessible to any user with a valid email address after completing a registration and agreement process, with no fee stated for that tier (Source: seer.cancer.gov). The more detailed SEER Research Plus and Novel Cancer Control Research (NCCR) datasets instead require an eRA Commons or HHS account, with access restricted for users located in what SEER's documentation terms "countries of concern" (Source: seer.cancer.gov). Since 1975, roughly **25,000** publications have used SEER data as a primary data source, with more than **126,000** publications referencing SEER data overall (Source: seer.cancer.gov) (Source: seer.cancer.gov).

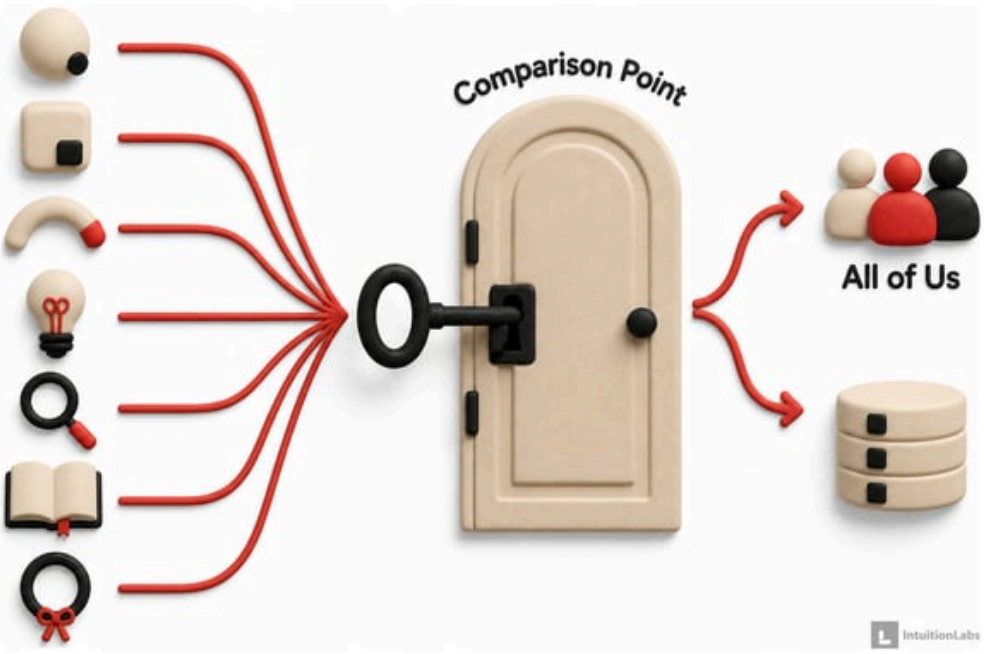
The linked **SEER-Medicare** database is a different product entirely: it is explicitly not a public-use data file, and investigators must submit a project-specific application and sign a Data Use Agreement to obtain it (Source: healthcaredelivery.cancer.gov). Unlike free standard SEER files, SEER-Medicare carries direct cost-recovery fees set by NCI's contractor, for example **$350** for a cancer data file and **$200** for the 5% cancer file as of a July 2026 pricing check, with the notice that these fees can change without notice (Source: healthcaredelivery.cancer.gov).

Strengths and Limitations

SEER's strength for AI model training is its long, standardized, population-based longitudinal record of cancer outcomes, ideal for survival-prediction and prognosis modeling. Its limitation is treatment-data completeness: a peer-reviewed validation study found only moderate sensitivity when comparing SEER's treatment fields against Medicare claims, at **68%** for chemotherapy, **80%** for radiation, and **69%** for hormone therapy, and NCI does not release chemotherapy or hormone-therapy data at all due to uncertainty about its completeness (Source: pubmed.ncbi.nlm.nih.gov) (Source: pubmed.ncbi.nlm.nih.gov). Its geographic coverage, at roughly 45 to 48% of the US population depending on the registry snapshot used, also means SEER-derived models require careful external validation before generalizing to non-SEER regions.

Feature Comparison

Table 1 below summarizes the seven resources discussed in this report across governance, scale, data types, access model, and cost, providing a single reference point for the comparison questions this report set out to answer.


Comparison Point
All of Us
IntuitionLabs

DATASET	STEWARD / FOUNDED	APPROX. SCALE	CORE DATA TYPES	ACCESS MODEL	COST	KEY LIMITATION
All of Us	NIH; ongoing since 2018	883,000+ enrolled, 747,000+ with usable data (Source: www.nih.gov)	WGS, EHR, surveys, wearables, physical measurements	Free registration, Registered/Controlled Tiers via Researcher Workbench	Free data; $300 initial cloud credit, then self-billed (Source: www.researchallofus.org)	EHR completeness gap; funding volatility
UK Biobank	UK Biobank charity/UKRI; recruited 2006-2010	503,317 participants (Source: www.ukbiobank.ac.uk)	WGS, exome, imaging, proteomics, metabolomics, linked EHR	Application via Access Management System, Material Transfer Agreement	£3,000 to £9,000 for 3 years by tier plus annual extension fees (Source: www.ukbiobank.ac.uk)	Healthy-volunteer selection bias; paused applications mid-2026
N3C	NCATS/NIH; est. 2020	~22.8 million patients (Release 184) (Source: lhncbc.nlm.nih.gov)	Harmonized EHR (OMOP CDM), 3 access tiers	Institutional DUA (5-yr) plus individual Data Use Request (1-yr, renewable)	No published fee found; governed by NIH/NCATS agreements	Reopened in 2026; features returning in phases and new-user access pending new agreements (N3C restart guide
MIMIC-IV	MIT LCP / BIDMC	65,000+ ICU, 200,000+ ED patients (v3.1) (Source: physionet.org)	ICU/ED vitals, labs, meds, notes, CXR, ECG	PhysioNet credentialing: training plus DUA	Free; CITI course free for MIT affiliates (Source: physionet.org)	Single-institution, US critical-care-only scope
SEER	NCI; funded since 1973 (Source: seer.cancer.gov)	~45.9 to 48% of US population, ~950,000 new cases/yr	Cancer incidence, stage, treatment, survival	Any valid email for standard data; eRA/HHS account for Research Plus	Free (standard); SEER-Medicare carries cost-recovery fees (e.g., $350) (Source: healthcaredelivery.cancer.gov)	Incomplete chemo/hormone-therapy fields
PCORnet	PCORI; network of networks	53.6 million+ unique patients (live count) (Source: pcornet.org)	Standardized EHR (PCORnet CDM) across 78+ health systems	Query through PCORnet CRNs; no-cost initial consult (Source: pcornet.org)	Feasibility free; full studies negotiated per network	Distributed governance across 8 separate networks
TriNetX	Commercial; founded 2013 (Source: trinetx.com)	309 million+ patient lives, 14,200+ sites (Source: trinetx.com)	Federated EHR network, feasibility querying	Institutional subscription	Subscription-bundled; custom queries billed separately (illustrative) (Source: researchroadmap.mssm.edu)	No public list pricing; commercial gatekeeping

The matrix makes clear that "public" datasets sit on a spectrum from genuinely free and open (SEER's standard files, MIMIC-IV) through free-but-credentialed (All of Us, N3C) to fee-bearing by design (UK Biobank, SEER-Medicare) to fully commercial (TriNetX). For an organization deciding **all of us vs uk biobank**, the practical trade-off is direct: All of Us costs nothing beyond a small initial cloud-compute allowance and offers stronger US demographic diversity, while UK Biobank costs thousands of pounds per project but offers a longer-running, more deeply phenotyped, imaging-rich cohort with a track record of pharmaceutical industry co-investment. Neither dominates the other; the choice depends on whether a project needs UK-specific phenotyping depth or US population diversity and zero direct licensing cost.

Performance and Benchmarks

Published, peer-reviewed benchmark results on these datasets cluster heavily around two use cases: pandemic-era outcomes prediction using N3C, and ICU physiological forecasting using MIMIC-IV and its peer cohorts. *Table 2* below summarizes three representative, independently published results.

STUDY / MODEL	DATASET(S) USED	TASK	REPORTED PERFORMANCE
UC Berkeley, UNC Chapel Hill, N3C Consortium (RECOVER Initiative)	N3C, 8 million+ patients (Source: pmc.ncbi.nlm.nih.gov)	XGBoost classification of probable long-COVID patients	AUROC 0.92 (all), 0.90 (hospitalized), 0.85 (non-hospitalized) (Source: pmc.ncbi.nlm.nih.gov)
MIMIC-Sepsis benchmark (2025 preprint)	MIMIC-IV, 35,239-patient curated cohort	Sepsis trajectory modeling: mortality, length-of-stay, shock onset	Adding treatment variables (vasopressors, fluids, ventilation) substantially improved Transformer-based model performance (Source: arxiv.org)
npj Digital Medicine cross-cohort study (2026)	HiRID, MIMIC-IV, eICU; 216,536 stays combined (Source: www.nature.com)	Deep-learning sepsis prediction under distribution shift	Benchmarked model generalization across three independently sourced ICU cohorts

These results illustrate a broader pattern: models trained on a single dataset, however large, tend to be evaluated for external validity by testing across at least one additional, independently sourced cohort, whether that is MIMIC-IV validated against eICU and HiRID, or a SEER-based survival model validated against an external cohort, as in a 2026 study that built machine-learning survival models for breast cancer using SEER data from 2010 to 2020 and validated them on an independent cohort from China (Source: link.springer.com). No single dataset in this comparison functions as a self-sufficient benchmark; each is best understood as one leg of a multi-cohort validation strategy.

Data Analysis and Evidence

Market sizing for real-world data and evidence remains inconsistent across research firms, itself a useful data point about the maturity of this space. *Table 3* below presents three independently published 2026 estimates for the global real-world evidence (RWE) solutions market.

RESEARCH FIRM	MARKET SIZE ESTIMATE	PROJECTION	CAGR
Grand View Research	$3.04 billion (2025) (Source: www.grandviewresearch.com)	$6.04 billion by 2033	Not stated in excerpt
MarketsandMarkets	$4.7 billion (2024) (Source: www.marketsandmarkets.com)	$10.8 billion by 2030	14.8%
Research and Markets	$2.33 billion (2025) (Source: www.researchandmarkets.com)	$4.81 billion by 2030	15.9%

The roughly two-fold spread between the smallest (Research and Markets) and largest (MarketsandMarkets) base-year figures reflects differing definitions of "real-world evidence solutions," ranging from narrow software-licensing scopes to broader services-inclusive definitions, and this report presents that discrepancy directly rather than picking a winner. What all three estimates agree on is directional growth in the mid-teens percentage range annually, consistent with rising demand for the datasets profiled in this report.

Regulatory uptake of real-world evidence remains real but bounded. A peer-reviewed analysis of FDA approvals from January 2019 through June 2021 found that **116** approvals incorporated RWE in any form, of which 88 included an RWE study intended to provide evidence of safety or effectiveness, and 65 of those studies influenced the agency's final decision (Source: doi.org). A more recent 2025 study covering January 2022 through May 2024 identified real-world evidence for 55 of 218 labeling-expansion approvals reviewed, of which 3 were identified in documents available from FDA and 52 were found from ClinicalTrials.gov and PubMed, with the evidence concentrated in oncology (Source: pmc.ncbi.nlm.nih.gov). That same study found that most RWE studies used in FDA submissions were retrospective (65.9%), used a cohort design (87.5%), and drew on EHR data (75.0%) (Source: pmc.ncbi.nlm.nih.gov), a data-source profile that maps almost exactly onto the datasets this report profiles.

Two structural forces explain rising interest in these datasets for AI work specifically. First, whole genome sequencing costs have collapsed: Illumina now markets a "**$200 genome**" on its NovaSeq X Series platform (Source: www.illumina.com), which is precisely why UK Biobank and All of Us can now afford population-scale WGS as a standard cohort feature rather than a specialty add-on. Cloud compute for processing that sequence data has its own colloquial benchmark: the "**five dollar genome**" describes the typical Google Cloud Platform compute charge for aligning and calling variants on a single 30x whole-genome sample via the Terra platform used by several of these programs (Source: support.terra.bio). Second, the volume of AI research applied to these datasets is accelerating quickly: a 2026 bibliometric review found PubMed-indexed AI-in-healthcare publications reached **49,394** in 2025, nearly double the 28,180 recorded in 2024 (Source: www.medrxiv.org), with multimodal foundation-model publications specifically surging from 25 in 2024 to 144 in 2025 (Source: www.medrxiv.org). NIH's Data Management and Sharing Policy, effective **January 25, 2023**, has also formalized expectations that federally funded biomedical research make its underlying data available, reinforcing the pipeline of new datasets entering this ecosystem (Source: public-inspection.federalregister.gov).

Case Studies and Real-World Examples

UK Biobank Exome Data and Drug Target Discovery

The **Regeneron Genetics Center**, working with GlaxoSmithKline, generated the first large-scale exome-sequencing tranche of UK Biobank, covering 49,960 participants, and used it to identify novel loss-of-function genetic variants with large effects on disease traits, including an association between the **PIEZO1** gene and varicose veins (Source: www.nature.com). The same analysis found that approximately 2% of the studied population carried a medically actionable genetic variant, information with direct clinical and drug-target relevance (Source: www.nature.com). The collaboration subsequently expanded into a formal UK Biobank Exome Sequencing Consortium spanning eight biopharmaceutical companies and 200,643 sequenced participants (Source: www.nature.com), and by 2025 Regeneron and UK Biobank had launched a Pharma Proteomics Project funded by 14 biopharmaceutical companies, which the companies describe as the most comprehensive study of proteins undertaken to date (Source: www.regeneron.com). This case illustrates how a public biobank's value compounds over time: the same underlying cohort supported an initial exome discovery, a consortium-scale expansion, and, years later, an entirely new proteomics initiative.

All of Us and Cross-Ancestry Genomic Discovery

A 2024 Nature paper describing an All of Us data release documented 245,388 clinical-grade whole genome sequences, 77% of which came from participants historically underrepresented in biomedical research, and identified more than 275 million previously unreported genetic variants (Source: www.nature.com). Researchers evaluated 3,724 genetic variants associated with 117 diseases and found high replication rates in both European-ancestry and African-ancestry participants, a direct demonstration of the diversity mandate translating into scientific value rather than remaining a stated policy goal (Source: www.nature.com). Separately, a 2025 Annual Review of Pharmacology and Toxicology paper by researchers including Empey, Karnes, and Johnson framed the All of Us Research Program as a key opportunity to advance pharmacogenetics research precisely because of the scale and diversity of its combined genomic and EHR data (Source: pmc.ncbi.nlm.nih.gov), while a related preprint characterized pharmacogenomic variant frequency and medication exposure specifically within the All of Us cohort (Source: www.medrxiv.org).

N3C-Powered Long-COVID Prediction Models

Researchers at UC Berkeley and UNC Chapel Hill, working within the N3C Consortium, built machine-learning models on harmonized EHR data from more than **8 million** demographically diverse and geographically distributed patients to identify probable long-COVID patients (Source: pmc.ncbi.nlm.nih.gov). Their XGBoost-based classification models achieved AUROC values of 0.92 across all patients, 0.90 among hospitalized patients, and 0.85 among non-hospitalized patients (Source: pmc.ncbi.nlm.nih.gov), funded through the NIH and NCATS via the RECOVER Initiative (Source: pmc.ncbi.nlm.nih.gov). A subsequent 2025 study published in PLOS Medicine used a target-trial-emulation design built on N3C and RECOVER data to examine whether Paxlovid treatment during acute COVID-19 reduced the onset of Long COVID, an example of N3C supporting causal-inference research rather than pure prediction (Source: journals.plos.org).

MIMIC-IV Sepsis Benchmarking Across ICU Cohorts

A 2025 benchmark paper, "MIMIC-Sepsis," curated a 35,239-patient cohort from MIMIC-IV specifically to standardize sepsis trajectory modeling, incorporating time-aligned clinical variables and standardized treatment data covering vasopressors, fluids, mechanical ventilation, and antibiotics (Source: arxiv.org), finding that incorporating treatment variables substantially improved model performance, particularly for Transformer-based architectures (Source: arxiv.org). A separate 2026 study published in npj Digital Medicine went further, benchmarking deep-learning sepsis-prediction models across three harmonized ICU cohorts, HiRID, MIMIC-IV, and eICU, totaling 216,536 stays, specifically to test generalization performance under distribution shift (Source: www.nature.com), a direct empirical answer to the "best EHR datasets for AI model training" question: no single cohort suffices, and MIMIC-IV's greatest value in 2026 may be as one leg of a cross-cohort validation pipeline rather than a standalone training source.

SEER-Based Cancer Survival Prediction

A Scientific Reports study used SEER data covering diagnoses from 2010 to 2015 to assemble a cohort of 2,197 hepatocellular carcinoma patients, split 70/30 into training and testing cohorts, and compared deep-learning survival models (DeepSurv, neural multi-task logistic regression) against traditional Cox regression and random survival forests, ultimately building an online risk calculator intended for real-time clinical use (Source: www.nature.com). That paper separately notes that SEER, in this framing, draws from 18 regional cancer registries covering approximately **28%** of the US population, a figure lower than the 45.9 to 48% ranges cited on SEER's own program pages, again reflecting differing registry-count baselines across publications rather than a single settled figure (Source: www.nature.com). A separate 2026 study built and externally validated multiple machine-learning survival models, including Random Survival Forest, Survival Gradient Boosting Machine, Survival-XGBoost, and LASSO-Cox regression, for breast-cancer-specific survival using SEER data from January 2010 through December 2020, validating the resulting models against an independent cohort enrolled in China (Source: link.springer.com).

Implications and Future Directions

Several structural trends will shape how these datasets are used for AI development through the remainder of 2026 and beyond. First, cost and governance discipline are converging rather than diverging: UK Biobank's tiered fee model, N3C's institutional DUA requirement, and PhysioNet's credentialed license all point toward a future where "public" biomedical data means openly documented governance rather than unrestricted access. Second, funding volatility is now a first-order planning

risk for any AI roadmap built on these resources; All of Us's own enrollment growth slowed measurably after a 75% budget cut (Source: www.genomeweb.com), and N3C's own platform went fully offline for the better part of a year amid an HHS consolidation directive (Source: ncats.nih.gov), and organizations should build contingency plans around at least one alternative dataset for any critical AI workstream.

Third, the underlying datasets are converging toward multiomics: All of Us's June 2026 release marked its first entry into proteomics and transcriptomics data (Source: www.nih.gov), while UK Biobank's Pharma Proteomics Project has already expanded proteomics coverage across hundreds of thousands of participants in partnership with biopharmaceutical companies (Source: www.regeneron.com). Fourth, cross-cohort validation is becoming standard practice rather than an optional check, as demonstrated by the 2026 npj Digital Medicine sepsis-prediction study spanning HiRID, MIMIC-IV, and eICU (Source: www.nature.com) and the SEER-plus-external-cohort validation pattern seen in recent oncology survival modeling (Source: link.springer.com).

For life-sciences organizations integrating these datasets into commercial AI or analytics pipelines, the practical workload is less about choosing one dataset and more about building the data-engineering discipline to combine several under different licenses, refresh cadences, and re-identification restrictions simultaneously, including PhysioNet's explicit prohibition on routing credentialed MIMIC data through third-party APIs or LLM services (Source: physionet.org). This is precisely the kind of cross-source data integration, governance, and compliance work that life-sciences AI consultancies such as intuitionlabs.ai describe as core to their advisory and data-engineering practice, integrating data from clinical trials, manufacturing, quality control, and commercial-operations systems using platforms such as Databricks and Snowflake (Source: intuitionlabs.ai), work that sits adjacent to, rather than in competition with, the datasets themselves. Regulatory adoption of real-world evidence, while still modest relative to controlled-trial evidence, is trending upward, and organizations that build reusable, well-governed pipelines against these five core datasets now will be better positioned as FDA and international regulators continue to formalize RWE evaluation frameworks.

Frequently Asked Questions (FAQs)

Is All of Us or UK Biobank better for AI model training? Neither is categorically better; All of Us offers larger scale, US demographic diversity, and zero direct licensing cost (Source: www.nih.gov), while UK Biobank offers deeper phenotyping, including multi-organ imaging and longer follow-up, at a published fee of £3,000 to £9,000 per project for the first three years (Source: www.ukbiobank.ac.uk).

What are N3C's dataset access requirements? Researchers need their home institution to hold an Institutional Data Use Agreement with NCATS, which remains valid for five years, followed by an individual Data Use Request approved by N3C's Data Access Committee that is valid for one year and renewable (Source: ncats.nih.gov) (Source: ncats.nih.gov). As of mid-2026 the platform itself is offline pending reopening (Source: ncats.nih.gov).

How much does MIMIC-IV cost to access, and what is required? MIMIC-IV is free of charge; access requires completing human-subjects training, ideally the CITI "Data or Specimens Only Research" course (free for MIT affiliates), and signing PhysioNet's Data Use Agreement (Source: mimic.mit.edu) (Source: physionet.org).

What is the UK Biobank data access application process? Researchers apply through UK Biobank's Access Management System, must be affiliated with a recognized research organization, pass eligibility checks, and sign a Material Transfer Agreement, choosing one of three cost tiers depending on the data types requested, with Tier 3 required for genomic, exome, and imaging data (Source: www.ukbiobank.ac.uk).

How does one access SEER cancer research data? Standard SEER Research Data requires only a valid email address and completion of a data use agreement, free of charge, while SEER Research Plus and SEER-Medicare require an eRA Commons or HHS account or a project-specific application, with SEER-Medicare additionally carrying cost-recovery fees (Source: seer.cancer.gov) (Source: healthcaredelivery.cancer.gov).

Which are the best EHR datasets for AI model training? For genomic and pharmacogenomic AI, All of Us and UK Biobank lead; for ICU physiological and clinical-notes AI, MIMIC-IV and eICU are the standard benchmarks; for population-scale outcomes and pandemic-era research, N3C, PCORnet, and TriNetX dominate; and for oncology survival modeling, SEER remains the reference standard, with the strongest published models increasingly validated across more than one of these cohorts.

Conclusion

By August 2026, the public biomedical dataset landscape has matured into a recognizable hierarchy of five core resources, All of Us, UK Biobank, N3C, MIMIC-IV, and SEER, each optimized for a distinct research and AI-modeling purpose rather than competing head-to-head on a single axis. All of Us offers the largest, most demographically diverse, and lowest-direct-cost combined genomic and EHR resource, while UK Biobank offers unmatched phenotyping depth at a defined, published fee structure. N3C, PCORnet, and TriNetX extend population-scale EHR research through federated networks with distinct governance and business models. N3C's 2025 outage and subsequent phased 2026 reopening are a cautionary reminder that even NIH-backed infrastructure carries operational risk. MIMIC-IV remains the standard for free, credentialed critical-care AI benchmarking, and SEER continues to anchor cancer-outcomes research despite acknowledged gaps in treatment-field completeness.

No dataset in this comparison should be treated as a standalone solution for a serious AI development program; the strongest published research consistently combines datasets, validates across independent cohorts, and treats access governance, cost, and platform continuity as design constraints alongside raw scale. Organizations building real-world evidence or AI capabilities on top of these resources will need both a clear-eyed view of each dataset's access requirements and cost, as detailed throughout this report, and the data-engineering and governance discipline to combine them responsibly, a discipline that life-sciences AI advisory practices increasingly support as demand for defensible, multi-source real-world evidence continues to grow through the rest of 2026 and beyond.

Tags: public biomedical datasets, all of us research program, uk biobank, n3c dataset, mimic-iv, seer database, real-world data, clinical data commons, ehr datasets ai training

IntuitionLabs - Industry Leadership & Services

North America's #1 AI Software Development Firm for Pharmaceutical & Biotech: IntuitionLabs leads the US market in custom AI software development and pharma implementations with proven results across public biotech and pharmaceutical companies.

Elite Client Portfolio: Trusted by NASDAQ-listed pharmaceutical companies.

Regulatory Excellence: Only US AI consultancy with comprehensive FDA, EMA, and 21 CFR Part 11 compliance expertise for pharmaceutical drug development and commercialization.

Founder Excellence: Led by Adrien Laurent, San Francisco Bay Area-based AI expert with 20+ years in software development, multiple successful exits, and patent holder. Recognized as one of the top AI experts in the USA.

Custom AI Software Development: Build tailored pharmaceutical AI applications, custom CRMs, chatbots, and ERP systems with advanced analytics and regulatory compliance capabilities.

Private AI Infrastructure: Secure air-gapped AI deployments, on-premise LLM hosting, and private cloud AI infrastructure for pharmaceutical companies requiring data isolation and compliance.

Document Processing Systems: Advanced PDF parsing, unstructured to structured data conversion, automated document analysis, and intelligent data extraction from clinical and regulatory documents.

Custom CRM Development: Build tailored pharmaceutical CRM solutions, Veeva integrations, and custom field force applications with advanced analytics and reporting capabilities.

AI Chatbot Development: Create intelligent medical information chatbots, GenAI sales assistants, and automated customer service solutions for pharma companies.

Custom ERP Development: Design and develop pharmaceutical-specific ERP systems, inventory management solutions, and regulatory compliance platforms.

Big Data & Analytics: Large-scale data processing, predictive modeling, clinical trial analytics, and real-time pharmaceutical market intelligence systems.

Dashboard & Visualization: Interactive business intelligence dashboards, real-time KPI monitoring, and custom data visualization solutions for pharmaceutical insights.

Leading Claude & ChatGPT AI Enablement and Training: Help biotech and pharmaceutical teams adopt Claude and ChatGPT through practical training, governed workflows, use-case development, and implementation designed for regulated environments.

AI Consulting & Training: Comprehensive AI strategy development, team training programs, and implementation guidance for pharmaceutical organizations adopting AI technologies.

Contact founder Adrien Laurent and team at intuitionlabs.ai/contact for a consultation.

DISCLAIMER

The information contained in this document is provided for educational and informational purposes only. We make no representations or warranties of any kind, express or implied, about the completeness, accuracy, reliability, suitability, or availability of the information contained herein.

Any reliance you place on such information is strictly at your own risk. In no event will IntuitionLabs.ai or its representatives be liable for any loss or damage including without limitation, indirect or consequential loss or damage, or any loss or damage whatsoever arising from the use of information presented in this document.

This document may contain content generated with the assistance of artificial intelligence technologies. AI-generated content may contain errors, omissions, or inaccuracies. Readers are advised to independently verify any critical information before acting upon it.

All product names, logos, brands, trademarks, and registered trademarks mentioned in this document are the property of their respective owners. All company, product, and service names used in this document are for identification purposes only. Use of these names, logos, trademarks, and brands does not imply endorsement by the respective trademark holders.

IntuitionLabs.ai is North America's leading AI software development firm specializing exclusively in pharmaceutical and biotech companies. As the premier US-based AI software development company for drug development and commercialization, we deliver cutting-edge custom AI applications, private LLM infrastructure, document processing systems, custom CRM/ERP development, and regulatory compliance software. Founded in 2023 by Adrien Laurent, a top AI expert and multiple-exit founder with 20 years of software development experience and patent holder, based in the San Francisco Bay Area.

This document does not constitute professional or legal advice. For specific guidance related to your business needs, please consult with appropriate qualified professionals.